Logic, Epistemology, and the Unity of Science 36

Theodora Achourioti
Henri Galinon
José Martínez Fernández
Kentaro Fujimoto *Editors*

Unifying the Philosophy of Truth

Springer

Logic, Epistemology, and the Unity of Science

Volume 36

Logic, Epistemology, and the Unity of Science aims to reconsider the question of the unity of science in light of recent developments in logic. At present, no single logical, semantical or methodological framework dominates the philosophy of science. However, the editors of this series believe that formal techniques like, for example, independence friendly logic, dialogical logics, multimodal logics, game theoretic semantics and linear logics, have the potential to cast new light on basic issues in the discussion of the unity of science.

This series provides a venue where philosophers and logicians can apply specific technical insights to fundamental philosophical problems. While the series is open to a wide variety of perspectives, including the study and analysis of argumentation and the critical discussion of the relationship between logic and the philosophy of science, the aim is to provide an integrated picture of the scientific enterprise in all its diversity.

More information about this series at http://www.springer.com/series/6936

Theodora Achourioti • Henri Galinon
José Martínez Fernández • Kentaro Fujimoto
Editors

Unifying the Philosophy of Truth

Springer

Editors
Theodora Achourioti
ILLC & AUC
University of Amsterdam
The Netherlands

Henri Galinon
PHIER
Université Blaise Pascal
Clermont Ferrand
France

José Martínez Fernández
Lògica Història i Filosofia de la Ciència
Universitat de Barcelona
Barcelona
Barcelona
Spain

Kentaro Fujimoto
Department of Mathematics
and Department of Philosophy
University of Bristol
UK

ISSN 2214-9775 ISSN 2214-9783 (electronic)
Logic, Epistemology, and the Unity of Science
ISBN 978-94-017-9672-9 ISBN 978-94-017-9673-6 (eBook)
DOI 10.1007/978-94-017-9673-6

Library of Congress Control Number: 2015938702

Springer Dordrecht Heidelberg New York London

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Contents

List of Contributors

Theodora Achourioti ILLC & AUC, University of Amsterdam, The Netherlands

Bradley Armour-Garb Department of Philosophy, HU-257 University at Albany-SUNY, Albany, NY, USA

Denis Bonnay Université Paris Ouest, IRePh & IHPST, Nanterre, France

Riccardo Bruni Università degli Studi di Firenze, Firenze, Italy

Andrea Cantini Dipartimento di Filosofia, Università degli studi di Firenze, Firenze, Italy

Cezary Cieśliński University of Warsaw, Warsaw, Poland

Pablo Cobreros Universidad de Navarra, Pampalona, Spain

John Collins University of East Anglia, Norwich, England

Sebastian Eberhard Institut für Informatik und angewandte Mathematik, Universität Bern, Bern, Switzerland

Paul Egré Institut Jean Nicod, Ecole Normale Supérieure-PSL Research University, Paris, France

Ali Enayat Department of Philosophy, Linguistics, and Theory of Science, University of Gothenburg, Gothenburg, Sweden

Pascal Engel EHESS, Paris, France

Martin Fischer Oak Park, Il, USA

Kentaro Fujimoto Department of Mathematics and Department of Philosopy, University of Bristol, Bristol BS8 1TW, UK

Henri Galinon PHIER, Université Blaise Pascal, Clermont Ferrand, France

Michael Glanzberg Northwestern University, Chicago/Evanston, USA

Volker Halbach New College, University of Oxford, Oxford, UK

Leon Horsten University of Bristol, Bristol, UK

Reinhard Kahle CMA and DM, FCT, Universidade Nova de Lisboa, Lisboa, Portugal

Graham E. Leigh Institute of Discrete Mathematics and Geometry, Vienna University of Technology, Vienna, Austria

José Martínez-Fernández Lògica, Història i Filosofia de la Ciència, Universitat de Barcelona, Barcelona, Barcelona, Spain

Friederike Moltmann IHPST (Paris 1/ ENS/ CNRS), Paris, France

Julien Murzi University of Salzburg, Salzburg, Austria

Graham Priest Departments of Philosophy, The Graduate Center, City University of New York, New York, USA, and the University of Melbourne, Australia

Stephen Read University of St Andrews, St Andrews, Scotland

David Ripley University of Connecticut, Storrs, CT, USA

Robert van Rooij University of Amsterdam, Amsterdam, Netherlands

Philippe de Rouilhan IHPST (Paris 1 – Panthéon-Sorbonne/ ENS/ CNRS), Paris, France

Lionel Shapiro University of Connecticut, Storrs, CT, USA

Michael Sheard Rhodes College, Memphis, TN, USA

Gila Sher University of California, San Diego, USA

Thomas Strahm Institut für Informatik und angewandte Mathematik, Universität Bern, Bern, Switzerland

Albert Visser Department of Philosophy, Bestuursgebouw, Utrecht, The Netherlands

Floris Tijmen van Vugt University of Music, Drama and Media, Hannover, Germany

IMMM, Hannover, Germany

Lyon Neuroscience Research Center, Université Claude Bernard Lyon-1, Villeurbanne, France

James A. Woodbridge Department of Philosophy, University of Nevada, Las Vegas, NV, USA

Elia Zardini LanCog, University of Lisbon, Lisbon, Portugal

Chapter 1
Introduction

Theodora Achourioti, Kentaro Fujimoto, Henri Galinon and José Martínez-Fernández

1.1 Presentation of the Volume

In 2011, and within only a few months, four international conferences on truth were independently organized in Amsterdam ("Truth be told", 23–25 March 2011), Barcelona ("BW7: Paradoxes of truth and denotation", 14–16 June 2011), Paris ("Truth at work", 20–23 June 2011) and Oxford ("Axiomatic theories of truth", 19–20 September 2011). This succession of events and the original work presented at them are evidence that the *philosophy of truth* is a lively and very diverse area of study. They saw a great variety of methodologies from philosophers, logicians and linguists, and even within these groups, a variety of problems and approaches to those problems. We think, however, that the interaction between the different research programmes was not as intense as it could have been. By collecting in one volume a wide range of the very latest research on truth, we hope to intensify the dialogue between philosophers and thus make a contribution to even better informed

T. Achourioti
ILLC & AUC, University of Amsterdam, Science Park 113, 1098 XG,
Amsterdam, The Netherlands
e-mail: t.achourioti@uva.nl

K. Fujimoto
Department of Mathematics and Department of Philosophy,
University of Bristol, University Walk, Clifton, Bristol BS8 1TW, UK
e-mail: kentaro.fujimoto@bristol.ac.uk

H. Galinon
PHIER, Université Blaise Pascal, Clermont Ferrand, France
e-mail: henri.galinon@univ-bpclermont.fr

J. Martínez-Fernández
Lògica, Història i Filosofia de la Ciència, Universitat de Barcelona, Montalegre 4,
08001 Barcelona, Barcelona, Spain
e-mail: jose.martinez@ub.edu

T. Achourioti et al. (Eds.), *Unifying the Philosophy of Truth,* Logic, Epistemology,
and the Unity of Science 36, DOI 10.1007/978-94-017-9673-6_1

research in the future. Hence the title of this volume, *Unifying the philosophy of truth*, which announces our project.

We are very glad that Springer agreed to host this volume within its 'Logic, Epistemology, and the Unity of Science' series. Although—as illustrated by the essays in this volume—contemporary research on truth is now mainly pursued in full independence of the early unity of science programme, this is a good place to recall that positivism played an essential role in the birth of contemporary research on truth in the first half of the twentieth century. The spirit of scientific empiricism and logical rigor promoted by positivism was at first at odds with the use of the concept of truth in philosophy and science. In the 1920s the philosophical notion of truth was under threat from various angles. First, it was not clear at that time what the appropriate conceptual analysis of the notion of truth should be. The traditional conception of truth as correspondence between discourse and what discourse is about was perhaps shared by many philosophers, but it was not seen as amenable to analysis in logical and empirical terms as positivists required for meaningful discourse. Thus, the metaphysical notion of truth was rejected. Second, it seemed all too evident to some philosophers[1] that the rehabilitation of the notion of truth would be a threat to the way they had conceived of their physicalist unitarian project. If the traditional notion of truth were accepted among scientific notions, that would pave the way for truth-conditional semantics, which would then threaten verificationist semantics as the scientific basis for the explanation of meaning. However, although it seems relatively clear that a sentence such as the then all too famous "Das Nichts selbst nichtet"[2] has no verification conditions, it is not equally clear that truth-conditions cannot be used instead to specify its meaning. In other words, the positivist critique of metaphysics would lose its bite if the language of metaphysics could be shown to be meaningful in terms of truth-conditions. Third, even in mathematics, truth-talk was not very fashionable in those days, as Hilbert's formalist programme was very prominent as a philosophy of mathematics.[3] And finally, the concept of truth had notoriously been involved in paradoxes for more than two thousand years. Even if it had not always been clear whether those paradoxes should be taken seriously or not[4], this could not have helped build trust in the notion of truth as a legitimate one, even less so at the beginning of the twentieth century which *was* a time—if any were—in which paradoxes were taken seriously. It is in this historical context that Tarski—following up on work within the Polish school[5]—published his celebrated essay on the notion of truth in formalized languages (Tarski 1983), giving birth to

[1] Neurath in particular. See, e.g. Mancosu (2009).
[2] See, e.g. Carnap (1931).
[3] Remember that in this context Gödel himself, despite his realist convictions in mathematics, carefully avoided use of the term 'truth' in his incompleteness paper. See Feferman (1989).
[4] But see Read's paper in this volume with respect to the Middle Ages.
[5] On the roots of Tarski's work in the Polish school, see Wolenski (2009) and Wolenski and Murawski (2008).

the contemporary research on truth.[6] Part of Tarski's philosophical work and long-standing success in the rehabilitation of the notion of truth is explained by his meeting the positivists' strictures[7] showing, in effect, how in many circumstances the notion of truth can be rigorously defined in a scientifically acceptable language: a paradox- and metaphysics-free language for science. He did so in a series of writings widely acknowledged as a model of balance between philosophical insight and scientific achievement.

Today, another close connection between truth theorizing and the unity of science programme is found in deflationism. Deflationism is probably the most discussed philosophical approach in contemporary philosophical research on truth and it appears in several places in this volume.[8] Some versions of deflationism have been motivated by concerns raised by forms of physicalism or radical empiricism with the nature of truth-theoretical explanations.[9] For it has been argued that physicalism implies that the property of truth is reducible to physical properties or, if not, that it must have no explanatory force. But a full reduction of the property of truth to empirical properties appeared to be hard to achieve[10] and this put some pressure on the physicalist to accept that the notion of truth has no explanatory force. Rather than return to a pre-Tarski state of affairs in which the notion of truth is removed from the language of science, the deflationist's twist is to maintain that the notion of truth is still legitimate in scientific talk, but for logical and not explanatory purposes. This position has in turn stimulated a number of discussions to which philosophy, logic and linguistics have all made contributions, some of which are discussed in detail in this volume.[11]

It seems safe to say that most philosophers, logicians and linguists do not adhere to the early—or late—positivist unity of science programme. But from that tradition we retain the goal of a scientifically informed philosophy of truth. Contemporary philosophy of truth has lain all along at the crossroads of logical, empirical and

[6] One has to add that in 1935 the context had already dramatically changed in the philosophy of mathematics after the publication of Gödel's incompleteness theorems in 1931.

[7] For more on Tarski's relationship with philosophers close to the Vienna Circle, see e.g. Mancosu (2008a, 2008b, 2009).

[8] This is not to say that most philosophers are deflationists. Indeed, Bourget and Chalmers' recent survey of professional philosophers (Bourget and Chalmers 2014), shows that correspondence theory is still the most widely shared view on truth. Within the sample faculty population (admittedly strongly biased towards North American philosophy departments), the results for the different conceptions of truth are as follows ("accept" and "lean towards" answers are aggregated): correspondence 50.8 %; deflationary 24.8 %; epistemic 6.9 %; other 17.5 %.

[9] We have in mind the deflationist tradition running from W. V. Quine (1970, 1960) to H. Field (2001), including the work of, e.g. S. Leeds (1978). P. Horwich (1998a, 1998b) could perhaps also be attached to this tradition, even though he seems to conceive of his deflationism as a philosophical elucidation in the tradition of Wittgenstein's philosophy of language.

[10] Thus, according to Field (1972), Tarski's work did not establish the reducibility of truth to empirical properties. Later attempts, such as e.g. Fodor (1989) in the context of a defence of "intentional realism", have also been criticized. See Loewer (1997) for an overview of naturalizing semantics. Field (2001) illustrates clearly Field's route from physicalism to deflationism about truth.

[11] See in particular Chapts. 4 and 5 of the present volume.

philosophical research, stimulating ever more interaction between them: applications of logical methods help decide philosophical matters (e.g. on the nature of truth[12]), philosophical reflection informs logic research (e.g. on paradoxes[13]) and so on. We hope that the present volume further encourages this ongoing dialogue between philosophy, logical methods and empirical work.

1.2 Organization of the Volume

Research in the philosophy of truth has expanded in many different directions over recent decades. The present volume could not possibly cover the full range of actively researched truth-related topics; however, it does provide an overview of some of the main themes that run through the work currently undertaken within the area in the analytic tradition. This is done directly, through the broad range of topics that the papers address, as well as indirectly, via the authors' reference to others' work that relates to their own.

We have grouped the papers into six chapters (2–7 of this volume). Here we introduce each of the chapters starting with a general presentation of the papers they contain in the order that they occur. This is to give the reader a first idea of what the papers in each chapter are about before we go on to introduce each of them separately. An introduction to each of the papers contained in the chapter then follows.

The topics addressed in different sections of this volume often relate to each other; by no means do we consider our organization of the volume the only possible one. Some of the connections are pointed out by the authors themselves, others we try to highlight in our introduction. A goal that we hope to have achieved by putting the papers side by side the way that we have is to help draw philosophical connections between the papers that go beyond the particular methodologies used. We have, therefore, opted for a divide that cuts across the usual distinctions—distinctions we do not really ascribe to—e.g. between philosophical, logical and linguistic papers. We hope that this, perhaps somewhat unorthodox, organization of the volume will prove helpful in further emphasizing connections between the papers by also drawing the readers' attention to work that they may not at first consider as immediately relevant to their own.

The introduction of the separate papers is not balanced. For each paper we present what we anticipate will be most useful to the non-specialized reader. For example, in cases where we judge that background knowledge of certain issues is required, we have tried to provide part of that background, occasionally at the expense of expanding on the paper's original ideas. In cases where the argumentation of the paper is rather complex, we have opted for a concise presentation of the argumentative structure or an exposition of the preliminary context which then makes it easier to

[12] See Ketland (1999) and Shapiro (1998) on deflationism and conservativity, and Chap. 5 in this volume.

[13] See, in this volume, Chaps. 6 and 7.

penetrate and appreciate the work. With this selective approach we hope to facilitate a comparative reading much more than could be achieved by a uniform exposition, which in some cases would favour the specialized reader.

Finally, it is often said that truth theorists should clarify their philosophical aims and presuppositions before delving into technical details; but we believe that philosophical deliberation and logical analysis should go hand in hand and complement each other. The interaction between ideas and formal techniques is generally a highly complex and intricate affair. Philosophical thinking may help clarify the ideas behind formalization, while reflection on technical work may lead to progress in philosophical thinking. This volume can be read as an illustration of this interactive process.

1.2.1 Truth and Natural Language

The first two papers in this volume offer a critical reflection on the transition from natural to formal language and to philosophical theories of truth. It is common for philosophers to use examples of sentences in natural language that contain the word 'true'. For example, deflationists often explain the non-substantive character of truth by noting that a sentence of the form 'it is true that A' has the same meaning as 'A' itself (an idea that goes back as far as Frege 1918). In her essay, Moltmann studies the linguistics of truth predication which she regards to be a complex phenomenon of natural language. A linguistic study of truth is of immediate relevance not only to philosophers who want to formalize the ordinary notion of truth but also to those who simply wish to align their theory with the behaviour of truth in natural language.[14] Moltmann's analysis shows that far from supporting a single philosophical theory, some of the ways truth predication manifests itself challenge views of truth that are as prominent as deflationism, or those that consider propositions to be the primary bearers of truth.

In a more foundational approach to the very endeavour of formalizing the notion of truth in natural language, Collins objects to a dilemma that seems to drive some of the contemporary discussions: either a consistent theory of untyped truth has to be developed or natural language is inconsistent because it contains a paradoxical notion of truth. Collins' view is that the paradoxical character of truth is inescapable and yet this does not imply that natural language is inconsistent, because questions of consistency can only meaningfully apply to formal theories, which natural language is not. In the essay following Collins', Sheard claims that it is still possible to identify consistent uses of an inconsistent notion and, furthermore, that such consistent uses are evidenced in the case of truth by the fact that paradoxes do not hinder speakers of ordinary language when communicating with each other when using the word 'true'.

[14] Think of questions concerning whether truth should be formalized as a predicate or an operator, whether it is an iterable notion, etc.

These consistent fragments of the use of the truth predicate in natural language can be analysed as inferential mechanisms wedded to specific communicative tasks. One can then study, in a spirit of truth-theoretic pluralism, which of the available axiomatic theories of truth offer the principles needed for carrying out separate tasks; thereby setting specific standards against which existing theories can be adjudicated (which is desirable nowadays given the number of interesting axiomatic theories of truth available).

1.2.1.1 "Truth Predicates in Natural Language" by Friederike Moltmann

Moltmann takes a close look at the appearance of truth in natural language and asks whether the linguistic data support known philosophical views of truth; or weaker than that, whether they are compatible with them. She does not focus her critical study on one particular philosophical theory, nor is there one such theory that is naturally favoured by the linguistic data provided by her study, although some philosophical positions are either excluded or significantly challenged. In fact, this paper is in line with so-called truth-theoretic pluralism: the view that there may be more than one viable notion of truth. Pluralism regarding truth is not a recent view; it has famously been defended, for example, by Lynch (2009). In the present volume, truth-theoretic pluralism is also found to be agreeable by Halbach and Horsten, who take their cue in this from Sheard (1994).

It is an essential assumption underlying Moltmann's analysis that truth predication be regarded as a phenomenon that extends beyond the occurrence of the word 'true'. Truth can be predicated by several expressions, here called 'apparent truth predicates', of which the standard truth predicate is only one. 'Apparent truth predicates' can either ascribe the property of truth (type 1 truth predicates), or express a relation of truth (type 2 truth predicates).

With respect to type 1 predicates, Moltmann argues that the semantics of natural language does not support an operator analysis. Such an analysis for 'is true' has been proposed, for example, by Grover et al. (1975), Grover (1992), Brandom (1994) and Mulligan (2010), and recent formal work on truth has again raised the question of whether truth should be formalized as an operator. Moltmann shows that a truth predicate does not exhibit distinctive sentential semantics such as one finds with expressions that are clear cases of operators, e.g. 'is possible'. She also shows that there is no reason to regard the linguistic form 'it is true that A' (that-clause in extraposition) as more representative of the occurrence of truth in natural language than its equivalent 'that A is true' (that-clause in subject position); whereas an operator approach does favour the former over the latter. Moreover, it is shown that there is no more reason to study constructions in which truth is predicated of that-clauses than nominal expressions; the latter almost universally neglected by philosophical theories of truth.

Under type 1 truth predicates, Moltmann also places normative predicates that are used to convey truth, such as 'is correct' or 'is right'. This gives rise to two notions of truth: representation-related and normative-related, which are combined in normative

truth-predicates and difficult to separate out. Moltmann's proposal is that studying the semantics of these normative predicates will provide insight into the nature of truth predication itself. She concludes, for example, that truth is predicated over intentional entities (attitudinal objects, see Moltmann 2003, 2013) rather than mind-independent entities, such as propositions (or sentences), which is what deflationists about truth traditionally claim (e.g. Horwich 1998). Note that the view that truth is predicated over intentional objects such as beliefs is already found in Ramsey (see Ramsey 1991, p. 8). It follows that the viability of the deflationist view for such predicates, given their semantics, depends on the possibility of distilling a purely representational role for truth.

1.2.1.2 "Truth and Language, Natural and Formal" by John Collins

Collins' essay is as much about the use of truth in natural language as about the paradoxes. For its starting point, recall that Tarski sees the paradoxes as the outcome of (1) the T-biconditionals that characterize the concept of truth, (2) the classical logic that we employ in reasoning about truth and (3) the fact that natural language can speak about everything and in particular it can speak about itself. This diagnosis leads Tarski to what has been called the 'inconsistency view' of truth: since the T-biconditionals are essential to define truth and classical logic should not be modified, one has to admit that paradoxes are produced because natural languages are universal, i.e. they contain their own truth predicate. This implies that natural languages use truth in an inconsistent way. The solution should therefore come from creating non-universal thoroughly-specified formal languages with rich expressive resources that can consistently incorporate a truth definition that implies all the T-biconditionals. And Tarski showed us how to do just that.

The Tarskian analysis of truth has been severely criticized and a new orthodoxy has been developed: the expressiveness of natural language should not be compromised, and the new goal in solving the paradox is to devise powerful formal languages that can speak about themselves to the extent of expressing all paradoxical sentences and still have a consistent truth predicate, even if this implies tinkering with classical logic.

In his paper Collins wants to challenge both sides of this discussion: he wants to defend, against most contemporary solutions to the paradox, the notion that Tarski was right in his diagnosis of the inherently paradoxical nature of the notion of truth in natural language; but he also wants to criticize defenders of an inconsistency view of truth for understanding natural language as inconsistent. Collins proposes a different interpretation of Tarski's view, one that respects the basic intuition that paradoxes are insoluble in natural language, while, at the same time, it does not see natural language as inconsistent.

The first part of the paper compares natural and formal languages. Formal languages are characterized by having an explicit stipulation of their syntax and a full transparent semantics. The full transparency of the semantics means that the syntactic conditions express the semantic properties of the language in such a way that

the semantic properties can be read off from the syntax. Formal languages are guaranteed by design to have these features. In contrast, natural languages are not fully transparent. Collins develops this idea by focusing on five linguistic phenomena: (i) ambiguity, which shows that syntactic structure is not always an accurate guide to interpretation; (ii) the presence of words that do not make any contribution to the sentences they appear in; (iii) the absence of words that should be present in a sentence; (iv) the abundance of positions in sentences that do not serve to predict the interpretation of their occupiers; and (v) the fact that there is no decidable notion of being a well-formed formula of natural language because the acceptance and the meaning of sentences depend on the psychological states of speaker-hearers of the language.

In the second part of the paper, after rejecting two objections to his understanding of natural languages, Collins focuses on the concept of truth in natural and formal languages. Collins claims that the concepts of consistency and inconsistency apply only to formal languages and cannot apply to natural languages because a natural language is not a set of fully transparent sentences such that we could have a consistent or inconsistent theory of it. He criticizes the opposing views of two contemporary defenders of the inconsistency view: Eklund and Patterson (Eklund 2001; Patterson 2008, 2009). Collins also criticizes the views of authors who propose formal languages that could model the universal aspect of natural language and still escape (or at least modulate) inconsistency. Against these views, Collins argues that paradoxical arguments are unavoidable in natural language, due to the inherent riskiness of the truth predicate. This riskiness is produced because, as Kripke (1975) pointed out, one can predicate truth of a set of sentences without knowing the content of the sentences themselves.

1.2.1.3 "Truth and Trustworthiness" by Michael Sheard

Sheard starts by observing that, as opposed to the theoretical case, the use of an untyped truth-predicate in real-life communication appears unproblematic: natural language users have no problem understanding each other when they use the word 'true', irrespective of their potentially different philosophical ideas concerning truth. Sheard proposes this as evidence of some kind of inferential semantics that operates on the surface level of language use and which is shared by language users. This inferential semantics must be consistent, since paradoxes do not seem to arise (or are somehow avoided) in everyday communication. The alternative to allowing for consistent uses of the truth predicate would be to consider language users as irrational beings, in which case empirical psychological work should explain how they manage to deal with inconsistencies; but, according to Sheard, this is a more general question which would not necessarily shed any light on the question: what is the inferential mechanism that is at work in specific situations when people use the word 'true'?

Sheard focuses on the use of 'true' in its function of conveying information. He constructs a simple scenario consisting of two (idealized) agents, a speaker and a hearer, where the speaker conveys a message with the help of the truth predicate by

means of an assertion, a denial or a generalization. The task of the hearer is to decode the speaker's message and to assimilate the knowledge it contains. Sheard then asks of each of three prominent axiomatic theories of truth, so-called FS, KF and VF[15], whether it provides a mechanism for the hearer to perform this decoding act. He observes that this very much depends on whether the hearer considers the speaker to be a trustworthy source, since if not, the hearer first has to check that the message does not lead to inconsistency before assimilating it. Decoding a message is pretty straightforward for all three systems and message forms in the case of a trustworthy source, with the exception of denial, which one has to formalize $T(\ulcorner \neg A \urcorner)$ instead of $\neg T(\ulcorner A \urcorner)$ (this is because decoding $\neg A$ from $\neg T(\ulcorner A \urcorner)$ requires the inference from A to $T(A)$ which is not generally available). Matters become complicated in the case of an untrustworthy source. Logically, KF and VF are equipped to deal with incoming inconsistencies since they are closed under *reductio ad absurdum*, but note that idealization assumptions become crucial here, i.e. the ability of the hearer to screen all logical consequences of existing knowledge for inconsistency.[16]

Decoding messages is discussed by Sheard as a seemingly simple example of a communicative task that allows a comparison between different axiomatic theories of truth. One should not, however, be surprised if a certain theory that fares well in this context does much worse than other theories once one changes the task at hand—Sheard gives another example to this effect. In fact, one should not expect there to be a single axiomatic theory that can account for all communicative uses of truth.

Sheard's approach is, therefore, compatible with both truth-theoretic pluralism and inconsistency theories of truth, since an inconsistent notion may allow for mutually incompatible, yet consistent, uses of it. There are two main reasons for engaging in this exercise of assessing axiomatic truth theories against simple communicative tasks: first, it offers more insight into the philosophical and inferential import of these theories; and second, it provides criteria for adjudicating between the theories. The theme of adjudicating between axiomatic theories of truth is also taken up later in this volume by Halbach and Horsten in chap. 12.

1.2.2 Uses of Truth

Emphasis is also placed on the non-paradoxical features of truth by Rouilhan who demonstrates how Davidsonian truth-theoretic meaning explanations could be used

[15] For an exposition of these theories, the reader can consult Halbach (2010).

[16] Sheard notes that FS presents the additional difficulty of not allowing for *reductio* reasoning due to its lack of a general deduction theorem. The best one can do, Sheard explains, is to provisionally accept a message and use the FS inference rules as a test mechanism in order to track potential inconsistencies while resisting the final step of the reductio argument, which means that instead of adding the negation of the provisionally accepted message to the database, the hearer simply dismisses the message altogether.

to debunk another paradox. 'Frege's paradox'—or rather, a general version of it—arises when meaning explanations for a language intended to be used as the language of science must make use of grammatical categories which clash with the logical structure of that language. Rouilhan's paper shows that appropriate uses of the notion of truth make it possible to give meaning explanations for a language of science that obey its type-theoretical logical structure, and thus comply with the universalism of founders of modern logic[17], that is, in a way that does not condemn these explanations to being nonsensical.

Kahle's 'Sets, Truth, and Recursion' illustrates how the notion of truth can be applied to foundational topics in mathematics, especially set theory. More specifically, Kahle's essay presents a set theory based on an axiomatic truth theory, where sets and the membership relations are *defined in terms of a truth predicate*. The set theory is a so-called *Frege structure*, roughly a way to restrict truth-theoretic assumptions and objects of the theory so as to maintain full comprehension. It is then a consequence of these restrictions that Frege structures support a non-classical concept of truth.

Still on the foundational side, Eberhard and Strahm explore the use of truth in 'unfolding' the content of arithmetic theories. The *unfolding programme* has famously been developed by Feferman and addresses a query of Kreisel's about the proof-theoretic commitments that one implicitly makes when accepting a certain theory. Eberhard and Strahm previously worked on theories of truth for feasible arithmetic, that is, arithmetic weaker than PA, which describes feasibly computable functions (usually identified with polynomial time algorithms). Here they consider the use and strength of such theories in carrying out the unfolding programme.

Finally, Bruni explores a fragment of the revision theory of truth, which was famously developed by Herzberger, Gupta and Belnap as a response to the truth-theoretic paradoxes. In particular, Bruni focuses on the finite use of a technique, the revision rule, which was primarily meant as a way to provide a natural semantics for predicates expressing circular concepts, such as the truth predicate. Finite revision falls short of giving a semantics for the truth predicate, yet it finds natural applications in, for example, game theoretic settings, where players base their decisions on what is rational for other players to do. Besides these applications, Bruni also highlights interesting connections between the finite fragment of revision theory and the FS axiomatic theory of truth.

1.2.2.1 "Putting Davidson's Semantics to Work to Solve Frege's Paradox on Concept and Object" by Philippe de Rouilhan

In his contribution, Rouilhan introduces the reader to what, in the paper itself, he calls for short *Frege's paradox* and the *generalized Frege's paradox*. These paradoxes are not part of the family of Russell's well-known paradoxes that afflicted Frege's logical

[17] See e.g. Rouilhan (2012) and references therein for a broader philosophical perspective on logical universalism.

system, nor part of the family of Frege's puzzle about identity statements; neither are they paradoxes concerning truth. Rather, for once, and as the title suggests, the concept of truth is not part of the problem, but part of the solution.

So, just what is Frege's paradox? It first arose as a consequence of Frege's conception of a language of science. For Frege, as is well known, a predicative term refers to a concept, a singular term refers to an object, and concepts are not objects. But then, to explain the semantics of a language of science, Frege felt that he was inevitably led to say things like, e.g. *the concept horse is not an object*. For this very sentence to have a meaning, according to Frege's logical grammar, the expression *the concept horse* must itself refer to an object. But if *the concept horse* refers to an object, it fails to refer to the concept horse itself (concepts are not objects!)- and so would any expression of the appropriate logical category to serve as subject for the predicate *is not an object*. Thus, what is intended, Frege thought, cannot properly be said, thereby leading to a kind of ineffability thesis.

Rouilhan generalizes Frege's conception of the logical grammar of a language of science. He calls 'generalized Frege's paradox' that which arises whenever one devises a putative language of science such that, in order to explain what this language means, one must resort to another language whose logical grammar clashes with the logical grammar of the language in question. Rouilhan thinks that, despite the fact that Frege himself was prepared to live with the paradox, falling prey to it is really anathema to any putative language of science. So the question is: is it possible to escape the paradox; and if so, what types of language of science do escape it?

Rouilhan argues that the paradox can be escaped for a wide range of plausible candidate languages of science. His starting point is to take up Davidson's idea that explaining what sentences in a language mean is to give their truth-conditions in the form of a 'recursive theory of truth à la Tarski' for that language. One has solved the generalized Frege's paradox for a putative language of science, Rouilhan argues, if one is able to construct a recursive definition of truth for the language in another language that complies with the logical grammar of the language under consideration. Hence the question arises: is it always possible to do so? If the logical basis of the language of science is ZFC, Rouilhan recalls that one can devise a recursive definition of truth for such a language in an extension of the language that shares the same logical basis and only has further extra-logical primitive vocabulary. The meaning-explanation of the language of science is thus carried out in accordance with its logical grammar, without any category mistake with respect to it, or any involvement of entities that were not taken from the start to be part of its ontology. But what about candidates for the status of language of science whose logical basis is type theory or a part of it? Do they escape the generalized Frege's paradox? The answer is far less straightforward and Rouilhan's contribution here is to show that it is still possible to show that they do.

True, it can be shown that one cannot construct a recursive truth definition for a language of infinite order in another language that shares the same logical basis. But Rouilhan submits that there are independent reasons why an infinite order language is a dubious candidate for the status of language of science anyway. The question of interest, then, concerns languages of given finite order. To illustrate, consider a

monadic language, L, of finite order n—that is, with typed variables ranging over individuals or over classes of individuals or over classes of classes of individuals, etc., up to classes of order n, without overlap, and nothing more. Is it possible to construct a recursive definition of truth for L in an extension of it of the same order, obtained by adding, at most, a few extra-logical constants? Rouilhan proves that it is possible if $n \geq 4$. Thus, one can explain the meaning of L in a Davidsonian manner if $n \geq 4$ and, in the end, the generalized Frege's paradox proves to be no threat to adopting such a language as the language of science. Of course, as the author reminds us, recursively defining truth for any language in another language sharing the same logical basis is not the same as giving any definition of truth for the language in question in itself: dealing with this latter difficulty brings us back to the familiar paradoxes of truth, and falls outside the solution of the generalized Frege's paradox.

1.2.2.2 "Sets, Truth, and Recursion" by Reinhard Kahle

In his contribution, Kahle presents a theory of sets by means of an axiomatic theory of truth by defining sets and membership relations in terms of a truth predicate: namely, he identifies an object a being a member of a set $\{x \mid P(x)\}$ for a predicate P, with the predicate P being true of an object a. This idea of Kahle's is based on the notion of so-called *Frege structures*, and he has published a series of papers on theories of Frege structures over *applicative base theories* (Kahle 1999, 2001, 2003, 2009, 2011). The present paper provides an overview of his work and presents the philosophical foundation of his framework.

The concept of Frege structure was introduced by Aczel to "isolate the structure of that part of Frege's *Grundgesetze* that we consider to be correct" (Aczel 1980, p. 38) and he illustrated a semantic construction of a Frege structure over models of lambda calculus. Beeson (1985) gave the first axiomatic system F of a Frege structure. The idea of a Frege structure in connection with the theme of this volume could be summarized as follows.

(1) The full comprehension axiom should hold for all propositional functions: namely, every propositional function f forms a set $\{x \mid fx\}$ such that $a \in \{x \mid fx\}$ is a true proposition iff fa is a true proposition.
(2) Russell's paradox is caused by Frege's assumptions that (i) propositions are either true or false and the conception of the truth of propositions is the classical bivalent one, and that (ii) every formula (or well-formed expression) gives a propositional function.
(3) A Frege structure determines what objects are propositions (and true propositions) and thereby provides the definitions of sets and membership relations so that the full comprehension axiom in the form of (1) above consistently holds by abandoning Frege's two assumptions (i) and (ii).

In the literature, a Frege structure is usually formulated over combinatory algebras (models of combinatory logic) or λ-structures (models of λ-calculus) which are special kinds of applicative structures. Applicative structures are meant to deal with certain abstract conceptions of functions ("functions as rules" in Barendregt 1984 or "functions as operation processes" in Hindley and Seldin 2008), although it is debatable precisely what this means. Each member a of the domain D of an applicative structure can be "applied" to any member $b \in D$ and thereby yields an output $ab \in D$; note, however, that elements of D are not functions in the set-theoretic sense because they are, so to speak, universal functions that can be applied to everything; see Barendregt (1984, Chap. 1) and Hindley and Seldin (2008, Chap. 3E), for further discussion. So, a Frege structure counts propositions and propositional functions among the objects of the domains of applicative structures; this assumption might be justified by arguing that propositional functions are "functions" anyway and propositions are the values of propositional functions. Hence, the intended domain of a Frege structure contains the bearers of truth (*not* sentences *but* propositions in this setting) together with various other mathematical objects and functions. It is usually assumed that propositions are logically suitably structured so that these structures possess some distinguished syntactical operations such as $\dot{\neg}$, $\dot{\wedge}$, etc., whose intended interpretations are the functions that send propositions to their negation, their conjunction, etc.; a discussion concerning the assumption of a logical structure of propositions in the context of formal theory of truth can be found, e.g. in Halbach (2010, § 2).

In the usual presentation of a Frege structure, as offered by Aczel, Beeson, Cantini (whose monograph Cantini (1996) is also an important reference on Frege structures) and Kahle, a set $\{x \mid fx\}$ for a propositional function f is defined as $\lambda x.fx$ by means of λ-abstraction, which is available in combinatory algebra, and a proposition $a \in \{x \mid fx\}$ is simply defined as $(\lambda x.fx)a$. Consequently, since $a \in \{x \mid fx\}$ is just identical to fa by definition (or β-reduction), the full comprehension axiom in the form of (1) above is in fact a trivial and immediate consequence of the definitions of sets and the membership relation $\in$. Hence, in the customary setting, the construction of a Frege structure essentially comes down to the question of how to give a sensible characterization of propositions and truth.

In axiomatizing Aczel's semantic construction of a Frege structure, Kahle adopts an *applicative theory* TON as the base theory of his theory FON of Frege structure. Applicative theories are first-order theories for applicative structures and are usually assumed to include combinatory logic as their core component; for an exposition of combinatory logic, see Barendregt (1984) and Hindley and Seldin (2008), or see Cantini (1996) for its connection to axiomatic truth theories. Then he introduces a truth predicate T as a primitive predicate symbol, and expresses "x is a proposition" by $Tx \vee T\dot{\neg}x$. Thereby, for example, the full comprehension axiom in the form of (1) above can be expressed by:

$$\forall a(Tfa \vee T\dot{\neg}fa) \rightarrow \forall a[T(a \in \{x \mid fx\}) \leftrightarrow T(fa)], \qquad \text{(FCA)}$$

where $\forall a(Tfa \vee T\dot{\neg}fa)$ expresses "the value of f is always a proposition", i.e. "f is a propositional function". As mentioned earlier, FCA trivially obtains by definition,

and the essence of Kahle's axiomatization of Frege structure lies in the postulation of appropriate axioms for T so that it properly expresses truth. However, since Frege's two assumptions must be restricted to sustain consistency, the conception of truth in a Frege structure is inevitably non-classical. In fact, truth in Kahle's theory FON behaves in accordance with the non-classical strong Kleene logic; namely, the "inner logic" of FON is strong Kleene logic. As another example, the inner logic of the truth in Aczel's original semantic construction and Beeson's axiomatization F is non-classical Aczel-Feferman logic (see Fujimoto 2010, where it is called Feferman Logic).[18]

Lastly, let us explain a notable difference between Kahle's FON (as well as Eberhard and Strahm's theories introduced later on) and the more traditional type of axiomatic theories of truth such as Leigh's and Cieslinski's in this volume. In the traditional setting, a truth predicate is conceived of as a predicate of sentences (i.e. sentences are taken to be the bearers of truth), and syntactical objects such as sentences are considered to be distinct objects from those of the subject matter of theories of truth (such as natural numbers and sets). Accordingly, a base theory B such as PA of the traditional type of axiomatic theories of truth has to perform two totally different roles at the same time (i.e. that of a theory of the subject matter and that of a theory of syntactical objects *via a certain coding system* such as Gödel numbering) and two totally different types of objects (i.e. the mathematical objects of the subject matter and syntactical objects) are entangled in the domain of discourse of B. This entanglement causes, for example, the following problem of axiomatic schemata: when a truth predicate is newly introduced into an arithmetical base theory B, one might want to expand the arithmetical induction schema for the augmented language so as to enable arguments or proofs by induction on the syntactical complexity of sentences, on the one hand; but one might also want not to expand the induction schema for the augmented language so as not to make further mathematical commitments from a deflationist point of view, on the other hand. In other words, it might be the case that one wants to expand the schema in regarding B as a theory of syntax but also wants to restrict the schema in regarding B as a theory of mathematics. For more detailed discussion, see Leigh and Nicolai (2013). There are two opposite directions one can take to resolve this entanglement:

(a) to clearly separate the domains (or sorts) of the bearer of truth and the object of the subject matter;
(b) to choose a subject matter whose domain of discourse intrinsically contains both the bearers of truth and mathematical objects altogether.

The first direction is taken by Heck (2011) and Nicolai (2014) for example (theory of truth with disentangled syntax from object-language). Kahle's theories and Eberhard

[18] In contrast to Kahle's FON, Beeson introduces a predicate expressing "x is a proposition" independently as another primitive predicate, and his theory F is based on another type of applicative theory EON. These differences yield no difference in proof-theoretic strength, and Kahle's FON and Beeson's F are in fact proof-theoretically equivalent.

and Strahm's theories, which take applicative theories as their base theories, can be regarded as taking the second direction, since the objects of the subject matter and the bearers of truth coexist in the intended domain of applicative theories.

1.2.2.3 "Unfolding Feasible Arithmetic and Weak Truth" by Sebastian Eberhard and Thomas Strahm

Eberhard and Strahm's contribution extends Feferman's *unfolding programme* to feasible arithmetic, and is a continuation of their previous study of theories of truth for feasible arithmetic (see Eberhard and Strahm 2012; Eberhard 2013).

Axiomatic theories of truth are traditionally based on Peano Arithmetic PA (or its equivalents such as Cantini's OP (1996)), but one may adopt a different kind of base theory. One way to go is to enrich a base theory to, say, a set theory ZF for example; see Fujimoto (2012). Eberhard and Strahm go instead in the opposite direction and weaken the base theory to the so-called feasible arithmetic.

In complexity theory, a branch of theoretical computer science, effective decision procedures or algorithms are classified into a hierarchy of various complexity classes. Some effective algorithms are 'efficient enough' and can be 'feasibly computed', while others are 'too inefficient' and take an intractable amount of time to terminate. Feasibly computable algorithms are often identified with polynomial time algorithms.[19]

Peano Arithmetic PA is too strong a base theory for a theory of truth for feasible arithmetic. This is because the class of definable functions of PA properly includes that of primitive recursive functions, and not all primitive recursive functions are feasibly computable.[20] Over the last few decades, we have seen the development of a variety of formal arithmetic theories associated with the class of feasibly computable functions, in the sense that the class of functions that the theory can 'describe' (in terms of provable totality, provable convergence, definability, etc.) coincides with that of feasibly computable functions. The authors previously presented a theory T_{PT} of truth of feasible strength (Eberhard and Strahm 2012) where the provably total functions are precisely polynomial time computable ones. The present paper presents a new theory $\mathcal{U}_T(\mathsf{FEA})$ of truth (as well as two other proof-theoretically equivalent theories) of feasible strength but this time in the form of Feferman's unfolding.

The notion of *unfolding* was presented by Feferman (1996) as his most recent answer to the following problem raised by Kreisel: 'What principles of proof do we

[19] Roughly speaking, a polynomial time algorithm for a computational problem P is an algorithm such that it can reach the solution of P for any input of length n within $F(n)$ steps of computation for some fixed polynomial function F. Granted this identification of feasibility and polynomial time computability, the $P = NP$ problem, a famous Millennium Prize problem, questions whether or not a certain class of computational problems is feasibly solvable. There are many good textbooks on complexity theory; e.g. see Garey and Johnson (2002).

[20] As a matter of fact, such a base theory must be even weaker than $\mathsf{I}\Sigma_1$, since the class of provably recursive (and thus definable) functions is in this case exactly that of primitive recursive functions.

recognize as valid once we have understood ... certain given concepts?' (Kreisel 1970).[21] Proof-theoretic analysis of the unfoldings of finitist arithmetic and non-finitist arithmetic are already given by Feferman and Strahm in their (2000, 2010) (all of which are significantly stronger than $\mathcal{U}_T(\mathsf{FEA})$). In principle, unfolding is applied to schematic theories which contain schematic axioms expressed in terms of free predicate variables P, Q, ... of each arity; for example, the induction schema is expressed by a single formula:

$$P(0,\vec{v}) \wedge (P(x,\vec{v}) \rightarrow P(x+1,\vec{v})) \rightarrow \forall x\, P(x,\vec{v})$$

by means of a predicate variable P in a schematic theory. The idea behind unfolding is as follows. An initial schematic theory S comes with basic operations and predicates of a subject matter from which we start the process of unfolding. We go on to define and introduce more and more operations and predicates to the initial schematic theory following the rules governed by a certain background theory of operation-forming and predicate-forming. Then, application of schematic axioms is expanded to those newly introduced operations and predicates by means of the Substitution Rule, which allows us to substitute anything (possibly containing new operations and predicates) for the predicate variables P, Q, ... in the schematic axioms of S. In general, unfolding systems comprise: (1) a schematic base theory, which determines the subject matter and universe of discourse of one's investigation; (2) a theory of operation—and predicate-formation, which determines what new operations and predicates we can construct and how they are constructed from the basic ones of S; and (3) a substitution rule, which enables us to apply the schematic axioms of S to newly constructed operations and predicates.

Put this way, it could be said that the essence of unfolding systems lies in the choice of the second component, i.e. its theory of operation—and predicate-formation. Feferman introduced two different types of unfolding: *operational* and *full* unfoldings, which differ in this second component. The former type only allows the introduction of new operations (over individuals); while the latter allows the introduction of both operations and predicates. According to Feferman and Strahm, '[w]hereas [the operational unfolding of S] addresses the question of which operations on $\mathbb{A}$ ought to be accepted given a schematic system S for a structure $\mathcal{A} = (\mathbb{A}, F_0, \ldots, F_n)$, the central question concerning [the full unfolding of S] can be stated as follows: which operations on and to predicates—and which principles concerning them—ought to be accepted if one accepts S?' (Feferman and Strahm 2000, p. 80).

In Feferman and Strahm's formulation in their (2000, 2010), full unfolding systems contain *terms* for 'predicates' for each arity and a binary predicate symbol $\in$ for the membership relation. For an n-ary predicate term X and n-tuple $(i_1, \ldots, i_n)$ of individuals, the formula $(i_1, \ldots, i_n) \in X$ means that $i_1, \ldots, i_n$ fall under the extension of the predicate (expressed by) X. A full unfolding system can thereby treat predicates as terms, and more and more predicates (as terms) are produced by manipulating or

[21] It might be worth noting here that perhaps the most famous axiomatic theory of truth KF was originally presented by Feferman (1991) as an answer to this question of Kreisel's.

combining those terms. Now, one can find here an intimate connection between the treatment of predicates in full unfolding systems and that in truth theories: to say that objects fall under the extension of a predicate is essentially to say that the predicate is true of the objects; cf., Kahle's contribution to this volume. In fact, in his 'pilot study' paper (Feferman 1996), Feferman originally formulated full unfoldings in terms of a truth predicate instead of the membership relation. Following Eberhard and Strahm, let us call this version of full unfolding *truth unfolding*; with respect to finitist and non-finitist arithmetic, truth unfolding and full unfolding are equivalent.[22]

In the present paper, Eberhard and Strahm present the truth unfolding, full unfolding, and operational unfolding of feasible arithmetic, and show that all three systems have the same feasible strength.[23] Through their research, Eberhard and Strahm open up a new subject of study: theories of truth for feasible arithmetic, and they provide a new perspective on the unfolding programme from the point of view of theories of truth.

1.2.2.4 "Some Remarks on the Finite Theory of Revision" by Riccardo Bruni

Paradoxes in general, as Quine noted in his popular (1976), have often stimulated reflection in new directions and have given rise to fruitful new methods and concepts that have been applied to other subjects. Paradoxes of truth are no exception here, and Bruni's essay can be seen as yet another application of the revision rule inspired by the revision theory of truth. The revision theory of truth, invented by both Gupta and Herzberger independently, and developed by Gupta and Belnap (1993), is one of the main contenders in the search for a solution to the Liar paradox. Revision theory identifies the source of the paradox as residing in the fact that the truth predicate admits a circular definition and pathological sentences are to be expected in the presence of circular concepts. Using an example from Gupta and Belnap (1993), suppose we define a predicate Gx as $x = \text{Socrates} \vee (x = \text{Plato} \wedge \neg Gx)$. From a classical perspective, no extension can be assigned to G, since to determine an extension we need to determine which elements of the domain satisfy the *definiens*, and, since G itself occurs in the *definiens*, we already need the extension of G in order to do that. However, the key intuition behind revision theory is that the circular definition gives us *hypothetical* information about the extension of G, and that this

[22] The full unfolding system in the formulation in Feferman and Strahm (2000, 2010) has one more important facility: the disjoin union operator *Join*, for predicate-formation. Without the *Join* operator, the resulting unfolding does not reach the strength of truth unfolding (and full unfolding) with respect to non-finitist arithmetic. In contrast, as Eberhard and Strahm show in the present paper, *Join* yields no difference in proof-theoretic strength with respect to feasible arithmetic (this is also the case with respect to non-finitist arithmetic).

[23] Eberhard and Strahm's theories are based on a certain type of applicative theory, just as Kahle's theories are; see the previous section introducing Kahle's paper for a discussion of the philosophical import of the applicative setting in comparison to the traditional setting. In general, applicative base theories are a natural set-up for pursuing the unfolding programme, because they provide us with a versatile and natural framework for term application and thus for operation- and predicate-forming.

information can be used to provide a rich theory of the content of G. The circular definition tells us at least this: assuming by hypothesis that Plato is not G, then both Socrates and Plato would satisfy the definiens of G and no other object would; however, assuming that Plato is G, then only Socrates would satisfy the definiens of G. Revision theorists keep track of this information as a rule of revision: if, by hypothesis, the extension of G is {Plato}, then it should be revised to {Socrates}; if it is {Aristotle}, then it should be revised to {Socrates, Plato}, etc. We see that Plato behaves, with respect to this definition, as the Liar does with respect to truth: if it is G, it should be not G and, if it is not G, it should be G.

The second key component of revision theory is the recipe for extracting *categorical* information from the revision rule associated with a circular definition. This is achieved by iterating the process of revision and paying attention to the sentences that have a fixed truth value when the process advances, no matter what the initial hypothesis. In our simple example, it is clear that, for any hypothesis, after the first revision, Socrates is classified as G, all other objects except Plato are classified as not G, and Plato is pathological. Hence, we could categorically assert 'Socrates is G' and 'Aristotle is not G', and we should refrain from asserting either 'Plato is G' or 'Plato is not G'. In general, if we have a language L with interpretation M and domain $|M|$, an expanded language L^+ is obtained from L by adding a new predicate G and a definition $Gx =_{\text{def.}} A_G(x, G)$, where $A_G(x, G)$ is a sentence in L^+. We then define the revision rule δ_A as follows: for any hypothesis H (i.e. for any subset of $|M|$), $\delta_A(H) = \{a \in |M| : (M, H) \models A_G(\overline{a}, G)\}$, where $\overline{a}$ is a constant that names a and (M, H) is the interpretation of L^+ obtained from M by adding H as the interpretation of G. Once we have a revision rule, a revision evaluation sequence is defined as $\delta_A^0(H) = H$, $\delta_A^{n+1}(H) = \delta_A(\delta_A^n(H))$. The revision process can be projected into the transfinite ordinals, defining a limit rule. There are several options available in the literature to do this. The key notion in extracting categorical content from a circular definition is the notion of a reflexive hypothesis. A hypothesis H is n-reflexive if $\delta_A^n(H) = H$, and it is reflexive if it is n-reflexive, for some $n > 0$. Once the revision process arrives at a reflexive hypothesis, all the subsequent iterations of the revision rule also produce reflexive hypotheses and form a cycle that repeats itself indefinitely. This is why Gupta and Belnap consider reflexive hypotheses the best candidates to determine the extension of G, and define a sentence B of L^+ as valid in M if it is true in all interpretations (M, H), where H is a reflexive hypothesis. B is valid if it is valid in all interpretations of L. In the general case, the notion of reflexivity has to be extended to a transfinite ordinal. (Gupta and Belnap 1993 consider other notions of validity.)

Bruni's paper analyses the class of finite definitions, a special class of circular definitions that satisfy the condition that for all interpretations M, there is a number k such that, for every hypothesis H, $\delta_A^k(H)$ is reflexive. Finite definitions are those that guarantee arriving at a reflexive hypothesis in a finite number of steps of revision for any hypothesis. The main source of examples of finite definitions is game theory, where the rational action for any player depends on what it is rational for the other players to do.

Bruni's paper evaluates finite definition semantics and compares it with standard (transfinite) revision semantics, highlighting three aspects. Firstly, finite revision semantics has less complexity than standard revision semantics. As an example of this, Bruni proves that every definable set on a circular predicate in the standard interpretation of arithmetic is at most $\prod^1_1$ in finite revision semantics; while (as proved by P. Welch 2003) it is at least Δ^1_2 in the transfinite case. Secondly, finite revision semantics has a sound and complete natural deduction calculus (due to Gupta and Belnap 1993) and Bruni presents an equivalent Hilbert calculus; while no complete calculus can be given for standard revision semantics. Thirdly, these calculi are not only technically interesting, but are also very natural, since they reflect the ordinary arguments one would make when reasoning with circular definitions.

Even though the scope of finite revision theory does not include the truth predicate, in the last part of the paper Bruni develops earlier work by Halbach (1994) that establishes a connection between truth as formalized in FS, and validity as codified in finite revision semantics. Bruni presents a syntactic version of Halbach's connection, showing that (when working in standard arithmetic) derivations from FS can be mimicked in a variation of the theory FS that uses indexed formulae, where the indices represent the stages in the revision evaluation sequence. These results raise the question of whether similar connections can be established in transfinite revision semantics.

1.2.3 Truth as a Substantial Notion

The first two papers in this chapter address common objections against substantial notions of truth and thereby pave the way for inflationary, as opposed to deflationary, theories of truth.

Sher addresses one of the main difficulties that correspondence theories traditionally face, which is none other than the need for a precise construal of the correspondence relation in a way that does not restrict the notion of truth to a single domain of discourse. In order to meet this requirement, Sher proposes what she calls 'composite' as opposed to 'direct' correspondence. She illustrates what this composite relation comes down to in the philosophy of mathematics; a most challenging domain for the correspondence theorist.

Glanzberg adopts an inflationary contextualist approach and argues that hierarchical accounts are necessitated by a reflection process that is meant to render explicit what is involved in our implicit grasp of the notion of truth. Alongside this general motivation, Glanzberg addresses specific arguments that have been put forward against hierarchical approaches to truth. Among these is what he calls the 'one concept' objection, which threatens the unity of the notion of truth; albeit in a very different way from the objection against correspondence theories of truth that Sher addresses in her paper.

Finally, rather than discussing specific arguments against substantial theories of truth, Engel shifts the burden of proof onto the other side by questioning the plausibility of construing the notion of truth as a non-substantial notion. Engel underlines the irreducibly normative role of truth in determining standards of correctness for belief and assertion. He subsequently identifies a tension for the deflationist who will not be able to account for such standards of correctness without admitting the non-deflationary normative content of truth.

1.2.3.1 "Truth as Composite Correspondence" by Gila Sher

A problem for traditional correspondence theories of truth is that it is hard to see how any precise expression of the correspondence relation between discourse and the world could ever account for the property of truth in its multifarious applications, from physical discourse to ethics and mathematics. Arguably, if a direct correspondence account is available for the truth of discourse about the physical world—with terms referring to physically identified objects and the truth of sentences built on reference as Tarski taught us (see also Field 1972)—such an account will give rise to insuperable difficulties when it comes to domains such as ethics and mathematics. It is not surprising then that, especially in mathematics, many philosophers have been tempted to give up on the idea that the truth of discourse amounts to correspondence with mathematical facts. Correspondence theorists thus face a problem. If they stick to the thesis that truth is correspondence, they will have a hard time providing a plausible account of correspondence that applies uniformly to the various realms of discourse. If, however, they admit that truth is correspondence in some domain, and not in others, then they compromise the unity of the notion of truth. So, either there need to be different meanings of the word "true", or "true" cannot be applied to the various domains of discourse to which we ordinarily, and unproblematically—or so it seems—apply it.

In her "Truth as Composite Correspondence", Gila Sher takes up the correspondence challenge in an attempt to overcome the above predicament. She does so in two ways. Firstly, by articulating the main lines of a renewed methodological programme for the development of a substantive theory of truth. The challenge here is twofold: (a) to alleviate what the author argues is the unjustified burden placed by a foundationalist stance on the possibility of developing any substantive account of truth; (b) to make room for an alternative construal of correspondence through a long-term holistic inquiry that would be faithful to the specifics of the various ways we access the world in different domains of discourse. Secondly, Sher sets her programme to work in the case of mathematical discourse. The author's main thesis is that one can conceive of mathematical discourse as being about facts of the world. But what kind of facts? Building on her earlier work[24], Sher argues that these facts consist of the world having some formal properties, where formal properties are in

[24] See Sher (1991).

turn explained in terms of invariance under some classes of transformations. In a nutshell, the invariance idea is here a generalization (a cross-domain generalization) of the idea put forward by Tarski (1986): just as the property of being red, say, can be seen as the property which is invariant under those transformations of the domain of the universe that leave red things red, so the formal properties, such as the (second-order) property of having cardinality 3, are properties that are invariant under *any* permutation of the domain. That much having been said, it remains problematic that mathematical discourse is, on the face of it, about individuals, namely numbers, and not about second-order properties. This is where composite correspondence comes in. First-order statements about numbers can, Sher argues, be said to correspond to facts involving second-order properties, via e.g. posits, if one allows for composite correspondence. And in fact, there are reasons for humans to have adopted a standard of composite correspondence rather than direct correspondence as a substantive norm for their discourse; namely, simplicity or cognitive tractability. There is thus some evidence that a standard of truth understood as composite correspondence is a plausible one. In the remainder of her paper, Sher goes on to sketch reasons for the fruitfulness of her approach in solving various puzzles in the philosophy of mathematics.

1.2.3.2 "Complexity and Hierarchy in Truth Predicates" by Michael Glanzberg

After Kripke's famous attack on the Tarskian hierarchy of languages, hierarchies have been viewed with suspicion. Kripke (1975) argued that a Tarskian hierarchy of truth predicates cannot be a good formalization of ordinary language, because it leaves out perfectly natural self-referential sentences. Even in theories with an untyped truth predicate—that is, theories especially designed to overcome these expressive problems (such as Kripke's own theory)—the eventual reappearance of hierarchies due to the revenge paradoxes is usually taken to be a defect that future research should overcome. In his "Complexity and Hierarchy in Truth Predicates" Glanzberg tries to dispel these worries by offering a sustained defence of hierarchies, and showing where and why they should be expected to appear.

The paper argues that inflationary theories of truth motivate the use of hierarchies, while deflationist theories do not. The key element that generates this difference is that while for deflationists truth is a simple property that is fully characterized by the transparency of truth (the intersubstitutability of any sentence A and $T(\ulcorner A \urcorner)$ in all non-opaque contexts), for inflationary theories truth is a potentially complex semantic property with internal structure. The complexity of truth originates in the different mechanisms involved in the determination of the truth value of sentences: semantic composition and facts about reference and satisfaction. Even if truth is a complex property, ordinary speakers have an implicit grasp of the truth predicate and Glanzberg argues that philosophers can make this implicit knowledge explicit by a process of reflection on our own abilities. The main claim of the paper is that this activity of reflection generates hierarchies.

The central part of the paper provides an overview of several processes of reflection, showing that all of them reveal the complexity of truth and some hierarchy is generated. The first case starts with the language of arithmetic, which one can reflect upon either model-theoretically (if one understands the language as interpreted in a model of arithmetic and then gives a Tarskian definition of truth) or proof-theoretically (if one starts with a theory such as PA and gives the compositional axiomatic truth theory CT). Complexity measures show that the complexity of the truth predicate obtained using a Tarskian definition of truth is greater than the complexity of the base theory. The second case is a language that contains its own truth predicate. From a model-theoretic perspective, Glanzberg here summarizes Kripke's construction of a fixed-point semantics as a process of approximation (what he calls the long iteration strategy for reflection), which shows up in almost all theories designed to provide a solution to the Liar paradox; such as revision, paraconsistent or paracomplete theories. The long iteration strategy is not able to produce a perfect theory of truth. Taking Kripke's construction as an example, when we get to a fixed-point interpretation for the truth predicate we see that the Liar sentence is neither in the extension nor the antiextension of the truth predicate. But then the Liar sentence is not true and we are back to a paradox. This means that the process of reflection is incomplete and has to start again, creating an open-ended hierarchy of new truth predicates. Glanzberg briefly presents his own contextualist solution to the Liar paradox as an example of this process of the generation of hierarchies and points to some recent results on the iteration of axiomatic theories of truth through suitable proof-theoretic ordinals. The fact that a completely successful theory of truth is not found is to be expected, given the extreme complexity of truth.

The last part of the paper defends the hierarchical approach from some common objections: the *one concept* objection (we only have one concept of truth, not many concepts generated at the different levels in the hierarchy), the *clumsiness* objection (some hierarchical theories cannot express some ordinary self-referential sentences) and the *weakness* objection (hierarchical theories of truth are very weak when it comes to mathematical purposes). The discussion of the one concept objection is of special interest. Here Glanzberg introduces the notion of stratification: "A concept is stratified if we cannot provide a single theory or definition for it. Instead, we provide a family of related theories or definitions, each of which is systematically connected to others. In effect, a concept is stratified if when we try to analyze it, we wind up with a hierarchy" (p. 211). Glanzberg compares the case of truth to other notions which are also stratified (such as the notion of mathematical proof), and finds that the one concept objection applies differently to different types of truth hierarchies. Finally, Glanzberg discusses how his own hierarchy stands with respect to the one concept objection. The general conclusion of the paper is that inflationary accounts of truth motivate the construction of hierarchical theories of truth; theories that are more natural than non-hierarchical ones.

1.2.3.3 "Can Deflationism Account for the Norm of Truth?" by Pascal Engel

Many influential deflationists, such as Quine and Field[25] have endorsed their conception of truth as part of a more general physicalist, or naturalist, philosophical framework. What these authors are primarily interested in is whether there is a place for a notion of truth in the language of natural science that would be suitable for the description or causal explanation of natural phenomena. Normative facts are not part of the picture—they are simply not natural phenomena. Science identifies a phenomenon in natural terms and seeks causal explanations; it is indifferent to the explanation of non-natural facts and of behaviour in terms of underlying reasons and norms. Against this backdrop, deflationists, with notable exceptions[26] however, have not paid much attention to the normative role of truth and the problem this role may pose to a general deflationary approach.[27] For, in holding that all there is to our understanding of the notion of truth is our understanding of the assertability of all of the T-sentences, the deflationist implies, among other things, that truth is not a distinctively normative property. It may be true that the notion of truth is not essentially a normative notion. By ascribing truth to the proposition that snow is white we do not thereby make a normative claim—as we would when ascribing, for example, goodness—but the notion of truth is still deeply involved, or so it seems, in explaining the norms that govern some of our actions and thoughts.

In particular, it seems to be part of our understanding of the norms that govern correct assertion and belief that they are correct only if true. Consequently, the notion of truth, even if not in itself normative, is in fact needed in order to account for the norm of correctness governing belief and assertion. This would perhaps not do much harm to the deflationist if it could be shown that our recourse to truth in this case is not to do with setting a distinctive and irreducible norm, and that truth is involved only in a shallow way; for instance, as a mere logical device. In his essay, Engel shows how uncomfortable the deflationist's position is here. As Engel recalls, there are many arguments showing that the standard of truth for assertion (and belief) is distinct from other standards, such as subjective standards (it is correct to assert that p if one believes that p) or epistemological standards, such as warranted assertability. At the same time, difficulties are lurking nearby if one chooses to renounce such a correctness norm. Engel sets out to evaluate the various deflationist strategies which downplay the normative role of the notion of truth and have been proposed in response to these challenges. His conclusion is that the following dilemma is robust: either deflationists have to eliminate the normative features of truth, but then they are unable to account for what constitutes the correctness of belief and assertion; or they grant that truth is involved in the normative account of belief and assertion, but then they are unable to account for the distinctive substantial norms intrinsically associated with truth.

---

[25] See e.g. Quine (1970, 1990) and Field (2001).
[26] See e.g. Horwich (2006). On a larger scale, see also Robert Brandom's work (Brandom 1994).
[27] One such problem was already pointed out early on by Dummett (Dummett 1959).

1.2.4 Deflationism and Conservativity

Halbach and Horsten are concerned with the general norms that a theory of truth should adhere to. Specifying such norms obviously depends on the notion of truth that one endorses, and for Halbach and Horsten those norms are motivated by their deflationary approach to truth. While Engel (see previous section) is interested in the way truth relates to other notions—namely, assertion and belief—Halbach and Horsten assume a perspective that is internal to a particular notion of truth, that is, reflexive truth, and propose a short list of general desiderata that axiomatic theories aimed at accounting for reflexive truth must satisfy. Their viewpoint is normative but also descriptive in that they are interested in describing what drives current work on truth theories. They do not propose their list in order to single out one theory as the best theory currently available; but they do hope that the norms will make it possible to compare theories and shed light on the choices made.

Deflationism famously advocates the non-substantial character of truth. One way to explicate the 'non-substantiality' thesis has been to take it to mean that a theory of deflationary truth should be conservative over its base theory, which in this context is commonly taken to be Peano Arithmetic (PA). A theory of truth is conservative over its base theory if it proves no theorems in the language of the base theory which are not already provable in the base theory alone; otherwise, it can be argued that the truth theory adds substantial content to that of the base theory. Conservativity as a criterion for deflationism was explicitly proposed by Horsten (1995), Ketland (1999) and Shapiro (1998) and is reminiscent of a long history of uses of this notion in the foundations of mathematics.[28]

Leigh's paper 'Some Weak Theories of Truth' follows this line of thought and examines which out of twelve principles of truth considered by Friedman and Sheard (1987) is to be blamed for non-conservativity over (PA). The truth theories that Friedman and Sheard consider are maximally consistent sets of these twelve principles modulo a theory $\mathsf{Base}_{\mathrm{T}}$. All of these theories, except one, are non-conservative extensions of (PA). Leigh observes that one of the principles, namely, U-Inf (the predicate version of the Barcan formula), is present in all of them and that in its absence, conservativity is restored. It follows that this is the principle responsible for the non-deflationary syntactic behaviour of the truth theories.

Fischer's paper 'Deflationism and Instrumentalism' changes the rules of the game somewhat by proposing a new understanding of deflationism and with it, a new way of assessing the conservativity requirement. Fischer construes deflationism as a form of instrumentalism in Hilbert's spirit[29]: carrying out the programme of instrumental deflationism means showing that it is possible to combine conservativity (i.e. truth-theoretic innocence) with instrumental utility (given by the expressive power of speed-up results). Fischer uses a weak theory of truth, PT^{-}, to show that this is possible. On top of this novel endorsement of the conservativity requirement,

[28] See e.g. Hilbert (1926), Field (1980) and Shapiro (1983).
[29] This understanding also underlies Ketland (1999).

Fischer offers additional support by addressing one of the main objections against conservativity according to which the truth theory should prove the soundness of its base theory, here PA.

Both Leigh's and Fischer's papers consider conservativity in its proof-theoretic sense, that is, they are concerned with the deductive power of the truth theory. In his 'Typed and Untyped Disquotational Truth', however, Cieśliński draws our attention to *semantic* conservativity and its relation to proof-theoretic conservativity for disquotational theories of truth, which are deflationary theories par excellence. It has been argued that semantic conservativity fits the deflationary spirit better (McGee 2006) as it evidences the lack of extra metaphysical commitments and for this reason it can be seen as an expression of the demand for an innocent non-substantial notion of truth. Cieśliński shows that conservativity in this sense is difficult to obtain for disquotational theories of truth. In preparation for this argument, Cieśliński offers an overview of disquotationalism and the problems inherent in devising a disquotational theory of truth, which have to do with deciding on adequate criteria for selecting a consistent subset of T-biconditionals, given that the full T-schema cannot be upheld.

Reading Enayat and Visser's paper 'New Construction of Satisfaction Classes' reminds us that the question of conservativity is also pressing for typed, and not just for untyped, theories of truth. The paper takes us back to Tarski's definition of truth in a model, for which the notion of satisfaction was famously introduced. As an axiomatization of this, Enayat and Visser propose $\mathsf{PA}^{\mathsf{FS}}$ as the base theory: a theory with a satisfaction predicate added to PA. It is reasonable to expect an axiomatic theory of Tarskian satisfaction to be conservative over its base theory and indeed the deflationist would be in an awkward position if conservativity fails in this case (although note that Enayat and Visser themselves do not draw any connections between their work and deflationism). However, the model-theoretic proof of conservativity (via the completeness theorem) with PA as the base theory is not easy, roughly because not every model of PA can be expanded to a model of $\mathsf{PA}^{\mathsf{FS}}$. Enayat and Visser offer a new, more simplified proof which is a clear improvement on previous results in this area.

1.2.4.1 "Norms for Theories of Reflexive Truth" by Volker Halbach and Leon Horsten

Given the proliferation of axiomatic truth theories in the last two decades, the question naturally arises of how one should adjudicate between them. Earlier attempts to answer this question include Sheard (1994) and Leitgeb (2005). The present paper extends that line of research by proposing a list of norms for axiomatic theories of the reflexive use of truth (or 'type-free' truth), as opposed to other uses, e.g. the use of truth in natural language. It is, of course, because there is no obvious way to circumvent the Liar paradox that there are many axiomatic truth theories suggested in the literature.[30]

[30] See also Halbach (2010) for an exposition of axiomatic truth theories.

The authors first offer some methodological remarks concerning the project they undertake. They doubt that a single property or cause can be identified from which all truth norms can be derived. However, they do not see the various desiderata as independent of each other, as they believe that satisfying truth-norms is not an all-or-nothing affair and that norms can be satisfied to a lesser or greater degree by different theories. This is where the authors distance themselves from the approach of Leitgeb (while mostly agreeing on the norms themselves) who considers truth theories as maximally consistent subsets of the norms that he lists.

Halbach and Horsten intend their norms to be underdetermined but also their list to be exhaustive in the sense that any desirable feature of an axiomatic theory of truth is somehow derived from the list. For example, in comparing their list to Sheard's, the authors take the necessitation rule to be a special case of the disquotational requirement, and the inference rule $\phi \rightarrow \psi \vdash T^{\ulcorner}\phi^{\urcorner} \rightarrow T^{\ulcorner}\psi^{\urcorner}$ as being derived from the compositionality norm and the identity between inner and outer logic, which also follows from the disquotational requirement. Since their list is very short, most of the work goes into determining which features of theories of truth are derived from which desiderata. An advantage of this approach is that it is open to novel ways (properties of the truth theory) of satisfying the desiderata. However, choosing what is fundamental and what is derivative is directly influenced by one's philosophical views. Under the deflationist approach that the authors adopt, the identity between inner and outer logic follows from the disquotational requirement and need not be listed as a separate norm; as it is, for example, by Leitgeb.

These, briefly are the norms given by Halbach and Horsten:

1. *Coherence*. A truth theory may be incoherent in relation to its base theory, if, for example, it contradicts theorems of the latter, or it is ω-inconsistent, or the induction schema cannot be extended to the language with the truth predicate. The truth theory may also be incoherent in its truth-theoretic part, if, for example, it proves $T^{\ulcorner}\phi^{\urcorner}$ for all ϕ.
2. *Disquotation and Ascent*. This norm stems from the deflationist's conception of truth as a disquotational devise or a devise for performing semantic ascent, and it entails that sentences ϕ and $T^{\ulcorner}\phi^{\urcorner}$ are in some sense equivalent. Ways to make this equivalence precise are (i) through the idea of transparency (Field 2008); meaning that ϕ and $T\phi$ are intersubstitutable without cost, or (ii) using the T-biconditionals; that is, material equivalences of the form $T(^{\ulcorner}\phi^{\urcorner}) \leftrightarrow \phi$. Both these ways lead to paradox under very weak assumptions, which means that the disquotational requirement cannot be met in full. The question then is how to weaken the disquotational requirement while still obtaining a non-trivial theory of truth. To restrict the T-schema, a straightforward proposal is to admit as many of its instances as possible. However, McGee (1992) shows that one cannot single out a unique maximally *consistent* set, and there is currently no uniform principle available to select a unique maximal set of admissible instances of the T-schema. To restrict the T-biconditionals, an obvious choice is to offer an alternative for the material conditional $\leftrightarrow$. Within classical logic the T-biconditionals may be consistently replaced by inference rules; that is, the truth theory may be closed under Necessitation $S \vdash \phi \Rightarrow S \vdash T(^{\ulcorner}\phi^{\urcorner})$ and Co-necessitation $S \vdash T(^{\ulcorner}\phi^{\urcorner}) \Rightarrow S \vdash \phi$. The possibilities proliferate in the case of

non-classical logics.
3. *Compositionality*. This norm requires that the T-predicate commutes with the connectives and the quantifiers (at least for vagueness-free fragments of the language). In itself commutation is not sufficient; one also needs the T-biconditionals for at least the atomic sentences in the ground language in order to get compositionality for the T-free fragment. Yet, it may not be possible to achieve full compositionality as it may clash with other norms. Motivation for restricting compositionality may then come from taking truth to be a partial concept; for example, one that does not apply to meaningless sentences such as the Liar paradox. One then naturally wants to reject commutation of the T-predicate with negation. Otherwise, negating the truth of the Liar sentence would mean that the negation of it is true, while this negation should be as meaningless as the Liar sentence itself. This is how things go with theories such as KF or Burgess' theory (Burgess 2009), where the restriction to positive compositionality gives rise to a grounded notion of truth. So groundedness is in some cases a derivative from the more general requirement of compositionality.
4. *Sustaining ordinary reasoning*. Feferman (1984) famously stated that a truth theory should sustain ordinary reasoning. There are many ways to interpret this norm; it rules out, for example, logics that do not allow the truth predicate in the induction scheme.
5. *A philosophical account*. This is a meta-norm that demands that the proposed norms are philosophically justified; since no truth theory can satisfy all norms in full, a philosophical story is needed to justify why certain axioms have been chosen and not others. Kripke's construction of the extension of the truth predicate as a learning process can be seen as one such philosophical story which justifies grounded truth. The story may not apply to other theories, and philosophical justification should be compatible with 'truth-theoretic pluralism' for the uses of the truth predicate in different contexts. The idea of truth-theoretic pluralism for axiomatic theories is already found in Sheard (1994) under the name of 'local truth analysis'. Finally, Halbach and Horsten add that, although desirable, this last norm should not hinder research on truth theories that is not motivated by the aim of providing a philosophical account for a particular use of the truth predicate.

1.2.4.2 "Some Weak Theories of Truth" by Graham E. Leigh

The question of whether a theory of truth is conservative or not over a base theory has acquired philosophical significance in the discussion of the nature of truth and the debate over deflationary conceptions of truth. In this connection, Leigh's essay on *Some Weak Theories of Truth* helps to circumscribe the theoretical possibilities that are open to the conservative deflationist, as it determines the nine *maximally conservative* sets of a collection of twelve principles of truth modulo a fixed theory $\mathsf{Base}_{\mathrm{T}}$. His research is motivated by the question: 'What assumptions about the nature of truth are responsible for deciding the proof-theoretic strength of a theory of truth?'. This motivation leads him to extend Friedman and Sheard's programme (Friedman and Sheard 1987).

Friedman and Sheard (1987) list twelve principles of truth, each of which is quite natural and plausible, but which taken together are inconsistent, and determine the nine *maximally consistent* combinations of them over a fixed theory $\mathsf{Base}_{\mathrm{T}}$ of truth including PA. The theory $\mathsf{Base}_{\mathrm{T}}$ is formulated over the language $\mathcal{L}_{\mathrm{T}}$ (the language of arithmetic plus a truth predicate T), and its axioms consist of those of PA with the expanded arithmetical induction schema for $\mathcal{L}_{\mathrm{T}}$ and the following three truth-theoretic axioms[31]:

(i) $\quad \forall\phi\forall\psi[T(\phi \to \psi) \to (T\phi \to T\psi)]$;

(ii) $\quad \forall\phi[(\phi \text{ is an axiom of } \mathsf{PRA}) \to T\phi]$;

(iii) $\quad \forall\phi[(\phi \text{ is a logically valid } \mathcal{L}_{\mathrm{T}}\text{-formula}) \to (\text{the universal closure of } \phi \text{ is true})]$.

Besides these basic axioms, Friedman and Sheard list twelve further truth-theoretic principles; see Table 1 in Leigh's paper. Then, they specify the following nine maximally consistent combinations $\mathcal{A}$–$\mathcal{I}$ of the twelve principles modulo $\mathsf{Base}_{\mathrm{T}}$:

$\mathcal{A}$. In, Intro, ¬Elim, Del, Rep, Comp, E-Inf, U-Inf.
$\mathcal{B}$. Rep, Comp, Cons, E-Inf, U-Inf.
$\mathcal{C}$. Del, Comp, Cons, E-Inf, U-Inf.
$\mathcal{D}$. Intro, Elim, ¬Intro, ¬Elim, Comp, Cons, E-Inf, U-Inf.
$\mathcal{E}$. Intro, Elim, ¬Intro, Del, Cons, U-Inf.
$\mathcal{F}$. Intro, Elim, ¬Elim, Del, U-Inf.
$\mathcal{G}$. Intro, Elim, ¬Elim, Rep, U-Inf.
$\mathcal{H}$. Out, Elim, ¬Intro, Del, Rep, Cons, U-Inf.
$\mathcal{I}$. Elim, ¬Elim, Del, U-Inf.

For each $\mathcal{X}$ among $\mathcal{A}$–$\mathcal{I}$, $\mathcal{X} + \mathsf{Base}_{\mathrm{T}}$ is consistent but adding any more of the twelve principles to $\mathcal{X} + \mathsf{Base}_{\mathrm{T}}$ results in an inconsistent theory, and every consistent (over $\mathsf{Base}_{\mathrm{T}}$) set of the 12 principles is included in some of $\mathcal{A}$–$\mathcal{I}$. In what follows, we identify each $\mathcal{X}$ and its induced theory $\mathcal{X} + \mathsf{Base}_{\mathrm{T}}$ for simplicity. The proof-theoretic analysis of these systems is given by Cantini (1990), Halbach (1994), and Leigh and Rathjen (2010), and their proof-theoretic ordinals are all known. Leigh and Rathjen also studied the Friedman-Sheard programme in intuitionistic logic; see Leigh and Rathjen (2012) and Leigh (2013).

Conservative theories of truth have a special status in deflationism. Among Friedman and Sheard's nine maximally consistent theories, only $\mathcal{A}$ is conservative over PA, and all the others go beyond PA; hence, if one accepts the conservativity requirement, $\mathcal{B}$–$\mathcal{I}$ are not deflationist theories of truth. So, which principle is responsible for the non-deflationary deductive power of these systems?

First, it is noted that the nine maximally consistent theories $\mathcal{A}$–$\mathcal{I}$ all contain the principle U-Inf. Also, E-Inf always comes together with Comp. Now, we can easily

[31] It is debatable whether $\mathsf{Base}_{\mathrm{T}}$ is a necessary basic part of theories of truth or even whether it is acceptable or not; some influential and popular axiomatic theories of truth such as KF (Feferman 1991) and DT (Feferman 2008) are inconsistent with the three axioms (i)–(iii) of $\mathsf{Base}_{\mathrm{T}}$.

verify that U-Inf implies E-Inf over $\mathsf{Base}_{\mathrm{T}}+\mathsf{Comp}$. Consequently, the following list gives the combinations of the principles that induce maximally consistent theories over $\mathsf{Base}_{\mathrm{T}}$ + U-Inf (rather than over $\mathsf{Base}_{\mathrm{T}}$):

A. In, Intro, ¬Elim, Del, Rep, Comp.
B. Rep, Comp, Cons.
C. Del, Comp, Cons.
D. Intro, Elim, ¬Intro, ¬Elim, Cons, Comp.
E. Intro, Elim, ¬Intro, Del, Cons.
F. Intro, Elim, ¬Elim, Del.
G. Intro, Elim, ¬Elim, Rep.
H. Out, Elim, ¬Intro, Del, Rep, Cons.
I. Elim, ¬Elim, Del.

Leigh's paper demonstrates that B–I yield a conservative theory over PA when they are adjoined to $\mathsf{Base}_{\mathrm{T}}$ (rather than $\mathsf{Base}_{\mathrm{T}}$ + U-Inf). This result indicates that U-Inf is responsible for the non-deflationary deductive power of $\mathcal{B}$–$\mathcal{I}$: that is to say, the culprit is spotted!

Leigh's paper sheds light on the subtle interactions between principles of truth and their effects on the truth-free consequences of a theory. For example, his results naturally raise the following interesting question: why does U-Inf add proof-theoretic strength? One may find a certain contrast between 'compositional' and disquotational principles here; U-Inf and its dual E-Inf may be called 'compositional' principles of truth because they partially axiomatize the compositional nature of truth where the truth of a sentence depends on the truth (or semantic value) of the constituents of that sentence; whereas the other principles, except Cons and Comp, are disquotational in the sense that they capture a fragment of the T-schema (for $\mathcal{L}_{\mathrm{T}}$).

1.2.4.3 "Deflationism and Instrumentalism" by Martin Fischer

At the root of deflationary conceptions is the suggestive, yet vague, idea that the notion of truth is 'useful' but not 'substantial'. This understanding of deflationism is underspecified as it leaves room for a variety of explanations, and indeed, deflationists disagree as to how to make it precise. On one understanding, truth is an 'expressive device' but not a natural property; on another, it is not a property at all. Some say that truth has no causal-explanatory force, while still others claim that it has no explanatory force whatsoever. Some think that all that is essential to our understanding of the concept of truth is our acceptance of Tarski's T-schema or some subset of its instances; others that the notion of truth is essentially a logico-syntactic device. Most of these claims, however, proved to be sufficiently vague to make it hard to assess, prove or refute them conclusively. To remedy this situation logical tools have had an increasing presence in discussions of deflationism in recent years. Philosopher-logicians, following the lead of Tarski, have discussed what would count as precise 'adequacy conditions' for a formalized theory, such that they would serve

as explications, in Carnap's sense, of the deflationists' claim. Such adequacy conditions are typically meant to ensure that the theory is a theory of *truth* (this is also a way to understand Tarski's material adequacy condition), that the theory accounts for a class of expected uses of the truth predicate (which account for its 'usefulness') and that truth is 'non-substantial' or innocent. The conclusions have not been overly favourable to the deflationist in that on the one hand, it has been claimed, against deflationism, that on some plausible precise explication of 'usefulness' and 'non-substantial' deflationary theories of truth cannot exist; while on the other hand, the deflationists have been slow to provide alternative consistent adequacy conditions.

In his essay, *Deflationism and Instrumentalism*, Fischer takes up the challenge on behalf of the deflationist. His philosophical starting point is the idea that deflationism about truth should be construed as a branch of instrumentalism. According to this view, vindication of deflationism depends on the possibility of carrying out an instrumentalist programme conceived in the spirit of Hilbert's programme that aims to show both the instrumental utility and the theoretical innocence of the truth predicate. In this paper, Fischer argues that this programme can be carried out successfully. The core of the argument is the proof that it is possible to devise a truth theory (Fischer's favourite being PT^-) which is conservative over PA (conservativity for innocence), and at the same time allows significant epistemic benefits by shortening of proofs (speed-up for epistemic usefulness). By proving conservativity and speed-up results for PT^-, Fischer lends credibility to his instrumental deflationism. Perhaps this would not have been entirely satisfactory if Fischer had not, in the same paper, also addressed the now classical criticism of conservative theories of truth due to Shapiro (1998) and Ketland (1999). The criticism is based, roughly speaking, on the notion that theories of truth should prove the soundness of their base theories and that a conservative theory of truth cannot do so (by Gödel's second incompleteness theorem). In the last section of his paper, Fischer argues that this criticism is not conclusive. Fischer does not directly take issue with the claim that an adequate theory of truth should prove soundness. Rather, he insists that, as is well known to logicians, the impossibility result strongly relies on soundness being formulated in the form of strong reflection principles. There are other ways to express soundness, however, and Fischer argues that some of them are such that: firstly, they constitute acceptable formulations of soundness; and secondly, a truth-theoretic conservative extension of PA can prove the soundness of PA so expressed. If Fischer is right in his conclusion, then his new way to deal with the 'conservativity argument' deserves serious consideration from both the deflationist and the non-deflationist. More generally, his essay proposes new logico-philosophical standards for assessing the value of various formalized theories of truth from a deflationist perspective, at the intersection between formal and philosophical reflections on truth.

1.2.4.4 "Typed and Untyped Disquotational Truth" by Cezary Cieśliński

Cieśliński's contribution studies a certain type of axiomatic theories of truth called *disquotational theories* of truth. In particular, it focuses on those disquotational

theories that are conservative over their base theory (PA in this setting), and for this reason generally thought to be consonant with the deflationists' claims about the nature and function of truth. This paper shows that there is more to the notion of conservativity that the deflationist ought to consider.

For the sake of simplicity, let us assume that the bearers of truth are sentences, though the following discussion would still apply for standpoints that take other objects—such as proposition—as the bearers of truth.

The core doctrine of *disquotationalism* holds that the content of the notion of truth is thoroughly captured by a certain collection of the so-called T*-biconditionals*, which are the sentences of the following form:

$$\sigma \text{ is true iff } \sigma. \tag{1.1}$$

Let us call statement 1.1 the T-biconditional *for* σ.

Tarski's famous Convention T contends that a predicate T is a materially adequate truth predicate of a language $\mathcal{L}$, when T validates the *full* T*-schema*: i.e.

$$T\ulcorner\sigma\urcorner \text{ iff} \sigma, \text{for } \textit{all} \text{ sentences } \sigma \text{ of the language } \mathcal{L}, \tag{1.2}$$

where $\ulcorner\sigma\urcorner$ denotes the name (or *structural descriptive* in Tarski's terminology) of sentence σ. However, Tarski's undefinability theorem tells us that no language $\mathcal{L}$ can contain a materially adequate truth predicate T for the language $\mathcal{L}$ itself. For, by taking a self-referential sentence λ such that $\neg T\ulcorner\lambda\urcorner$ iff λ, we can immediately derive a contradiction from the particular T-biconditional for λ, which is an instance of the full T-schema.

This fact forces disquotationalists to place restrictions on the full T-schema and impels them to find a suitable proper subschema of the full T-schema by excluding the contradictory T-biconditionals. Cieśliński's paper begins with a brief exposition of the philosophical background of the disquotational approach, and the reader may also refer to Halbach (2010) and Horsten (2011).

Now, as we have seen, at least the T-biconditional for the above particular λ ought to be excluded from any disquotational theory of truth. However, we cannot block inconsistency simply by expelling λ, since many other T-biconditionals yield inconsistency in various manners.[32] Thus the question arises of which subschema of the full T-schema correctly characterize the disquotationalist conception of truth.

An obvious candidate for a suitable disquotational theory of truth is that obtained by restricting the full T-schema to sentences that do not contain any occurrence of T; the resulting theory is often denoted by $\mathsf{TB}\restriction$ (or TB^{-}) in the literature[33], which stands

[32] For instance, let us consider a pair of sentences ρ_0 and ρ_1 such that $T(\ulcorner\rho_0 \leftrightarrow \rho_1\urcorner)$ iff ρ_0 and that $T(\ulcorner\rho_1 \leftrightarrow \neg\rho_0\urcorner)$ iff ρ_1; for a formal construction of such sentences in arithmetic, we refer readers to Boolos (1993) or Hájek and Pudlak (1993). Then, we can easily derive a contradiction from the T-biconditionals for ρ_0 and ρ_1.

[33] When the base theory is PA, the result of furthermore adding the expanded arithmetical induction schema for the expanded language (obtained by adjoining the truth predicate T to the language of PA as a new primitive predicate symbol) to $\mathsf{TB}\restriction$ is more simply called TB.

for "Tarski Biconditionals".[34] The theory $\mathsf{TB}\restriction$ is known to be conservative over its base theory, the proof of which is originally due to Tarski (1983): any consequence of $\mathsf{TB}\restriction$ in the base language can be derived in the base theory without using any T-biconditionals.

This restriction of the full T-schema results in a *typed* conception of truth and thus would be renounced by disquotationalists who like to encapsulate (part of) the self-applicative character of the notion of truth in their theories. Hence, disquotationalists in favour of a self-applicative conception of truth would have to search for another suitable "untyped" subschema of the full T-schema.

It is sometimes wrongly thought that disquotational theories must be deductively fairly weak. This misconception may perhaps have been wrongly deduced from Tarski's conservativity result for $\mathsf{TB}\restriction$ or derived from the redundancy theoretic point of view; redundancy theorists claim that "is true" can be eliminated by means of the equivalence between $T\ulcorner\sigma\urcorner$ and σ and thus "is true" is redundant. However, as is also mentioned in Cieśliński's paper, it follows from McGee's trick (McGee 1992) that any sentence, regardless of whether it contains the truth predicate or not, is equivalent to some T-biconditional over Peano arithmetic, and thus any set of axioms can be reaxiomatized over PA as a disquotationalist theory of truth, i.e. a set of T-biconditionals.[35] Consequently, "untyped" disquotational theories can have arbitrary strength. To make matters worse, McGee (1992) also showed that there are uncountably many mutually inconsistent maximally consistent subschemata of the full T-schema. Furthermore, it follows from Theorem 2 of McGee (1992) that there are uncountably many mutually inconsistent maximally sound (or ω-consistent) subschemata of the full T-schema. Later, Cieśliński (2007) even showed that there are uncountably many incompatible maximally conservative subschemata of the full T-schema. So, how should we select a consistent subschema and what kind of subschema is to be chosen from this vast variety of options? Unless we stick to a Tarskian-type distinction like $\mathsf{TB}\restriction$, the main challenge for the disquotationalist is the problem of specifying a sensible set of T-biconditionals.

Although in principle they are independent of each other, disquotationalism is often correlated to (or even subsumed in) deflationism in the philosophical literature; we may perhaps say that the disquotationalist tenet that the statement "it is true that . . . " is a mere paraphrase of the statement ". . . " gave the prototype of deflationism of truth. For instance, Halbach (2010, § 21) counted disquotationalism as one of the

[34] Alternatively, we may allow *parameters* in T-biconditionals and obtain a disquotational theory of truth by restricting the full T-schema with parameters to the sentences containing no occurrence of T; the resulting theory is often called $\mathsf{UTB}\restriction$ (or UTB^{-}), which stands for "Uniform Tarski-Biconditionals"; for more details, see Halbach (2010). As in the case of $\mathsf{TB}\restriction$, the result of adding the expanded arithmetical induction schema to $\mathsf{UTB}\restriction$ is called UTB, which corresponds to UTB_1 in Cieśliński's paper.

[35] Indeed, it follows from the arithmetized completeness theorem that any recursive consistent theory over any recursive language can be interpreted in some disquotational theory. Hence, for example, disquotational theories (over PA) can have even greater consistency strength than ZF plus the existence of very large cardinals.

core doctrines of deflationism. According to Halbach, another core doctrine of deflationism consists of the insubstantiability of truth: "truth is a thin notion in the sense that it does not contribute anything to our knowledge of the world" (Halbach 2010, p. 310). As mentioned earlier (c.f. § 2.4), some argue that this second doctrine translates into the conservativity requirement, according to which a deflationary theory of truth must yield no consequence that is not already a consequence of the base theory.

A common formal formulation of the conservativity requirement is:

If a theory S of truth over a base theory B is deflationary,

then $\mathsf{S} \vdash \phi$ implies $\mathsf{B} \vdash \phi$ for all sentence ϕ of the language $\mathcal{L}_0$ of B. (1.3)

However, there is another formulation of "conservativity". We say a theory S of truth is *conservative in the semantic sense* (or *semantically conservative* in Cieśliński's paper) over a base theory B iff every model of B is expandable to a model of S: So from this, due to the completeness theorem of first-order logic, conservativity in the semantic sense implies conservativity in the ordinary sense (or in the proof-theoretic sense (McGee 2006)); while the converse does not generally hold. For example, McGee (2006) argues that conservativity in the semantic sense is preferable from the deflationist point of view and has more philosophical significance, since a move from a base theory B to a conservative theory of truth over B in the semantic sense makes no difference on what the world is like and there is no metaphysical cost in such a move. Hence, it may be worth considering an alternative formulation of the conservativity requirement in the semantic sense: i.e.

If a theory S is deflationary, then S is conservative in the semantic sense over B. (1.4)

Cieśliński's paper shows that some disquotational theories are conservative in the proof-theoretic sense but not in the semantic sense. More precisely, he shows that only a recursively saturated model of PA can be expanded to a model of any of these disquotational truth theories. His paper is expected to shed more light on the distinction between the two formulations, 1.3 and 1.4, of the conservativity requirement, which has attracted less attention from philosophers than it should have, and is expected also to awake more technical interest in the problem of the semantic conservativity of theories of truth.

1.2.4.5 "New Constructions of Satisfaction Classes" by Ali Enayat and Albert Visser

Also concerned with conservativity, Enayat and Visser's paper "New Construction of Satisfaction Classes" presents a new proof of the conservativity of the axiomatic theory of a *full satisfaction class* over PA.

The modern definition of truth that we use today in model theory is usually credited to Tarski and Vaught (Tarski 1983; Tarski and Vaught 1956). They defined truth in

a model-theoretic structure $\mathfrak{M}$ via the definition of satisfaction for $\mathfrak{M}$. Given a structure $\mathfrak{M}$ for a language $\mathcal{L}$, we first define when each $\mathcal{L}$-formula ϕ is "satisfied" by a variable assignment α in $\mathfrak{M}$, which is a function from variables to the domain M of $\mathfrak{M}$, and thereby define that an $\mathcal{L}$-sentence is true in $\mathfrak{M}$ iff it is satisfied by at least one assignment α (or, equivalently, by all assignments α) in $\mathfrak{M}$. Hence, we may say that a theory of (Tarskian) satisfaction subsumes a theory of (Tarskian) truth. Enayat and Visser present a theory $\mathsf{PA}^{\mathsf{FS}}$ ("FS" for "full satisfaction class") of satisfaction over a base theory PA and in their paper prove that $\mathsf{PA}^{\mathsf{FS}}$ is conservative over PA for the language $\mathcal{L}_{\mathsf{PA}}$ of PA. This means that any arithmetical theorem of $\mathsf{PA}^{\mathsf{FS}}$ (i.e. an $\mathcal{L}_{\mathsf{PA}}$-sentence derivable from $\mathsf{PA}^{\mathsf{FS}}$) is already derivable in PA.

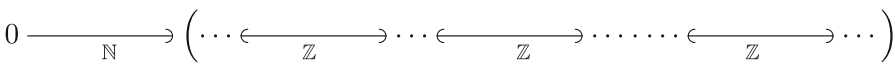
The theory $\mathsf{PA}^{\mathsf{FS}}$ is an axiomatic characterization of the Tarskian definition of model-theoretic satisfaction (for models of PA). It is formulated over the language $\mathcal{L}_{\mathsf{PA}}$ of PA plus a new binary predicate symbol S which takes a code or Gödel number of an $\mathcal{L}_{\mathsf{PA}}$-formula as its first argument and a variable assignment (of codes of variables to natural numbers) as its second argument, where $\mathsf{S}(x, y)$ expresses "a formula (coded by) x is satisfied by a variable assignment (coded by) y". For example, the $\mathsf{PA}^{\mathsf{FS}}$-axiom for negation is expressed as:

$$\forall x \forall y \forall u \big[\text{“}x \text{ is the code of the negation of the code } y \text{ of an } \mathcal{L}_{\mathsf{PA}}\text{-formula”} \wedge \text{“}u \text{ is a variable assignment”} \rightarrow \big(\mathsf{S}(x, u) \leftrightarrow \neg \mathsf{S}(y, u)\big)\big]. \tag{1.5}$$

For a model $\mathfrak{M}$ of PA with the domain M and for a binary relation $S \subset M \times M$, we say that S is a *full satisfaction class* for $\mathfrak{M}$ when $(\mathfrak{M}; S)$ is a model of $\mathsf{PA}^{\mathsf{FS}}$ in which S is interpreted by S. Precisely what Enayat and Visser prove in the present paper is that, for any model $\mathfrak{M}$ of PA, we can construct an elementary extension $\mathfrak{M}'$ of $\mathfrak{M}$ with the domain M' and a set $S \subset M' \times M'$ such that $(\mathfrak{M}'; S)$ is a model of $\mathsf{PA}^{\mathsf{FS}}$. This immediately entails the conservativity of $\mathsf{PA}^{\mathsf{FS}}$ over PA for $\mathcal{L}_{\mathsf{PA}}$ due to the completeness theorem.

Now, we know that Tarskian satisfaction can be defined for *any* structure. So, one might expect that $\mathsf{PA}^{\mathsf{FS}}$ is conservative over PA by reasoning in the following (wrong) way: since $\mathsf{PA}^{\mathsf{FS}}$ is an axiomatization of satisfaction that can be defined for every $\mathcal{L}_{\mathsf{PA}}$-structure, every model $\mathfrak{M}$ of PA can be expanded to a model $\mathfrak{M}^+$ of $\mathsf{PA}^{\mathsf{FS}}$ simply by interpreting S by the so-defined satisfaction for $\mathfrak{M}$, and we get the desired conservativity by the completeness theorem. It is true that $\mathsf{PA}^{\mathsf{FS}}$ is conservative over PA but this reasoning is fallacious; the proof of this conservativity is never that simple, and this fallacious reasoning is only valid when $\mathfrak{M}$ is the standard model of arithmetic.

By the compactness theorem, PA has a non-standard model and its non-standard part is ill-founded. In general, a non-standard model of PA has the following structure:



What we have here is many linear orderings of "non-standard numbers" *order-isomorphic to* $\mathbb{Z}$ topped up on the initial standard part that is order-isomorphic to $\mathbb{N}$.

In fact, a countable non-standard model of PA is order-isomorphic to $\mathbb{N} + \mathbb{Z} \cdot \mathbb{Q}$; see Kaye (1991, Chap. 6.2) for more details. In particular, each $\mathbb{Z}$-part of a non-standard model of PA has no end-points and is ill-founded with respect to the less-than relation $<$; and a non-standard number that lies on any $\mathbb{Z}$-part is greater than all standard numbers lying on the initial standard $\mathbb{N}$-part. Hence, we may informally say that non-standard numbers are "infinite" numbers. Since each non-standard $\mathbb{Z}$-part of a non-standard model of PA is ill-founded, the induction principle does not hold for these ill-founded $\mathbb{Z}$-parts. That is, for a non-standard model $\mathfrak{M}$ with the domain M, even if we have $\forall x \in M(\forall y \in M(y < x \rightarrow y \in X) \rightarrow x \in X)$ for $X \subset M$, we do not necessarily have $X = M$ (but $X = M$ holds under this assumption when X is $\mathcal{L}_{\mathsf{PA}}$-definable, since $\mathfrak{M}$ was assumed to be a model of PA). Another notable feature of non-standard models of PA is the "overspill" phenomenon. This is such that if an $\mathcal{L}_{\mathsf{PA}}$-definable property P is satisfied by unboundedly many standard numbers in the initial $\mathbb{N}$-part of a non-standard model $\mathfrak{M}$ of PA, then P is satisfied by some non-standard number of $\mathfrak{M}$ as well (or so to speak, the class $\{x \mid Px\}$ "spills over" the $\mathbb{N}$-part); see Kaye (1991, Chap. 6.1) for proof of this and more details. Consequently, a non-standard model $\mathfrak{M}$ of PA contains "non-standard $\mathcal{L}_{\mathsf{PA}}$-formulae". Precisely, for an $\mathcal{L}_{\mathsf{PA}}$-formula $\mathsf{Form}(x)$ that expresses "x is a code of $\mathcal{L}_{\mathsf{PA}}$-formula", $\mathfrak{M}$ contains a non-standard number a such that $\mathfrak{M} \models \mathsf{Form}(a)$; see Halbach (2010, § 8.3) for more detailed expositions. Similarly, a non-standard model of PA contains "non-standard variables", "non-standard syntactic complexities of formulae", etc.

Now, recall that the Tarskian definition of satisfaction for a structure $\mathfrak{M}$ is arrived at by recursion on the syntactic complexity of formulae (or the number of logical constants, etc.): we start the definition of satisfaction from the simplest formulae, i.e. atomic formulae; then, we define satisfaction for more and more complex formulae step by step using the definition of satisfaction already given for less complex formulae. Here, it is crucial that this definition applies directly to formulae and *not* to their "codes" in $\mathfrak{M}$. In contrast, however, the predicate S is interpreted in any model $\mathfrak{M}^+$ of $\mathsf{PA}^{\mathsf{FS}}$ as a relation over the domain M of $\mathfrak{M}^+$: the interpretation of S is a relation of M-elements satisfying a formula $\mathsf{Form}(x)$ in $\mathfrak{M}$ (i.e., M-elements *coding* $\mathcal{L}_{\mathsf{PA}}$-formulae in $\mathfrak{M}$) and M-elements representing variable assignments. In other words, the ordinary model-theoretic definition of satisfaction is for objects external to the structure $\mathfrak{M}$; whereas a full satisfaction class is to be defined for objects *in* the structure $\mathfrak{M}$. Hence, when $\mathfrak{M}$ is non-standard, the relation S may take a non-standard formula a whose syntactic complexity b is also non-standard. In such a case, an ordinary recursion or inductive definition of syntactical complexity (as an element of M) cannot reach the definition for a with complexity b, since b lies on an ill-founded $\mathbb{Z}$-part not accessible from the least element 0 by such a recursion process in a step-by-step manner.

This explains why the proof of the conservativity of $\mathsf{PA}^{\mathsf{FS}}$ is not as easy as it looks.[36] Furthermore, Lachlan (1981) showed that a non-standard model $\mathfrak{M}$ of PA only has a full satisfaction class when $\mathfrak{M}$ is *recursively saturated* (see Kaye 1991 for a definition of this). Therefore, not every model of PA has a full satisfaction class and therefore neither can they all be expanded to a model of $\mathsf{PA}^{\mathsf{FS}}$, since not all non-standard models of PA are recursively saturated.

As explained in Enayat and Visser's paper in more detail, the first conservativity proof for $\mathsf{PA}^{\mathsf{FS}}$ is given by Kotlarski et al. (1981). They show that every recursively saturated model of PA can be expanded to a model of $\mathsf{PA}^{\mathsf{FS}}$. So, since every model of PA has a recursively saturated elementary extension, this result yields the conservativity of $\mathsf{PA}^{\mathsf{FS}}$. To our knowledge, since that proof was offered by Kotlarski, Krajewski and Lachlan, there has been no other proof of this conservativity result (except for variants or extensions), and it is indeed quite technical and complicated. Enayat and Visser provide a simpler and more versatile proof of this conservativity result.[37]

1.2.5 Truth Without Paradox

The three papers in this chapter discuss solutions to the truth-theoretic paradoxes that escape sentences that are problematic in either their semantics or their syntax. Armour-Garb and Woodbridge argue for an approach in which some sentences employing the truth predicate, such as the Liar sentence, are to be seen as semantically defective. The authors defend the meaningless status of semantically problematic sentences, i.e. the fact that they lack wordly content, by means of a fictionalist account of truth-talk. Fictionalism in this context takes truth-talk to be part of a game of make-believe and the proper use of the truth predicate to be determined by the rules of that game. These rules establish certain worldly conditions as prescriptive for the pretenses displayed in (non-pathological) instances of truth-talk, so the talk thereby functions as an indirect means for specifying those conditions. The authors show how their account not only blocks the Liar paradox but also the reasoning underlying its revenge.

The view that some sentences that employ the truth predicate are semantically defective is also shared by Bonnay and Van Vugt; for them, however, semantic defectiveness amounts to lack of groundedness. In an attempt to make this idea

[36] The proof of the conservativity of the theory of satisfaction or Tarskian truth over set theory ZF is relatively easy. It was first model-theoretically shown by Krajewski (1976); and an elementary proof-theoretic proof can even be found in Fujimoto (2012).

[37] As Enayat and Visser state in their paper, the proof they present here still has some limitations in its application. For example, their proof assumes that the language of arithmetic is purely relational without any constant or function symbols. They announce in the paper that their techniques can be suitably modified for many other settings with functional languages and much weaker base theories. This immediately entails, for instance, that the theory $\mathsf{CT}\restriction$ of "compositional truth (with restricted induction)", also known as TC^{-} of "Tarskian inductive clauses (without expanded induction)", formulated in the standard functional language of arithmetic is also conservative over PA.

more precise, they compare two different ways of conceptualizing groundedness, as found in Kripke (1975) and Leitgeb (2005), and study the extent to which they are extensionally equivalent. In order to get a good grasp of the difference between the two approaches, Bonnay and Van Vugt take a closer look at Leitgeb's notion of conditional dependence, which they elucidate using the notion of groundedness according to the supervaluational scheme.

We saw earlier that in order to counter the so-called 'one concept objection' (according to which a hierarchy furnishes many concepts of truth when there is in fact only one), Glanzberg introduces the notion of stratification and argues that a concept may be stratified yet unique. In his contribution to this volume, Cantini proposes an axiomatization of the Tarskian hierarchy which essentially uses stratification in Quine's sense in order to represent a consistent theory of untyped truth. We should recall that the traditional theories of typed truth have been criticized for not being capable of representing the untyped ordinary notion of truth, and as a result various theories of untyped truth have been presented, such as Feferman's KF and Cantini's own VF. Cantini adds a new kind of theory of untyped truth to the existing variety of such theories; it is a theory of stratified—hence, not generally self-referential—untyped truth.

1.2.5.1 "Truth, Pretense and the Liar Paradox" by Bradley Armour-Garb and James A. Woodbridge

In their contribution Armour-Garb and Woodbridge propose a new fictionalist framework to articulate deflationism about truth. They argue that their construal of deflationism allows for a way out of the truth-theoretic paradoxes. According to the brand of fictionalism that the authors defend—pretense-involving fictionalism, as they call it— truth-talk always invokes a background game of make-believe in its functioning, in virtue of which these sentences serve as an indirect means for specifying (as obtaining or not obtaining) the worldly conditions that the game's rules establish as prescriptive for the pretenses that the sentences display. These constitute the meaning-conditions specified by the sentences. Thus, for instance, our truth-involving language game compels us to pretend that expressions such as 'is true' are descriptive predicates—which they no more are than children pretending to be cowboys in the courtyard really are cowboys—and also that the pretenses displayed in a utterance of 'It is true that snow is white' are prescribed if and only if snow is white, and so on. The rules continue in a similar manner for other classical conditions that deflationists take to govern the use of the truth predicate. The authors claim there are two main benefits of their approach over other deflationist frameworks. The first is methodological: it allows for unification of the deflationist treatment of truth-talk with widely used fictionalist strategies in other areas of philosophy. More specifically, in this paper the authors argue for a second benefit: that their pretense account of truth is immune to the Liar paradox and its revenge. This is because, the authors argue, their account allows for a successful version of a "meaningless" strategy against these paradoxes.

Regarding the simple Liar sentence, the rough idea of the strategy is as follows. The Liar sentence is an instance of truth-talk, and as a case of pretense talk, any

meaning it had would be constituted by whatever worldly conditions the rules of the pretense establish as prescriptive for the pretenses it displays. However, the Liar sentence manifests a kind of 'ungroundedness' because the rules of the pretense do not make any worldly conditions prescriptive for the pretenses it displays, and so the Liar sentence specifies no meaning-conditions. In this way, sentences like the Liar suffer a kind of semantic defectiveness that the authors go on to partially characterize in their essay. Furthermore, because such sentences do not specify any meaning-conditions, there are no conditions that are prescriptive for the pretenses invoked in asserting that the Liar sentence is true, or that it is false, or even that it is not true. All such instances of truth-talk are also semantically defective. It follows that sentences like the Liar are not truth-evaluable at all. What about the revenge? It is well known that many solutions to the Liar paradox that accept partitioning the domain of sentences into true sentences, false sentences, and sentences with a third kind of truth value, I, are generally vulnerable to a revenge paradox, by way of reasoning involving a sentence which says of itself that it is false or I. Let λ be the sentence: λ is false or λ is I. The usual reasoning is: (1) λ cannot be true, since this would imply that it is false or I (both of which contradict the hypothesis that λ is true). (2) If λ is false, then its first disjunct is true, hence λ is true after all—again contradicting the hypothesis. (3) Then finally, if we hold λ to be I, λ must be true in virtue of its second disjunct, which now contradicts its being I—and we are caught in the revenge paradox.

In the authors' framework, no third truth-value is involved, but one could construct a corresponding tentative revenge sentence in the form of a sentence λ which says that λ *is not true or* λ *is semantically defective*. The authors argue that this sentence is semantically defective. Moreover, and this is where their work on the meaning-conditions of truth-talk pays off, it is argued that even though the sentence 'λ is semantically defective' is true, that does not imply that the sentence 'λ is not true or λ is semantically defective' (i.e., λ itself) is true. This is because: first, λ being devoid of content means that its first disjunct is an instance of truth-talk that (as in the simple Liar case) fails to specify any meaning-conditions; and second, disjoining a meaningless sentence with a true sentence results in a meaningless whole, in virtue, roughly speaking, of the compositional features of meaning constitution. Since a meaningless sentence cannot be a consequence of anything, it follows that λ being semantically defective blocks the inference from the claim that it is semantically defective to λ itself. Thus, the air of paradox is dispelled as, specifically, the semantic defectiveness of the problematic sentence λ no longer implies its truth. Carrying on with their earlier work on meaning and understanding in connection with liars, the pretense-involving fictionalism defended by the authors here contributes to giving new bite to "meaningless strategies" against the paradox.

1.2.5.2 "Groundedness, Truth and Dependence" by Denis Bonnay and Floris Tijmen Van Vugt

Gupta and Belnap (1993) argued that in order to make a correct diagnosis of the truth-theoretic paradoxes one should not focus on the paradoxical sentences themselves,

but instead try to understand the ordinary behaviour of the truth predicate. In the same spirit, Bonnay and van Vugt are interested here not in determining which sentences are pathological, but in characterizing unproblematic ones. The aim of their paper is to compare two prominent characterizations of non-pathological sentences due to Kripke (1975) and Leitgeb (2005). Both characterizations share the basic intuition that unproblematic sentences are those that are grounded in the world, i.e. those whose truth value ultimately depends on what the world is like. But Kripke and Leitgeb propose two different accounts of groundedness that are not extensionally equivalent and here Bonnay and van Vugt wish to identify the parameters responsible for their divergence.

Kripke follows an indirect route in order to determine the collection of grounded sentences, in the sense that he first defines the extension of the truth predicate via a fixed-point construction and then he defines grounded sentences as those that have a truth value at the least fixed point. Of course, which sentences are grounded also depends on the scheme of evaluation used. Leitgeb, in contrast, uses a direct route; he determines the class of grounded sentences as a fixed point of a construction that directly identifies sentences whose truth value does not depend on the truth predicate. In what follows, J^K denotes the Kripke jump for the strong Kleene scheme of evaluation; then K-grounded sentences are those grounded in Kripke's sense according to the least fixed point of J^K. J^V, and V-grounded are the corresponding notions for the supervaluational scheme. L-grounded sentences are those grounded according to Leitgeb's characterization.

L-groundedness does not imply K-groundedness; nor vice versa. As an example of the difference between L-grounded and K-grounded sentences, the authors consider the sentence $Tr\ulcorner 2+2=4\urcorner \vee \lambda$, where λ is the Liar sentence. As $2+2=4$ is true, $Tr\ulcorner 2+2=4\urcorner$ is also true, which makes $Tr\ulcorner 2+2=4\urcorner \vee \lambda$ a K-grounded sentence. However, according to Leitgeb's definition of dependence, the sentence $Tr\ulcorner 2+2=4\urcorner \vee \lambda$ depends both on $2+2=4$ and on λ, hence it is L-ungrounded; although $2+2=4 \vee \lambda$ does not depend on any other sentence and so it is L-grounded. This is a result that the authors find counterintuitive and they explore ways to avoid this. They consider Leitgeb's notion of conditional dependence, according to which sentences that are declared grounded in one step of the construction depend on the grounded sentences in the previous step and on the partition of those into true and false sentences. The main result of the paper, which contributes to a better understanding of the connection between these two notions of groundedness, is that the set of grounded sentences for conditional dependence coincides with the V-grounded sentences.

1.2.5.3 "On Stratified Truth" by Andrea Cantini

Cantini's contribution is motivated by the problem of "finding a consistent axiomatization of the Tarskian hierarchy, where stratification is understood in Quine's sense", which he accredits to Feferman. Let us start by explaining what "stratification in Quine's sense" means formally.

Russell's and other set-theoretic paradoxes suggest that "circularity" must be somehow restricted in set theory for the sake of consistency. This view led to the

idea of type hierarchy, according to which, each set is assigned a "type" and may contain only sets of lower types. "Circularity" is thereby excluded by stipulating that the types form a well-founded hierarchy; for then the type of any set a cannot be lower than itself.

A natural formal implementation of this idea is the so-called theory of types. In the theory of types, each item of vocabulary of a language is syntactically assigned types. For instance, each variable is indexed by a natural number $i < \mathbb{N}$, indicated by writing v^i, where the hierarchy of the types is given by the less-then relation of natural numbers, and an expression $v^i \in v^j$ is syntactically well-formed only when $j = i+1$. Then each compound formula is uniquely assigned a type in a straightforward manner by taking the maximum of the types that occur in the formula, and the axiom schema of comprehension is reformulated as

$$\exists x^{i+1} \forall z^i (z^i \in x^{i+1} \leftrightarrow \phi(z^i)), \text{for a formula } \phi \text{ of type } i+1.$$

Quine's New Foundation, NF, uses a different method to formally implement the idea of type distinction. In NF, each item of vocabulary, such as a variable and the membership relation $\in$, is *not* indexed by any type *at the syntax level*; the language of NF is an ordinary first-order language of set theory with a single sort. However, instead of syntactical type distinctions, NF restricts the axiom schema of comprehension to formulae that *could* be typed by natural numbers; those potentially "typable" formulae are said to be *stratified* in this context. For example, $x \in x$ is not stratified since there is no possible type assignment in which the type of x is higher than that of x. In contrast, $x \in y$ is stratified (provided that $x \neq y$) since we can assign any natural number n to x and then assign $n+1$ to y; so NF reformulates the axiom's schema of comprehension as follows:

$$\exists x \forall y (x \in y \leftrightarrow \phi(x)), \text{for a stratified (i.e. typable by natural numbers) formula } \phi.$$

At first glance, the difference between the two formal implementations of the idea of type distinction looks superficial, and one might expect that NF is essentially the same as the theory of types. This is not the case, however. For a typical example, the universal set provably exists in NF since $x = x$ is obviously stratified; in contrast, there is no such set in the ordinary theory of types. Interestingly, it turned out that NF is quite a complex theory and, in fact, there is no consistency proof for it yet.

Truth predicates are applied to terms that refer to sentences, and the Liar paradox is (usually diagnosed to be) caused by a self-referential application of a truth predicate; specifically, it is caused by a sentence in which a truth predicate is applied to a term that refers to the sentence itself. Given this, the Liar paradox and its cousins suggest, in essentially the same way as Russell's paradox suggests for set theory, that "self-reference" must be somehow restricted in theories of truth in order to avoid inconsistency. The argument of Tarski's undefinability theorem, which is a formalization of the Liar paradox within formal systems, tells us that the following so-called T-schema

$$T(\ulcorner \sigma \urcorner) \leftrightarrow \sigma, \text{for all sentences } \sigma,$$

cannot be consistently sustained if we allow σ to contain the truth predicate T. This naturally suggests the view that application of a truth predicate should be restricted to sentences that do not contain that predicate.

This view gave rise to the theory of typed (or ramified) truth. The language of a theory of typed truth contains a stock of more-than-one truth predicates each of which is indexed by its type, i, and the application of the truth predicate of type i is restricted to sentences that only contain truth predicates of types lower than i. As in the case of the theory of types, each formula of a theory of typed truth is uniquely assigned its type. So the truth theoretic axioms such as the T-schema are reformulated to:

$$T_i(\ulcorner \sigma \urcorner) \leftrightarrow \sigma, \text{ for a sentence } \sigma \text{ of type lower than } i.$$

Now we can see a clear analogy between the theory of types and the theory of typed truth in their remedy for the paradoxes.

Cantini's new theory SFT takes an alternative route to consistently restrict "self-referential" applications of a truth predicate in the same way as Quine's NF does for consistently restricting "circularity". Cantini adopts a language with a single universal truth predicate T without any index or "typing", but introduces the notion of *stratified* formulae of truth theory in the same spirit as Quine's notion of stratified formulae of set theory: a stratified formula is a formula in which we could suitably type the occurrences of terms and the truth predicate T by natural numbers; see Definition 1.3 of the paper. Thereby, Cantini restricts the T-schema as well as the other truth-theoretic principles to stratified sentences (more generally, stratified formulae); e.g.

$$T(\ulcorner \sigma \urcorner) \leftrightarrow \sigma, \text{ for a stratified sentence } \sigma.$$

For instance, the Liar sentence λ has the property $\ulcorner \lambda \urcorner = \ulcorner \neg T \ulcorner \lambda \urcorner \urcorner$, but a term such as $\ulcorner \lambda \urcorner$ is not stratified and properly excluded from the range of application of the truth predicate in the stratified T-schema. Cantini then proceeds to provide a proof of the relative consistency of his SFT to NF. Hence, if NF is consistent, SFT is consistent; although we do not yet know whether NF is consistent or not.

1.2.6 Inferentialism and the Revisionary Approach

The title of the previous chapter, 'Truth without Paradox', fits the papers collected here well too, since they present work primarily motivated by the goal of evading the truth-theoretic paradoxes. The reason for grouping these papers into a separate section is to emphasize their interest in the inferential substrate underlying the derivation of the paradoxes, whether it is to do solely with principles governing truth or it extends to the logic. All the papers can be said to follow a revisionary approach; albeit not in the same sense. Theories of truth are usually called revisionary if they revise the underlying logic to avoid compromising naive properties of truth. A well-known example of such an approach can be found in Priest's contribution at the end of this chapter where a paraconsistent logic is proposed to replace classical logic. However,

according to Murzi and Shapiro, as well as Zardini, such revisions are only partly successful. Those authors take an even more radical approach in proposing a revision of the structural rules of the underlying logic; and in fact both of their papers draw our attention to the rule of contraction. Also on the substructural level, Cobreros, Egré, Ripley and van Rooij challenge the rule of transitivity, though only for reasoning that involves truth. This focus on truth-related reasoning[38] is shared by Read who is, however, primarily concerned with the (in)validity of specific principles of truth, as opposed to the logic.

Read exploits a solution to the Liar paradox due to the medieval philosopher Bradwardine, according to which the Liar sentence comes out false. The solution is based on Burley's semantics of signification and truth. Read crucially explores the semantics and logic underlying this solution to the paradox which he shows to have some attractive features. In particular, and unlike the following papers, the logic itself remains intact, while some problematic truth principles are proved by Read not to follow from the semantics; in particular, T-OUT (i.e. $T^{\ulcorner}\phi^{\urcorner} \rightarrow \phi$) and commutation of truth with negation and conditional. This gives a principled way for motivating a consistent notion of truth.

Inspired by their treatment of vague predicates, Cobreros, Egré, Ripley and van Rooij propose a novel consequence relation, their so-called strict-to-tolerant consequence relation, which is permissive in that it allows for tolerantly asserted conclusions (conclusions that take the value 1 or 1/2) from strictly asserted premises (premises that take the value 1). This consequence relation basically blocks the Liar paradox by invalidating transitivity for inference steps whose conclusions are only tolerantly accepted; while, given its permissive character, it preserves transparency for truth. The authors explain why dropping transitivity for some inferences that involve truth is not as implausible as one may think at first. As with Read's paper, the underlying logic here is not affected in the absence of the truth predicate, since strict-to-tolerant and classical consequence coincide in this case.[39]

From permissive consequence, Murzi and Shapiro take us back to the traditional intuitive notion of validity as truth preservation (VTP) which is often dismissed by revisionary theorists of truth. The authors believe that just as the revisionary theorist is interested in guarding the naive properties of truth, importantly the transparency (or intersubstitutivity) of truth, naive properties of validity should also be safeguarded. They rehearse the usual reasons for repudiating VTP and argue that these are based on a solution to the semantic paradoxes that is, however, problematic. Showing then that rejecting the rule of contraction offers a better solution to the paradoxes—a rule the authors believe in any case to be in tension with naive principles governing the intuitive notion of validity—the reasons for dismissing VTP fall through.

---

[38] Recall that the term 'revision' is also used in the sense of applying only to truth by Halbach and Horsten in this volume.

[39] It is interesting to observe that transitivity is also what is at stake in the medieval sophism that Read discusses at the beginning of his paper as the starting point for developing an alternative semantics for signification and truth.

Zardini is on the same page as the previous two papers in considering the transparency of truth a necessary requirement for an adequate theory of truth. His endorsement of a contraction-free logic comes from comparing it with paracomplete and paraconsistent theories in their capacity to provide a solution to three paradoxical arguments that he presents. Zardini shows that not only is his proposal able, unlike the other theories, to deal with all three arguments, but that it does so by offering a unified solution to them. This conclusion is then philosophically strengthened even further by metaphysical motivation for rejecting the rule of contraction.

All the papers in this chapter endorse a unified approach to the semantic paradoxes. We find, for example, that Read, Murzi and Shapiro, and Zardini all discuss Curry's paradox. For Read, Curry's paradox plays a crucial role in improving Bradwardine's logic so that there is a principled reason (other than simply evading the Liar) for excluding commutation of truth not only with negation, but also with implication. Murzi and Shapiro, crucially base their support for a contraction-free based logic on the failure of alternative proposals to deal with Curry's paradox; as does Zardini. Finally, Cobreros, Egré, Ripley and van Rooij extend the idea of a unifying approach from the semantic paradoxes to the paradoxes of vagueness.

It is, therefore, fitting to conclude this chapter, and with it the volume, with Priest's paper which is a systematic expression of the idea of a uniform solution to the paradoxes. Priest revisits his Inclosure Schema which provides a general form underlying self-referential paradoxes. He tests and strengthens the validity of this schema by showing that a new intensional paradox, Kripke's thought paradox, falls under it. Subsequently, in accordance with his so-called Principle of Uniform Solution, he shows that dialetheism, which he has earlier defended as the solution to the Liar paradox, applies in this case too.

1.2.6.1 "Truth, Signification and Paradox" by Stephen Read

In his contribution Read engages in the discussion of Bradwardine's solution to the Liar paradox; one of the medieval solutions to the semantic paradoxes that is the focus of renewed interest today. The topic is introduced with a sophism that attracted some attention in the middle ages: "If I say that you are an ass, I say that you are an animal. And if I say that you are an animal I say something true. Therefore, if I say you are an ass, I say something true". The conclusion seems obviously false; so, if the premises are true, then the validity of the transitivity of the conditional (suffixing, in Read's terminology) is threatened. Read surveys several authors who all respond to the sophism by distinguishing two notions of "saying that". One corresponds more or less to the literal notion of meaning: the second premise is true because if I say that you are an animal I say something true; but the first premise is false because if I say that you are an ass I do not literally say that you are an animal. According to the second notion, when I say that you are an ass, I also say all the consequences of that statement; so the first premise is true, but then the second premise is false because when I say that you are an animal I do not always say something true (for example, if I say that you are an animal by means of saying, falsely, that you are an ass).

Read focuses on Burley's account of these ideas. Burley distinguishes four notions of proposition: the written proposition, the spoken proposition, the mental

proposition and the real proposition. The last two are a composition (or division) of concepts or real things, respectively. A written or spoken proposition signifies a mental proposition which is true when it corresponds to a true real proposition, i.e. when it composes concepts that stand for things which are really united in reality or divides concepts that correspond to things really divided in reality. The two key aspects of Burley's semantics are that signification is closed under consequence (hence, if I say that you are an ass, I also say that you are an animal), and that truth requires that all the things I say are in reality as signified (so that sometimes I may be saying something false when I say that you are an animal, because I may say so by saying that you are an ass).

The second section of the paper explains how these two theses are used by Bradwardine to develop an original solution to the Liar paradox. In the same way that a proposition that signifies that you are an ass also signifies that you are an animal, a proposition that signifies that itself is not true, also signifies that itself is true, because this follows from it. Hence the Liar proposition signifies both that it is true and that it is false and, therefore, the Liar is simply false, because not everything that the Liar signifies is true. Read explains in detail the derivation that Bradwardine gives of this result and also proves that from his postulates it follows that every sentence signifies its own truth, as was defended by several medieval logicians.

The last section of the paper studies the logic implicit in Bradwardine's solution to the paradoxes. He accepts half of the T-schema (T-OUT: if p is true, then p) but not the converse, since from the fact that p obtains, it does not follow that everything that p signifies obtains. Read then focuses on the analysis of the compositional principles of truth that fail for Bradwardine's solution; a problem shared with other solutions that deny T-IN (such as Kripke 1975 and Maudlin 2004). Bradwardine just adds the compositional principles for disjunction and conjunction as axioms (a conjunction is true iff each part is true, and it is false iff one of its parts is false; analogously for disjunction). But that does not satisfy Read, because, for instance, if we added similar compositional principles for negation or the conditional, we would arrive at paradoxes (the Liar and Curry paradoxes, respectively). Hence, how do we know that adding the compositional principles for conjunction and disjunction do not create new paradoxes? (Think of simple paradoxical sentences such as '1 + 1 = 2 and this whole sentence is false'.) Read offers an argument on behalf of Bradwardine showing that the compositional principles for conjunction and disjunction can be proved from the basic principles of logic that Bradwardine accepts. He defends this solution as attractive because it "preserves those truth principles which are unaffected by the paradoxes, without sacrificing any logical principles".

1.2.6.2 "Vagueness, Truth and Permissive Consequence" by Pablo Cobreros, Paul Egré, David Ripley and Robert van Rooij

In their contribution, Cobreros, Égré, Ripley and van Rooij adopt a non-standard notion of logical consequence in order to provide an adequate semantics for vague predicates as well as for the truth predicate. A predicate is vague when it satisfies the principle of *tolerance*: sufficiently small variations in two objects cannot make

a difference in the application of the predicate. For instance, if one person is only 1 cm shorter than another person, then either both of them are tall or neither of them is. The principle of tolerance, together with classical logic, produces the sorites paradox: consider a series of people from someone who is clearly tall to someone who is clearly not tall, such that any person immediately following another person in the series is just 1 cm shorter. Given that the first person in the series is tall, by a repeated application of the principle of tolerance it follows that the last person in the series is also tall, contradicting the initial hypothesis that the last person is not tall. The analogous key semantic feature governing the truth predicate is the *transparency* principle which states that a sentence 'A' is intersubstitutable with 'A is true' in all extensional contexts without any cost for the validity of arguments. Transparency together with classical logic produce the Liar paradox and its kin.

There have been well-known attempts in the literature to solve these paradoxes by the use of three-valued logics that admit gappy sentences (i.e. whose truth value is neither true nor false) in order to deal with borderline cases of vague predicates or the truth predicate (Kripke 1975 and Tye 1994). Those solutions have been subjected to close scrutiny and often criticized (see, for example, Keefe 2000 and Gupta and Belnap 1993). In this paper the authors try to overcome the limitations of the standard solutions not by introducing a new three-valued scheme of interpretation (they follow the standard strong Kleene interpretation), but by modifying the definition of logical consequence. The new definition is based on two modes of assertion: *strict* and *tolerant*. In a three-valued setting, a sentence is strictly assertible provided it takes the value 1, and tolerantly assertible provided it does not take the value 0. Given the strict and tolerant standards of assertion, the authors explore the consequence relations that arise from varying these standards for premises and conclusion. They show in particular how to combine the two modes to get a relation of permissive or st-consequence consequence (strict to tolerant), which requires premises to be asserted strictly and the conclusion only tolerantly. This consequence relation combines features of both Kleene's Strong Logic (K3) and of its dual, the Logic of Paradox (LP).

The notion of permissible consequence has some desirable properties that, according to the authors, make it appropriate for providing a good solution to the paradoxes of vagueness and truth. For a language that does not contain vague predicates or the truth predicate, st-consequence coincides with that of classical logic; which also means that it has a well-behaved conditional that satisfies modus ponens and the deduction theorem. In the presence of vague predicates or the truth predicate, it also satisfies the tolerance and transparency principles. So what prevents the sorites and Liar arguments from going through is the fact that st-consequence lacks *transitivity*. Although for every object a_n in a sorites series, it follows that: if a_n is tall, then a_{n-1} is also tall, we cannot chain all these reasoning steps together to conclude that if the first object in the series is tall, the last one is too. As an example of non-transitivity in the case of truth, even though any sentence has as an st-consequence the Liar, and the Liar has as an st-consequence any sentence, explosion does not follow.

Alongside applying st-consequence to the cases of vagueness and truth, the authors discuss two concerns that are naturally raised by their proposal. On the one hand, since transitivity is a basic structural rule, it is difficult to justify how we

can do ordinary reasoning without it[40]. On the other hand, they address the standard concerns of how to solve issues of higher-order vagueness or strengthened Liar paradoxes.

With respect to the first problem, the distinction between strict and tolerant assertion comes to rescue. Transitivity still holds when we go from strictly accepted premises to strictly accepted conclusions, but fails when we get only tolerated conclusions that cannot subsequently be used as strictly accepted premises. This means that in the absence of vagueness or semantic predicates, we can safely reason transitively.

With respect to the second problem, they consider revenge paradoxes expressed with the determinateness operator: we want to say that, if an object is a borderline case of tallness, then it is neither determinately tall nor determinately not tall; and we also want to say that the Liar paradox is neither determinately true nor determinately false. In a three-valued logic, the natural determinateness operator is an operator D such that DA is true when A is a true sentence and false when A is either a false or a gappy sentence. But then if a sentence A says that an object is a borderline case of tallness, then A is gappy and, given the definition of D, the sentence $D(\neg DA \wedge \neg D\neg A)$ is true. This means that it is determinately the case that the object is a borderline case of tallness. Hence, one cannot express the existence of second-order vagueness in the language (i.e. the existence of objects that are not determinately tall but also not determinately borderline cases of tallness). In the case of truth, if D belongs to the language, then the sentence that says of itself that it is not determinately true is a new paradox that the three-valued semantics cannot consistently evaluate. The authors of the paper explore two strategies for coping with these revenge paradoxes and claim that their theory is compatible with both. The first strategy, which is the one they find most congenial to their theory, is to argue that those operators should not be included in the object language. The second strategy consists of modifying the semantics to include some versions of determinateness operators.

1.2.6.3 "Validity and Truth-Preservation" by Julien Murzi and Lionel Shapiro

In their contribution, Murzi and Shapiro address one of the unpleasant consequences of revisionary approaches to paradox, which is that some naive principles governing the intuitive notion of validity are invalidated. The authors consider the standard definition of validity as truth-preservation, so-called VTP: 'If an argument is valid, then, if all its premises are true, then its conclusion is also true'. Validity relies on truth according to VTP, so it is to be expected that attempts to secure a consistent notion of truth may have consequences for VTP. The authors first rehearse the reasons why VTP has come to be unpopular among revisionary theorists. First, invalidating

[40] Recall Feferman's relevant dictum which in their contribution to this volume Halbach and Horsten present as one of the desiderata for type-free theories of truth.

VTP is a corollary of the standard revisionary approach to the semantic paradoxes of truth in that revisionary approaches typically aim to preserve naive properties of truth (crucially, the T-schema and intersubstitutivity) and propose a conditional to replace the so-called detaching conditional (i.e. one that satisfies MP), in order to deal with the Liar paradox. In order to secure a uniform approach to the semantic paradoxes, this new conditional is typically chosen to be such that it also blocks the derivation of the c-Curry paradox (c for conditional). This is achieved by a conditional that does not satisfy the $I \rightarrow$ rule, thereby invalidating the law of contraction, which is widely held responsible for the c-Curry paradox. But without $I \rightarrow$, VTP is also invalidated, for example, in its simplest reconstruction offered by Field; the so-called Validity Argument (Field 2008).

Independent arguments have also been put forward to show that VTP cannot be consistently asserted in the object language. The authors believe that VTP is a factive statement, and that a semantic theory should be able to affirm what we know to be true. They thus counterpose their own independent arguments in defence of VTP. These come down to two main points. First, the authors believe that a 'naive view of semantic properties' is in the spirit of the revisionary approach to semantic paradox in general and not only for truth. To make the 'naive view of validity' precise, the authors list two principles that underwrite it: the so-called VP and VD, which resemble the necessitation rule and the **T** axiom for the necessity operator respectively. Their claim is that just as dealing with the Liar paradox does not call for revision of the semantic properties of truth by, for instance, invalidating the T-schema, so the revisionist should not abolish the naive VP and VD either when seeking a way out of the paradoxes of validity. Second, the standard revisionary way of dealing with the Curry paradox, rejecting the operational rule of $I \rightarrow$, does not apply in the case of the v-Curry paradox (v for validity), since the rule is not used in the derivation. In contrast, the structural rule of contraction *is* used and since this paradox of validity is as genuine a paradox as c-Curry is, a uniform solution to both versions of Curry's paradox naturally calls for dispensing with contraction.

The authors proceed to inquire into the consequences of their proposal; that is, the consequences that the rejection of contraction has on the main challenges to VTP that they have identified. They show that these challenges to VTP rely on the way premise aggregation is represented for multiple-premise arguments. Moreover, they show that rejecting contraction opens up different non-classical options to represent premise aggregation (as in multi-based logic with multiplicative conjunction or dual-bunching logic). The authors then offer a detailed analysis of where the arguments underlying the challenges to VTP break down with respect to different substructural choices. Rather than arguing in favour of one such choice, the authors' aim is to establish that VTP is not incompatible with a revisionary approach to paradox; especially since rejecting contraction is no obstacle to supporting a naive theory of truth.

1.2.6.4 "Getting One for Two, or the Contractors' Bad Deal. Towards a Unified Solution to the Semantic Paradoxes" by Elia Zardini

In his contribution, Zardini proposes a new solution to the Liar paradox based on a substructural logic (see also Zardini 2011). He takes the transparency of the truth predicate to be a requirement of any acceptable theory of truth and argues that the basic law that has to go in order to restore the consistency of a truth predicate in a self-referential language is the structural rule of contraction. In its most basic form, that rule says that, given sentences ϕ, ψ, if $\phi, \phi \vdash \psi$, then $\phi \vdash \psi$. The version of contraction-free transparent theory of truth Zardini uses is called IKT and is presented in sequent calculus style. The paper compares the paracomplete and paraconsistent treatment of the paradoxes (that reject LEM—the Law of Excluded Middle—and LNC—the Law of Non-Contradiction—respectively) with IKT, and argues that the latter offers an important advantage over its competitors: it provides a unified solution to different paradoxical arguments, while paracomplete and paraconsistent solutions need to concoct different modes of justification to solve different paradoxical arguments.

The bulk of the paper is devoted to presenting three paradoxical arguments followed by a detailed analysis of the diagnosis that the different theories could give of them. The first two arguments use the Liar, i.e. a sentence λ identical to $\neg T(\ulcorner \lambda \urcorner)$, to produce unacceptable (even for a dialetheist) conclusions. In the first argument both LEM and LNC are used, so paracomplete and paraconsistent solutions have a principled way to reject it. The second argument, however, does not use LEM. This would force the paracomplete theorist to reject another rule, typically the metarule of the single-premise reduction theorem (if $\phi \vdash f$, then $t \vdash \neg\phi$, where t expresses the conjunction of all logical truths and f the disjunction of all logical falsehoods). Zardini gives a detailed analysis of the justification for this rule to show that its rejection does not follow from the denial of LEM. A parallel of this dialectics is presented, offering a version of the second argument that does not use LNC and that would force the dialetheist to reject the metarule of the single-conclusion demonstration theorem (if $t \vdash \phi$, then $\neg\phi \vdash f$) for reasons that do not follow from the denial of LNC. The third form of argument analysed in the paper is Curry's paradox, which uses neither LEM nor LNC. In this case, paracomplete and paraconsistent theorists would have to reject the single-premise deduction theorem (if $\phi \vdash \psi$, then $t \vdash \phi \rightarrow \psi$), but only for reasons that do not follow from the denial of either LEM or LNC. In contrast to paracomplete and paraconsistent theories, IKT accepts as valid all the principles dismissed by them and offers a unified solution: reject contraction, which is used in all three arguments. Of course this advantage by itself does not show why contraction fails. Although this question is not tackled here, the paper ends by giving a few hints of an answer that relies on a picture of metaphysical reality as unstable. Zardini considers that contraction fails for sentences that express unstable states of affairs, meaning states of affairs that lead to consequences with which they need not co-obtain.

1.2.6.5 "Kripke's Thought-Paradox and the 5th Antinomy" by Graham Priest

As a consequence of the central role that the Liar paradox occupies in the philosophy of truth, the reasoning form of the paradox has also become an object of study in its own right, independent of underlying theories. Since it can be classified as a self-referential paradox, a way to further our understanding of it is to study whether there are features shared by self-referential paradoxes which license a common description of them. The so-called Inclosure Schema has been proposed by Priest (2002) as such a description. The schema is a reworking of Russell's description of set-theoretic paradoxes given in 1905: 'On some difficulties in the theory of transfinite numbers and order types' (Russell 1906). Priest generalizes Russell's description to self-referential paradoxes (which, for example, include the Liar and König's paradox). He is careful to note that not all reasoning that shares this underlying form should be considered as paradoxical (that is, the schema should not be read as laying out sufficient conditions); this can already be seen in his discussion of the Barber paradox in his (2002), but Curry's paradox also presents a different challenge. What the extra conditions are for such self-referential reasoning to count as paradoxical is still under discussion, and as a consequence, so is the meta-theoretic status of the schema (for instance, whether it can be seen as playing a heuristic role).

Priest's interest in having a general schema that delineates a class of paradoxes is strongly linked to his belief that paradoxical arguments that share the same form call for a uniform solution. He introduced this idea in his (1994) under the name of the 'Principle of Uniform Solution'. In his contribution to this volume, both the Inclosure Schema and the Principle of Uniform Solution are put to the test by considering a new intensional self-referential paradox, namely, Kripke's thought paradox. In the first part of the paper Priest shows how the Inclosure Schema can accommodate that paradox and in the second part he argues for a dialetheist solution to it. Since the same solution can be given for the Liar paradox, this paper contributes, on the one hand, to a uniform representation and solution of self-referential paradoxes; as well as providing an additional argument for dialetheism, on the other. We saw that the preceding papers depart from the specific paraconsistent approach as the solution to the paradoxes. Yet, Priest's work constitutes a systematic expression of the philosophical idea of unification that drives most of the work in this area of revisionary approaches to truth.

Acknowledgements We would like to thank Springer and especially Shahid Rahman for hosting this collection of papers within the series 'Logic, Epistemology and the Unity of Science'; we are thankful to the editorial office of Springer for all their work, and especially Christi Lue and Rajdeep Crest Roy. We are most grateful to the authors of the papers in this volume for honouring us with their contributions, as well as their patience, cooperation and trust while we have been preparing this volume. Finally, we are grateful to all those who have acted as anonymous referees for their careful reading and valuable feedback.

Theodora Achourioti would like to thank Peter van Ormondt for his help with the organization of the conference 'Truth be told' (Amsterdam, 23–25 March 2011). She would also like to acknowledge the generous financial and technical support offered by: the NWO open competition

project ‘The Origins of Truth and the Origins of the Sentence’; the Institute for Logic, Language and Computation; the Evert Willem Beth Foundation; the NWO VICI project ‘Unsupervised Learning with the DOP Model’; and the Philosophy Department of the University of Amsterdam.

Kentaro Fujimoto would like to thank Volker Halbach as co-organizer of the “Axiomatic Theories of Truth” conference (Oxford, 19–20 September 2011) as well as the other members of the AHRC project “Inexpressibility and Reflection in the Formal Sciences” for their cooperation; and also to acknowledge financial support for the conference by the Arts and Humanities Research Council (AH/H039791/1).

Henri Galinon would like to thank Pierre Wagner and Denis Bonnay as co-organizers of the “Truth at Work” conference (20–23 June 2011), the members of the Institut d’Histoire et de Philosophie des Sciences et des Techniques for their help in organizing the conference, as well as Sebastien Gandon and the PHIER for their support during the preparation of this volume. The organization of the conference was financially supported by the ANR-funded project ‘Logiscience’ managed by Pierre Wagner at the IHPST.

José Martínez would like to thank Genoveva Martí and Sergi Oms as co-organizers of the “BW7 Conference: Paradoxes of Truth and Denotation” (14–16 June 2011) as well as the other members of the Logos Research Group for their cooperation. He would also like to acknowledge funding received from the Spanish *Ministerio de Economía y Competitividad* for project FFI2010-11447-E, R+D Project FFI2011-25626 (Reference, Self-reference and Empirical Data) and the Consolider Ingenio Program, CSD2009-00056, (Perspectival Thoughts and Facts); and from the *Generalitat de Catalunya* for project 2010ARCS10160.

References

Aczel, P. (1980). Frege structures and the notions of proposition, truth and set. In J. Barwise, H. Keisler, & K. Kunen (Eds.), *The Kleene symposium* (pp. 31–59). Amsterdam: North-Holland.

Barendregt, H. (1984). *The Lambda calculus, its syntax and semantics (studies in logic and the foundations of mathematics, volume 103)*. Amsterdam: Elsevier.

Beeson, M. (1985). *Foundations of constructive mathematics*. Berlin: Springer.

Boolos, G. (1993). *The logic of provability*. Cambridge: Cambridge University Press.

Bourget, D., & Chalmers, D. (2014). What do philosophers believe. *Philosophical Studies,170*(3), 465–500.

Brandom, R. (1994). *Making it explicit. Reasoning, representing, and discursive commitment*. Cambridge: Harvard University Press.

Burgess, J. P. (2009). Friedman and the axiomatization of Kripke’s theory of truth. Paper delivered at the Ohio State University conference in honor of the 60th birthday of Harvey Friedman.

Cantini, A. (1990). A theory of formal truth arithmetically equivalent to ID1. *The Journal of Symbolic Logic, 55,* 244–259.

Cantini, A. (1996). *Logical frameworks for truth and abstraction*. Amsterdam: Elsevier.

Carnap, R. (1931). Überwindung der Metaphysik durch logische Analyse der Sprache. *Erkenntis*, *2*(1), 219–241.

Cieśliński, C. (2007). Deflationism, conservativeness and maximality. *Journal of Philosophical Logic, 36,* 695–705.

Dummett, M. (1959). Truth. *Proceedings of the Aristotelian Society, 59,* 141–162.

Eberhard, S. (2013). Weak applicative theories, truth, and computational complexity. PhD Thesis. University of Bern.

Eberhard, S., & Strahm, T. (2012). Weak theories of truth and explicit mathematics. In U. Berger, H. Diener, P. Schuster, & M. Seisenberger (Eds.), *Logic, construction, computation* (pp. 157–184). Berlin: De Gruyter.

Eklund, M. (2001). Inconsistent languages. *Philosophy and Phenomenological research, 64*(2), 251–276.

Feferman, S. (1984). Towards useful type-free theories I. *Journal of Symbolic Logic, 49,* 75–111.
Feferman, S. (1989). Kurt Gödel : Conviction and caution. In S. G. Shanker (Ed.), *Gödel's theorems in focus* (pp. 96–115). London: Routledge.
Feferman, S. (1991). Reflecting on incompleteness. *The Journal of Symbolic Logic, 56,* 1–49.
Feferman, S. (1996). Gödel's program for new axioms: Why, where, how and what? In P. Hájek (Ed.), *Gödel '96*, volume 6 of *Lecture notes in logic* (pp. 3–22). Berlin: Springer.
Feferman, S. (2008). Axioms for determinateness and truth. *Review of Symbolic Logic, 1*(2), 204–217.
Feferman, S., & Strahm, T. (2000). The unfolding of non-finitist arithmetic. *Annals of Pure and Applied Logic, 104*(1–3), 75–96.
Feferman, S., & Strahm, T. (2010). Unfolding finitist arithmetic. *Review of Symbolic Logic, 3*(4), 665–689.
Field, H. (1972). Tarski's theory of truth. *Journal of Philosophy, 69,* 347–375.
Field, H. (1980). *Science without numbers: A defense of Nominalism*. Princeton: Princeton University Press.
Field, H. (2001). *Truth and the absence of fact*. Oxford: Oxford University Press.
Field, H. (2008). *Saving truth from paradox*. Oxford: Oxford University Press.
Fodor, J. A. (1989). *Psychosemantics: The problem of meaning in the philosophy of mind.* Cambridge: MIT Press.
Frege, G. (1918). Der gedanke: Eine logische untersuchung. *Beitrge zur Philosophie des Deutschen Idealismus, 1,* 58–77.
Friedman, H., & Sheard, M. (1987). An axiomatic approach to self-referential truth. *Annals of Pure and Applied Logic, 33,* 1–21.
Fujimoto, K. (2010). Relative truth definability of axiomatic theories of truth. *The Bulletin of Symbolic Logic, 16,* 305–344.
Fujimoto, K. (2012). Classes and truths in set theory. *Annals of Pure and Applied Logic, 163,* 1484–1523.
Garey, M., & Johnson, D. (2002). *Compters and intractability: A guide to the theory of NP-completeness*. New York: Freeman.
Grover, D. (1992). *A prosentential theory of truth*. Princeton: Princeton University Press.
Grover, D., Camp, J., & Belnap, N. (1975). A prosentential theory of truth. *Philosophical Studies, 27*(2), 73–125.
Gupta, A., & Belnap, N. (1993). *The revision theory of truth*. Cambridge: MIT.
Hájek, P., & Pudlak, P. (1993). *Metamathematics of first-order arithmetic*. Berlin: Springer.
Halbach, V. (1994). A system of complete and consistent truth. *Notre Dame Journal of Formal Logic, 35,* 311–327.
Halbach, V. (2010). *Axiomatic theories of truth*. New York: Cambridge University Press.
Heck, R. (2011). The strength of truth theories. (2011).
Hilbert, D. (1926). Über das unendliche. *Mathematische Annalen, 95,* 161–190.
Hindley, R., & Seldin, J. (2008). *Lambda calculus and combinators: An introduction*. Cambridge: Cambridge University Press.
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* (pp. 173–187). Tilburg: Tilburg University Press.
Horsten, L. (2011). *The Tarskian turn: Deflationism and axiomatic truth*. Cambridge: MIT Press.
Horwich, P. (1998a). *Meaning*. New York: Oxford University Press.
Horwich, P. (1998b). *Truth*. New York: Oxford University Press.
Horwich, P. (2006). The value of truth. *Nous, 40*(2), 347–360.
Kahle, R. (1999). Frege structures for partial applicative theories. *Journal of Logic and Computation, 9*(5), 683–700.
Kahle, R. (2001). Truth in applicative theories. *Studia Logica, 68*(1), 103–128.
Kahle, R. (2003). Universes over Frege structures. *Annals of Pure and Applied Logic, 119*(1–3), 191–223.

Kahle, R. (2009). The universal set—A (never fought) battle between philosophy and mathematics. In O. Pombo & Á. Nepomuceno (Eds.), *Lógica e Filosofia da Ciência*, volume 2 of Colecção Documenta (pp. 53–65). Lisboa: Centro de Filosofia das Ciências da Universidade de Lisboa.

Kahle, R. (2011). The universal set and diagonalization in Frege structures. *Review of Symbolic Logic, 4*(2), 205–218.

Kaye, R. (1991). *Models of Peano arithmetic*. Oxford: Clarendon Press.

Keefe, R. (2000). *Theories of vagueness*. Cambridge: Cambridge University Press.

Ketland, J. (1999). Deflationism and Tarski's paradise. *Mind, 108,* 69–94.

Kotlarski, H., Krajewski, S., & Lachlan, A. (1981). Construction of satisfaction classes for nonstandard models. *Canadian Mathematical Bulletin, 24,* 283–293.

Krajewski, S. (1976). Non-standard satisfaction classes. In W. Marek, M. Srebrny, & A. Zarach (Eds.), *Set theory and hierarchy theory*, Lecture notes in mathematics 537 (pp. 121–144). Berlin: Springer.

Kreisel, G. (1970). Principles of proof and ordinals implicit in given concepts. In A. Kino, J. Myhill, & R. Vesley (Eds.), *Intuitionism and proof theory* (pp. 489–503). Amsterdam: North-Holland.

Kripke, S. (1975). Outline of a theory of truth. *Journal of Philosophy, 72,* 690–716.

Lachlan, A. (1981). Full satisfaction classes and recursive saturation. *Canadian Mathematical Bulletin, 24,* 295–297.

Leeds, S. (1978). Theories of reference and truth. *Erkenntis, 13*(1), 111–129.

Leigh, G. (2013). A proof-theoretic account of classical principles of truth. *Annals of Pure and Applied Logic, 164,* 1009–1024.

Leigh, G., & Nicolai, C. (2013). Axiomatic truth, syntax and metatheoretic reasoning. *Review of Symbolic Logic, 6*(4), 613–636.

Leigh, G., & Rathjen, M. (2010). An ordinal analysis for theories of self-referential truth. *Archive for Mathematical Logic, 49,* 213–247.

Leigh, G., & Rathjen, M. (2012). The Friedman-Sheard programme in intuitionistic logic. *Journal of Symbolic Logic, 77,* 777–806.

Leitgeb, H. (2005). What truth depends on. *Journal of Philosophical Logic, 34,* 155–192.

Loewer, B. (1997). A guide to naturalizing semantics. In B. Hale & C. Wright (Eds.), *A companion to the philosophy of language* (pp. 108–126). Oxford: Blackwell Publishers.

Lynch, M. (2009). *Truth as one and many*. Oxford: Oxford University Press.

Mancosu, P. (2008a). Quine and Tarski on nominalism. In D. Zimmerman (Ed.), *Oxford studies in metaphysics* (Vol. 4). New York: Oxford University Press.

Mancosu, P. (2008b). Tarski, Neurath, and Kokoszynska on the semantic conception of truth. In D. Patterson (Ed.), *New essays on Tarski and philosophy*. New York: Oxford University Press.

Mancosu, P. (2009). Tarski's engagement with philosophy. In S. Lapointe, J. Wolenski, M. Marion, & W. Miskiewicz (Eds), *The golden age of Polish philosophy*, volume 16 of *Logic, epistemology, and the unity of science* (pp. 131–153). Amsterdam: Springer Netherlands.

Maudlin, T. (2004). *Truth and paradox*. Oxford: Oxford University Press.

McGee, V. (1992a). Maximal consistent sets of instances of Tarski's schema (T). *Journal of Philosophical Logic, 21,* 235–241.

McGee, V. (1992b). Maximal consistent sets of instances of Tarski's schema (t). *Journal of Philosophical Logic, 21,* 235–241.

McGee, V. (2006). In praise of the free lunch: Why disquotationalists should embrace compositional semantics. In T. Bolander, V. F. Hendricks, & S. A. Pedersen (Eds.), *Self-reference* (pp. 95–120). Stanford: CSLI Publications.

Moltmann, F. (2003). Nominalizing quantifiers. *Journal of Philosophical Logic, 32,* 445–481.

Moltmann, F. (2013). *Abstract objects and the semantics of natural language*. Oxford: Oxford UP.

Mulligan, K. (2010). The truth predicate vs the truth connective. On taking connectives seriously. *Dialectica, 64,* 565–584.

Nicolai, C. (forthcoming). Deflationism and the Ontology of Expressions: An Axiomatic Study. DPhil Thesis, University of Oxford.

Patterson, D. (2008). Understanding the liar. In J. C. Beall (Ed.), *Revenge of the liar: New essays on the paradox* (pp. 387–422). New York: Oxford University Press.
Patterson, D. (2009). Inconsistency theory of semantic paradox. *Philosophy and Phenomenological Research, 79*(2), 387–422.
Priest, G. (1994). The structure of the paradoxes of self-reference. *Mind, 103*(409), 25–34.
Priest, G. (2002). *Beyond the limits of thought*. Oxford: Oxford UP.
Quine, W. v. O. (1960). *Word and object*. Cambridge: MIT Press.
Quine, W. v. O. (1970). *Philosophy of logic*. Cambridge: Harvard University Press.
Quine, W. v. O. (1976). *The ways of paradox, and other essays*. Cambridge: Harvard University Press.
Quine, W. v. O. (1990). *The pursuit of truth*. Cambridge: Harvard Unviersity Press.
Ramsey, F. (1991). On truth. *Episteme, 16,* 1–16.
Rouilhan, Ph. de. (2012). In defense of logical universalism: Taking issue with Jean van Heijenoort. *Logica Universalis, 6*(3–4), 553–586.
Russell, B. (1906). On some difficulties in the theory of transfinite numbers and order types. *Proceedings of the London Mathematical Society,* series *2*(4), 29–53.
Shapiro, S. (1983). Conservativeness and incompleteness. *Journal of Philosophy, 80*(9), 521–531.
Shapiro, S. (1998). Proof and truth: Through thick and thin. *Journal of Philosophy, 95*(10), 493–521.
Sheard, M. (1994). A guide to truth predicates in the modern era. *Journal of Symbolic Logic, 59,* 1032–1054.
Sher, G. (1991). *The bounds of logic*. Cambridge: MIT Press.
Tarski, A. (1983). The concept of truth in formalized languages. In J. Corcoran (Ed.), *Logic, semantics, metamathematics* (2nd ed., pp. 152–278). Indianapolis: Hackett.
Tarski, A. (1986). What are logical notions? *History and Philosophy of Logic, 7,* 143–154.
Tarski, A., & Vaught, R. (1956). Arithmetical extensions of relational systems. *Compositio Mathematica, 13,* 81–102.
Tye, M. (1994). Sorites paradoxes and the semantics of vagueness. In T. E. Tomberlin (Ed.), *Philosophical perspectives 8: Logic and language* (pp. 189–206). Atascadero: Ridgeview.
Welch, P. (2003). On revision operators. *Journal of Symbolic Logic, 68,* 689–711.
Wolenski, J. (2009). The rise and development of logical semantics in Poland. In S. Lapointe, J. Wolinski, M. Marion, & W. Miskiewicz (Eds.), *The golden age of polish philosophy*, volume 16 of *Logic, epistemology, and the unity of science*. Dordrecht: Springer.
Wolenski, J., & Murawski, R. (2008). Tarski and his Polish predecessors on Truth. In D. Patterson (Ed.), *New essays on Tarski and philosophy*. Oxford: Oxford University Press.
Zardini, E. (2011). Truth without contra(di)ction. *Review of Symbolic Logic,* 4, 498–535.

Part I
Truth and Natural Language

Chapter 2
'Truth Predicates' in Natural Language

Friederike Moltmann

Abstract The aim of this paper is to take a closer look at the actual semantic behavior of what appear to be truth predicates in natural language and to re-assess the way they could motivate particular philosophical views. The paper will draw a distinction between two types of apparent truth predicates: type 1 truth predicates such as in English *true* and *correct* and type 2 truth predicates such as English *is the case*. It will establish the following points:

1. Type 1 truth predicates are true predicates, predicated of a representational objects of some sort, such as sentences, propositions, and entities of the sort of beliefs and assertions.
2. *That*-clauses with type 1 truth predicates do not act as referential terms, referring to propositions as truth bearers, but rather specify the content of contextually given attitudinal objects, such as 'John's belief that S' or 'Mary's claim that S'.
3. Type 2 'truth predicates' do not in fact act as truth predicates, but rather express the relation of truthmaking, relating a situation or 'case' to the content of a *that*-clause.

2.1 Introduction

The notion of truth has given rise to a great variety of philosophical views. Some of those views have been motivated by appeal to natural language, in particular the linguistic status and semantic behavior of what appear to be truth predicates, such as *true* in English. The aim of this paper is to take a closer look at the actual semantic behavior of apparent truth predicates in natural language and to re-assess the way they could motivate particular philosophical views.

True is not the only expression in English acting as an apparent truth predicate, and the paper will discuss other expressions in English (and German) as well that appear to convey truth. They include *correct* and *right* (in some of their uses), as well as *is the case*. Note that I call the relevant expressions 'apparent truth predicates' since it is

F. Moltmann
IHPST (Paris 1/ ENS/ CNRS), Paris, France
e-mail: fmoltmann@univ-paris1.fr

T. Achourioti et al. (Eds.), *Unifying the Philosophy of Truth,* Logic, Epistemology, and the Unity of Science 36, DOI 10.1007/978-94-017-9673-6_2

controversial whether they really act as predicates predicating truth, and in fact only one among two types of expressions that I will discuss turns out to consist in actual truth predicates. The label 'truth predicate' will just serve to simplify the discussion.

The paper argues for a sharp distinction among two types of apparent truth predicates. *Is true* belongs to what I will call 'type 1 truth predicates'; *is the case* belongs to what I will call 'type 2 truth predicates':

Type 1 truth predicates
(1) That S is true.

Type 2 truth predicates
(2) That S is the case.

The paper will also discuss *right* and *correct* as type 1 truth predicates, since *right* and *correct* despite their more general normative meaning convey truth with *that*-clauses as well as certain noun phrases. Type 2 truth predicates in English also include the predicates *occur* and *is so*.

The paper will establish several important points about the two types of truth predicates:

1. Type 1 truth predicates cannot be viewed as 'operators', 'connectives', or 'anaphoric devices', as has been claimed in some of the philosophical literature. Rather they are predicates predicated of a representational object of some sort. Such representational objects include sentences, propositions, beliefs, and assertions.
2. *That*-clauses with type 1 truth predicates do not act as referential terms, referring to a proposition as the truth bearer. Rather they have the function of specifying the content of a contextually given representational object. With *correct* and *right*, such objects cannot be sentences or propositions, but must be mind-dependent objects of the sort 'John's belief that S' or 'Mary's claim that S' (or perhaps kinds of such objects), that is what I call 'attitudinal objects' (Moltmann 2003a, 2013).[1] There is evidence that the same holds for *true*.

Point 2 is important since it would mean that propositions as abstract, mind-independent objects are not involved in the semantics of *that*-clauses with type 1 truth predicates, as on deflationist or 'modest' accounts of truth. Instead it opens the door for philosophical views tying the notion of truth primarily to the intentionality of mind-dependent objects, such as beliefs and claims.

Furthermore, the use of normative predicates such as *correct* to convey truth is at least compatible with a view according to which truth as constitutive of the normativity of beliefs (and related attitudinal objects) is a notion prior to the notion of truth applicable to sentences and propositions. The semantic behavior of other type 1 truth predicates that will be discussed is particularly suggestive of such a view.

[1] Note that attitudinal objects are not mental or illocutionary acts. They differ, most importantly, in that they have truth- or satisfaction conditions. They are thus proposition-like, but yet mind- and agent-dependent, see Moltmann (2003b, 2013).

3. Type 2 'truth predicates' do not in fact act as truth predicates, but rather express the relation of truth-making, relating a situation, or rather a 'case', to the content of a *that*-clause. Natural language thus does not just reflect the notion of truth, but also that of truth-making.

The appendix will briefly discuss two other apparent truth predicates in English, namely *is a fact* and *is the truth*. It will argue that they involve a more complex syntactic structure than what is apparent and do not serve to predicate truth.

2.2 Type 1 Truth Predicates

2.2.1 Basic Properties of Type 1 Truth Predicates

Is true is the truth predicate in English that has received the most philosophical attention. But there are other truth predicates that behave in relevant respects alike and thus classify as type 1 truth predicates. In particular, the normative predicates *is correct* and *is right* act as truth predicates when applied to *that*-clauses. We will later see that taking into account such predicates will be important for understanding truth predication in general.

In what follows, I will not discuss particular philosophical views about the status of *is true* in detail, but restrict myself to discussing the adequacy of a number of assumptions or claims that have been made in the philosophical literature about the linguistic status of *is true.*

Let us start with some very general linguistic properties of *is true*.[2] First, *is true* allows both for clausal subjects, as in (3a), and for extraposition, as in (3b):

(3) a. That the sun is shining is true.
b. It is true that the sun is shining.

Moreover, *is true* allows for certain quantifiers and pronouns in place of *that*-clauses in subject position, such as *everything* and *that*:

(4) a. Everything is true.
b. That is true.

This does not hold for extraposed clauses, though (as is always the case, whatever the predicate):

(5) a. It is true that S.
b. * It is true everything/that.

The reason is that noun phrases (NPs) can never appear in that position.

Another important fact about *is true* is that it allows for referential NPs in subject position, namely NPs referring to entities such as propositions, sentences, beliefs, or claims:

(6) a. The proposition that S is true.

[2] As can easily be verified, the negative truth predicate *is false* exhibits the very same properties.

b. The sentence 'S' is true.
c. John's belief that S is true.
d. John's claim that S is true.

Related to that is the (often overlooked) fact that *true* can act as an adnominal modifier of those same NPs:

(7) a. the true proposition that S
b. the true sentence 'S'
c. John's true claim that S

Various philosophers have developed views of the notion of truth focusing on the clausal structures in (3a, b). Some philosophers in particular have proposed views concerning the formal status of *true* in the clausal construction. Thus, Ramsey (1927) held that *is true* in that construction is simply redundant. That is, *that* S *is true*, on that view, means the very same thing as S. Grover et al. (1975) proposed that *is true* it is simply an anaphoric device. Roughly, on their view, *that is true* in the discourse context *It is raining. That is true.* is simply a device permitting re-use of the preceding sentence. Finally, there is the view according to which *is true* is an operator or connective (or part of an expression acting that way, an expression that would include *that*), a view recently defended by Mulligan (2010).

Such views all give priority to the clausal construction over the construction in which *is true* applies to a referential NP, or they focus entirely on the clausal construction. The operator/connective view of *is true* moreover gives priority to the extraposition structure. In the next sections, we will see that the assumptions that the clausal structure takes priority is untenable, as is the assumption that the extraposition structure takes priority over the subject clause structure.

2.2.2 The Priority of the Clausal Construction

True in predicate position accepts both *that*-clauses and ordinary referential or quantificational NPs in subject position, and it naturally occurs as an adjectival modifier of the latter. In general, it seems, type 1 truth predicates come both with clausal and nominal constructions, and if they involve an adjective (like *true*), the adjective will have an application as an adnominal modifier. There is no evidence for the priority of the clausal construction. Moreover, the semantic contribution of *true* appears exactly the same in the clausal and the nominal construction.

There are adjectives that like *true* can appear in predicate position with clausal subjects, but with which the clausal construction displays a distinctive 'sentential' semantics, unlike with *true*. Examples are the adjectives *possible* and *probable*. *Possible* and *probable* in predicate position behave like sentence adverbials in two respects. First, *possible* and *probable* have adverbial counterparts that act as sentence adverbials. Thus (8a) and (8b) are equivalent:

(8) a. Possibly/Probably, John will be late.

b. It is possible/probable that John will be late.

Second, the subject *that*-clause with *possible* and *probable* cannot be replaced by an explicit proposition-referring NP without change in meaning:

(9) The proposition that John will be late is possible/probable.

(9) does not mean what (8a) and (8b) mean. Rather it states that the existence of a proposition as an abstract object is possible/probable. Failure of substitution of a coreferential term is a good indication of the nonreferential status of a *that*-clause. With *possible* and *probable* as predicates, *that*-clauses do not act like singular terms referring to propositions and they do not specify the content of any object whatsoever to which *possible* and *probable* could apply as predicates. Rather, in predicate position, *possible* and *probable* appear to retain the very same semantic function that they have when acting as sentence adverbials.

The same two diagnostics for a sentence-adverbial function of an adjective in predicate position fail to apply to type 1 truth predicates. First of all, *true* lacks a sentence-adverbial counterpart, though it has an adverbial counterpart modifying the VP:

(10) a. John truly believes that S.
b. John truly said that S.

Given the common Davidsonian analysis of VP adverbials, *truly* here acts as a predicate, namely of the Davidsonian event argument of *believe* and *say*—or an entity closely related to it, an 'attitudinal object' of the sort of a belief or a claim, just as *true* does in (11a, b):[3]

(11) a. John's belief that S is true.
b. John's claim that S is true.

Furthermore, with *true*, a replacement of the subject clause as in (12a) by a referential NP referring to a proposition (or another suitable object) as in (12b) is generally possible:

(12) a. That S is true.

---

[3] For a discussion of uses of *truly* as in (10a, b) see Aune (1967). In Moltmann (2013, Chap. 4), I argue that acts and states, such as a 'John's act of claiming' or 'John's state of believing' do not have truth conditions; only the corresponding attitudinal objects do, that is, entities of the sort 'John's claim' or 'John's belief'. This may be a problem for the Davidsonian account of *truly* in (10). The Davidsonian account appears problematic anyway, though, because the adverbial use of *truly* as in (10) does not seem to be available in all languages. For example, it is not available in German, which lacks an adverbial form of *wahr* 'true' with the right meaning. German has the adverbial form of *richtig* 'correctly'. But as an adverbial *richtig* cannot convey truth. Thus, (ia) is impossible, even though *richtig* can act as an adverbial with other predicates, as in (ib) and (ic):

(i) a.??? Hans glaubt richtig, daβ es regnet.
'John believes correctly that it is raining.'
b. Hans hat das Wort richtig geschrieben.
'John has written the word correctly.'
c. Hans hat das Wort richtig verwendet.
'John used the word correctly.'

b. That proposition that S is true.

There is another important difference between *probable*/*possible* and type 1 truth predicates. With *that*-clauses, *true* displays an *anaphoric effect* that *possible* and *probable* don't. Thus, (13) suggests that *that* S has been claimed or considered by someone in the context of discourse, whereas this is not the case for (14a, b):[4]

(13) It is true that Mary is guilty.
(14) a. It is possible that Mary is guilty.
b. It is probable that Mary is guilty.

Unlike (13), (14a, b) are perfectly acceptable in a context in which 'Mary is guilty' was not under discussion or has not been entertained by anyone. The anaphoric effect indicates that *is true* is in fact predicated of a contextually given attitudinal object, let's say a claim, supposition or 'acceptance'.

There is also a somewhat weaker effect than a strictly anaphoric one that *is true* and *is correct* may convey, and that is a concessive effect. That is, (13) may just concede that Mary is guilty (continuing then with *but* ....), without requiring that to have been maintained by someone in the context. This effect can be considered a special case of the anaphoric effect, involving accommodation rather than a link to the previous context of conversation. That is, it will involve adding a suitably general kind of attitudinal object to the 'common ground', of the sort 'the thought that Mary is guilty' (which need not require a particular agent to actually have entertained the thought that Mary is guilty). On this use, *is true* is predicated of a hypothetical supposition, which amounts to an act of conceding.

To summarize then, there *are* predicates allowing for *that*-clauses in English that display a distinctive 'sentential' semantics, but type 1 truth predicates do not belong to them.[5]

2.2.3 Modifiers of Type 1 Truth Predicates

A further argument against the priority of the clausal construction is the interpretation of modifiers of type 1 truth predicates. Modifiers such as *partly* and *to some extent* are equally applicable with *that*-clauses and with NPs in subject position:

(15) a. That the students are intelligent is partly true.

---

[4] An anaphoric effect is also noticeable with *is possible* and *is probable* when the *that*-clause is in subject position:

(i) That John is inexperienced is possible/probable.

A plausible explanation is that *that*-clauses do not actually occur in subject position, but only in topic position (Koster 1978) (see also Fn11).

[5] Note that subject clauses with *possible* and *probable* allow for a replacement by *everything* or *that*, an indication that such quantifiers and pronouns do not go along with a referential function of the *that*-clause. See also Fn 10.

b. It is partly true that the students are intelligent.

(16) a. That John is incompetent is to some extent true.
b. It is to some extent true that John is incompetent.

The modifiers *partly* and *to some extent* are modifiers that relate to the part-whole structure of the object of which the predicate is predicated, in this case the content, in a suitable sense, of the *that*-clause.[6] The semantics of such modifiers is hard to account for on a 'sentential' semantic analysis of the *is true*-construction.

2.2.4 The Apparent Priority of the Extraposed Form and the Referential Status of the Subject Clause

Since extraposed clauses cannot be replaced by quantifiers or anaphora, the extraposition structure seems to reflect the logical form of a sentence in which *true* plays the role of an operator or connective, rather than acting as a predicate (Mulligan 2010).[7,8]

There is not much linguistic support for the extraposition structure being prior to the subject-clause structure, however. First of all, extraposition is always possible with (one-place) predicates allowing for a subject clause, regardless of the content of the predicate. This includes predicates such as *is interesting*, *is shocking,* or *was the subject of a great debate*, for which true predicative status is hardly implausible. Second, extraposition is equally available with infinitival clauses, which arguably do have the status of referential terms, unlike *that*-clauses. Unlike *that*-clauses, infinitival clauses can 'flank the identity sign', one of Frege's criteria for referential terms:

(17) a. * That John lives is that John works.
b. To live is to work.

Infinitival clauses arguably stand for action types.[9] As such, they are replaceable by explicit descriptions of actions, at least with predicates such as *correct*, *right,* and

[6] For an account of partial truth see Yablo (2014).

[7] More precisely, *true* will have to be considered part of an expression acting that way, namely *is true that* (Mulligan 2010).

[8] Sometimes *it is true (that)* cannot just be a connective, for example when it hosts tense, which may require a particular temporal interpretation, as well as temporal or modal adverbials:

(i) a. This was true.
b. This may be true.
(ii) Last year it was still true that S.

True can go along with other copular verbs than *be*:

(iii) That S became true/remained/seems true.

Thus, the view that *it is true* acts as an operator/connective may have to restrict itself to only part of the semantic function of that expression. But see the discussion in Grover et al. (1975).

[9] See, for example, Portner (1997) for such a view.

wrong. Below we see that those predicates allow for infinitival clauses both in subject position and extraposed:

(18) a. To address Mike as 'Sir' is correct.
b. It is correct to address Mike as 'Sir'.
(19) a. To take advantage of others is wrong.
b. It is wrong to take advantage of others.

(20) shows the possibility of replacing the infinitival clauses by explicit descriptions of actions:

(20) a. Actions of addressing Mike as 'Sir' are correct.
b. Actions of taking advantages of others are wrong.

Clearly then, whether a clause is in subject position or extraposed does not bear on its referential status.

2.2.5 The Referentially Dependent Status of that-Clauses

Subject clauses can be replaced by certain quantifiers and pronouns, such as *something* or *that.* However, there is evidence that *that*-clauses in subject position, as elsewhere, are not themselves referential.[10,11] In particular, in subject position *that*-clauses are not referentially independent. This is an important point, though generally not acknowledged in the semantic literature. First of all, a *that*-clause in subject position is not by itself a proposition-referring term. A *that*-clause in subject position can also stand for a fact or a possibility, and what kind of entity it stands for depends strictly on the predicate. This can be seen with the evaluative predicate *nice* below:

(21) That S is nice.

(21) allows only for a reading on which *nice* evaluates a fact, even though *nice* could in principle evaluate a proposition (as in *the proposition that* S *is nice*) or a possibility (as in *the possibility that S is nice*). Only in the presence of a suitable predicate can a *that*-clause in subject position stand for a proposition, as in (22a), or a possibility, as in (22b):

(22) a. That S/The proposition that S implies that S'.
b. That John might get elected/The possibility that John might get elected is excluded.

---

[10] Quantifiers and pronouns like *everything* and *that* themselves in fact are not indicators of the referentiality of the expression they may replace. See Moltmann (2003a, 2004, 2013) for discussion.
[11] In fact, *that*-clauses in apparent subject position, it has been argued, are actually not in subject position but rather in topic position (Koster 1978). The topic position is not a referential position, as seen below, where *really happy* appears in topic position:

(i) Really happy, he will never be.

This means that in subject position a *that*-clause does not on its own refer to a proposition, a possibility, or a fact. Rather it serves to only characterize a proposition, a possibility, or a fact depending on the predicate. The semantic role of a *that*-clause is that of specifying the content of a proposition-like object of the kind required by the predicate, an object to which the property expressed by the predicate can then apply.

The semantic role of subject clauses to only partially characterize the argument of the predicate goes along well with the account of the anaphoric effect of *is true* given earlier. With *is true*, a *that*-clause specifies the content of the relevant (contextually given or accommodated) attitudinal object (or kind of attitudinal object), of which *true* is then predicated.

2.2.6 Consequences for Deflationist View of the Content of 'True'

One general issue in the philosophical discussion of truth is the question of the status of *true* as a predicate expressing a property. On the face of it, *true* appears no different from an ordinary predicate. Deflationists deny that *true* expresses a true property, but they do not necessarily make claims about the syntactic status of *true*. Thus, Horwich's (1990) version of deflationism is sufficiently carefully formulated so as to not make direct claims about the linguistic status of *true*. The view maintains only that what constitutes having the concept of truth is the knowledge of the equivalence schema below, where [] is a nominalization function (roughly corresponding to *that*):[12]

(23) [S] is true iff S.

Yet some assumptions about the semantics of sentences with the predicate *is true* are made nonetheless. Most importantly, the account gives priority to the clausal construction: (23) is applicable only when *true* applies to a *that*-clause and not when it applies to a referential NP. (23) moreover treats a *that*-clause as a proposition-referring term (with the aim of giving justice to the possibility of replacing the *that*-clause by quantifiers like *something*, anaphoric pronouns like *that*, and descriptions of the sort *what John said*). Given (23), the application of the truth predicate amounts to the denominalization of the proposition-referring term (a *that*-clause) and the use of the sentence thus obtained.

In view of the lack of referential independence of *that*-clauses discussed in the preceding section, the deflationist view faces the problem that the subject clause by itself just could not stand for a proposition. This is not a serious problem, though, since in (23) just one particular nominalization function introducing propositions may have been chosen in the presence of the predicate *true*. More of a challenge is the anaphoric effect associated with *is true*, which indicates that it is not a proposition, but a contextually given or accommodated attitudinal object (or kind of attitudinal

[12] For a closely related view see Künne (2003).

object) that *true* is predicated of. There is further support for such an account of *that*-clauses with *true* as a type 1 truth predicate and that comes from normative truth predicates such as *correct* and *right*. With *that*-clauses, normative truth predicates simply *could* not apply to propositions.

2.2.7 Normative Truth Predicates

Correct and *right* (as well as their negative counterpart *wrong*) act like truth predicates in some of their uses, but they obviously have a more general normative meaning. They differ in that respect from *true*, which I will call a *representation-related truth predicate*. The semantic behavior of normative truth predicates is significant in that it bears both on the analysis of *that*-clauses with truth predicates in general and on the question of the priority of different notions of truth.

Correct and *right* are predicates that appear to express truth with *that*-clauses (in subject position and when extraposed):

(24) a. That John is the director is correct/right.
b. It is correct/right that John is the director.

In that role, *correct* and *right* display the very same anaphoric effect as *true*, illustrated by the contrast with *possible* and *likely* below:

(25) a. It is correct/right that John is inexperienced.
b. It is possible/likely that John is inexperienced.

Correct and *right* also act as truth predicates with referential NPs referring to attitudinal objects such as beliefs and claims:

(26) a. John's belief that S is right/correct.
b. John's claim that S is right/correct.

Correct and *right* have a more general normative meaning, though. This is what allows *correct* and *right* to apply also to decisions, punishments, movements, proofs, and conclusions:

(27) a. John's decision was right.
b. John's punishment was right.
c. The dancer's movements were correct.
d. The proof was correct.
e. The conclusion that Mary is guilty is correct.

With their more general meaning, *correct* and *right* express the fulfillment of the relevant norm (be it a moral value, a rule, an instruction, or logical validity).

The normative aspect is apparent also when *correct* and *right* are predicated of certain types of truth bearers such as explanations, and answers, in which case they do not simply predicate truth:

(28) a. The explanation that Mary was tired was correct.

b. The answer that Paris is the capital of France is correct.

For an explanation to be a correct explanation, it does not suffice for its content to be true; it also needs to explain what is to be explained. Similarly, for an answer to be a correct answer, it does not suffice for its content to be true; it also needs to respond to the question. In fact, *true* is not applicable to explanations, and in some languages, for example German, its counterpart does not apply to answers:[13]

(29) a.?? The explanation that Mary was tired was true.
b.?? Die Antwort, dass Paris die Hauptstadt von Frankreich ist ist wahr.
'The answer that Paris is the capital of France is true.'

Like *true, correct* and *right* can also be predicated of sentences:

(30) This sentence is correct/right.

However, when predicated of sentences, *correct* and *right* evaluate grammaticality rather than truth. This certainly is due to the more general normative meaning of *correct* and *right*. The norm associated with a syntactic object such as a sentence is grammaticality not truth.

An important (and related) observation is that *correct* and *right,* unlike *true,* cannot felicitously be predicated of propositions:

(31)?? The proposition that it is raining is correct/right.

This has an important consequence for the semantic analysis of *that*-clauses with *correct* and *right* as predicates, and in fact for the semantics of *that*-clauses in general. A *that*-clause with *is right* or *is correct* as predicate could not serve to specify a proposition, but only an attitudinal object of the sort of a belief or a claim—or a kind of attitudinal object in the context of a concessive use.[14]

The fact that *that*-clauses with *correct* or *right* need to specify the content of an attitudinal object (or a kind of attitudinal object) but could not stand for a proposition is rather remarkable. The more familiar cases in which a *that*-clause could not stand for a proposition are those in which the *that*-clause is required to specify a fact (*nice*) or a possibility (*is excluded*). If *that*-clauses have to specify attitudinal objects or kinds of them with *correct* and *right*, then this makes it rather plausible that

[13] In fact, a linguistic act being an answer presupposes that it addresses the question. An answer may then be 'correct' or not depending on whether its content is true. Note that the extent to which an answer addresses the question cannot be conveyed by *correct*, but only by *good*. An answer that truly addresses the question may be considered a 'good answer', whereas an answer that evades the question a 'bad' one. Obviously, an answer cannot be identified with a proposition or an assertion: propositions and assertions have different normative profiles.

[14] A concessive use involving accommodation of a kind of attitudinal object is actually harder to get with *right* and *correct* than with *true*, but it is not impossible, let's say in a suitable context with the sentence below:

(i) It is right that John is inexperienced. Yet he should be given a chance.

(ii) shows that *correct* and *right* can apply to kinds of attitudinal objects:

(ii) The claim/The assumption that John is inexperienced is correct.

they will do that with *true* as well. It would in both cases explain the anaphoric effect. Moreover, it would go along well with philosophical views that consider the primary truth bearers to be mind-dependent objects of the sort of beliefs, rather than abstract propositions, that is, it would suit well philosophical views that tie truth to intentionality.

The use of *right* and *correct* as truth predicates displays a notion of truth according to which truth is constitutive of the norm associated with beliefs: if one ought to believe p, then p (Boghossian 2003; Gibbard 2005). It is the notion of truth as the aim of belief just as the fulfillment of moral values is what certain actions and decisions should aim for. The notion of truth displayed by *true,* by contrast, is that of a property of representational objects: sentences, abstract propositions, as well as attitudinal object such as beliefs and claims.

Obviously, a deflationist account that invokes the denominalization of a proposition-referring *that*-clause is not applicable to *correct* and *right* when they are used as truth predicates with *that*-clauses. With *correct* and *right*, *that*-clauses could not stand for propositions but only for entities like beliefs and assertions. This need not be an objection to deflationism as such, though. In fact, it is not plausible that *correct* and *right* with objects like beliefs and assertions just convey truth. Rather, truth is treated as a consequence of the fulfillment of the norm associated with beliefs and assertions, a consequence of what ought to be believed or what ought to be asserted. If the content of a belief or assertion is separated from the belief or assertion itself, then the fact that the correctness of a belief or assertion implies the truth of its content is compatible with a deflationist view of truth.

There is an alternative view, however, according to which a truth-related norm is inseparable from the notion of content itself, with belief being in fact the most fundamental propositional attitudes related to the content of propositional attitudes in general (as suggested by Boghossian 2003). On such a view, the norm-related notion of truth conveyed by *correct* and *right* when applied to beliefs would not be explained in terms of the representation-related notion, but rather considered primitive, constitutive of the notion of mental content itself. The representation-related notion conveyed by *true* when predicated of sentences and abstract propositions would instead be explained in terms of the more primary normative notion. Roughly, *true* would hold of a proposition in virtue of that proposition making up the content of a (potential) belief fulfilling its norm. *True* would hold of a sentence in virtue of that sentence expressing a proposition of which *true* holds.

This is not the place for a more elaborate philosophical discussion of the two views. The purpose of the preceding remarks was mainly to clarify the philosophical options compatible with the linguistic facts. What we can certainly conclude from the linguistic facts that natural language as such does not support the priority of a notion of truth involving abstract propositions, as, for example, the deflationist account would have it.[15] The semantic behavior of representation-related and especially norm-related truth predicates does not go along well with the view that

[15] For a critique of abstract propositions as semantic values of *that*-clauses see also Boghossian (2010).

that-clauses act as proposition-referring terms. It gives much better support for the view that *that*-clauses with type 1 truth predicates serve to specify the content of contextually given or accommodated attitudinal objects (or kinds of them).

The distinction in English between the normative predicates *correct* and *right*, which can be used as truth predicates, and the representation-related truth predicate *true* raises the question of how general that distinction is. A quick look at two truth predicates of one other language (German) shows that the distinction is not very clear cut, which in turn suggests that the normative notion of truth should not be explained in terms of the representation-related one.

First, in German, there is only a single negative predicate for both the representation-related and the normative notion, namely *false*:

(32) a. Die Schluβfolgerung ist falsch.
'The conclusion is false/
b. Der Satz ist falsch.
'The sentence is false.'
c. Die Behauptung ist falsch.
'The claim is false'.
d. Die Entscheidung war falsch.
'The claim/The decision was wrong.'
e. Die Tanzschritte waren falsch.
'The dance steps were wrong.'

Interestingly, *falsch* when predicated of sentences as in (32b) is not ambiguous, but means only 'false', not 'grammatically wrong'. To convey ungrammatically requires explicitly negating *korrekt* or *richtig*:

(33) Der Satz ist nicht richtig/nicht korrekt/inkorrekt.
'The sentence is not right/not correct/incorrect.'

This indicates that *falsch* has a single meaning (which manifests itself as 'false' when applied to sentences), rather than being ambiguous between two meanings (though what exactly that meaning is remains to be spelled out).

Second, the German 'truth verb' *stimmen* 'be right', also a type 1 truth predicate, combines what appears to be a representation-related and a normative use. *Stimmen* conveys a very different norm-related notion, though, than *correct* and *right*.

Stimmen can be predicated of sentences as well as assumptions and assertions with the meaning 'true':

(34) a. Der Satz stimmt.
'The sentence is true.'
b. Maria's Annahme/Anna's Behauptung stimmt.
'Mary's assumption/Ann's claim is right'.

Stimmen also means 'true' with *that*-clauses:

(35) Daβ es regnet, stimmt.
'That it is raining is right'.

Stimmen in addition applies to certain actions or their products, conveying that they conform to the relevant rules or conditions:

(36) a. Die Tanzschritte stimmen.
'The dance steps are right.'
b. Der Beweis stimmt.
'The proof is right.'

As such *stimmen* also applies to answers, conveying that their content is true:

(37) Die Antwort stimmt.
'The answer is right.'

However, *stimmen* does not apply to actions or decisions whose associated norms are moral values:

(38) a.??? Maria's Entscheidung stimmt.
'Mary's decision is right.'
b.??? Anna's Bestrafung stimmt.
'Ann's punishment is right.'
c.??? Anderen zu helfen stimmt.
'To help others is right.'

Stimmen generally cares only about fairly local conditions and not more general action-guiding moral values. This manifests itself also in the fact that *stimmen* is perfectly natural with statements of personal taste, with which *richtig* is infelicitous, as it is with *wahr* 'true' (Kölbel 2008):

(39) a. Maria's Behauptung, dass Skifahren Spaβ macht, stimmt./?? ist richtig/??? ist wahr.
'Mary's claim that skiing is fun is right/correct/true.'
b. Daβ die Schokolade fantastisch schmeckt ist, stimmt/?? ist richtig/??? ist wahr.
'That the chocolate tastes fantastic is right/correct/true.'

The general meaning of *stimmen* appears to be that of meeting rather specific conditions associated with the entities to which *stimmen* can apply. In the case of sentences, those conditions concern truth rather than grammaticality. In the case of assertions, they concern truth as well as intersubjective sharability.

Given these observations, we can conclude that there is no strict division among representation-related and norm-related truth predicates. German *false* and *stimmen* have both representation-related and norm-related uses, and they impose rather different requirements on the norms associated with the objects they can apply to. In both cases, though, truth is treated as the norm for sentences as well as beliefs and assertions.

2.2.8 The Nominalization Truth

A further important linguistic fact about the adjective *true* is that it has a nominalization, *truth*. *Truth* can help form a relational NP that is of the very same form as a term referring to a particularized property or trope, such as *the wisdom of Socrates* or *the beauty of the landscape*.[16] In such an NP, *truth* will take as its complement a referential NP to which *true* can also apply as a predicate and an adnominal modifier such as *the proposition, the belief, the claim,* or *the sentence*:[17,18]

(40) a. the truth of the proposition
b. the truth of the belief/claim
c. the truth of the sentence

Truth otherwise displays typical occurrences as an abstract mass noun, namely as a bare mass noun apparently referring to a quality in (41a) (which is parallel to (41a')) and as a mass quantifier ranging over quality instances (or tropes) in (41b) (which is parallel to (41b')):

(41) a. The topic of the seminar was truth.[19]
a'. The topic of the conversation was beauty.
b. There was little truth in what he claimed.
b'. There was little beauty in the photograph.

---

[16] Terms of this sort are used as standard examples of trope-referring terms in the relevant philosophical literature. See Moltmann (2007, 2013) for a discussion of trope-referring terms in natural language.

[17] It is obvious from the behavior of predicates that *the truth of the proposition that* S cannot refer to the same thing as *the proposition that* S:

(i) a. The proposition that S might have been false.
b. ??? The truth of the proposition that S might have been false.

Truth is essential to 'the truth of the proposition that S', but not generally to 'the proposition that S'.

[18] Hinzen (2003) emphasizes the 'possessor' relation (the relation of inalienable possession) that is manifest in the application of the nominalization *truth*, as in (ia) and especially in (ib):

(i) a. There is some truth in his claim that S.
b. The claim that S has some truth in it.

This is the very same relation that may also apply to the referents of adjective nominalizations such as *wisdom*, where it is traditionally considered the relation of a trope to its bearer:

(ii) a. There is some wisdom in his remark.
b. His remark has some wisdom in it.

This relation is also involved in the interpretation of 'ordinary' trope-referring terms formed with *truth* or *wisdom*:

(iii) a. the truth of his claim.
b. the wisdom of his claim.

[19] Coherence theorists would consider the quality-referring term *truth* as expressing the primary notion of truth, prior to that expressed the predicate *true* or the relational use of *truth*.

Given its semantic behavior, the nominalization *truth* thus treats truth as a particularized or general quantity. If this is considered an indication as to the concept of truth itself, this poses considerable difficulties for philosophical views of truth that focus entirely on the clausal construction, such as the redundancy theory, the anaphoric theory, and the theory of *true* as (part of) a connective or operator. Natural language not only treats *true* as a property-ascribing predicate. It also treats NPs formed with its nominalization as trope-referring or trope-quantifying terms or as terms standing for a quality.

This also poses a challenge to the deflationist view. The deflationist account of *true* explains the application of *true* in terms of the use of a denominalized sentence and thus could not account for modifiers of *true* such as *partly* or *to some extent*, which relate to the part-whole structure of the content of the truth bearer.

2.3 Type 2 Truth Predicates

Is the case is a type 2 truth predicate. The construction *is the case* is often considered synonymous with *is true*. More obviously than *is true*, *is the case* appears to act as a semantically redundant sentence operator, serving at best the purpose of hosting negation, as in (42a), or as permitting quantification, as in (42b), or anaphoric reference, as in (42c):

(42) a. That it is raining is not the case.
b. Several things he said are not the case.
c. That is not the case.

The equivalence of the sentences below, with subject clauses and with extraposition, seems to show the synonymy of *is true* and *is the case*:

(43) a. That S is not true.
a'. That S is not the case.
b. It is not true that S.
b'. It is not the case that S.

However, the two constructions are in fact fundamentally different both syntactically and semantically.

A first difference consists in that (44a) is perfectly fine as it is, whereas (44b) is quite peculiar:

(44) a. That it is raining is true.
b.?? That it is raining is the case.

Is the case seems to require negation, as in (42a), or else an adverbial (*that it is raining is often the case*).

Another difference is that unlike *is true*, *is the case* does not accept full NPs in subject position:[20], [21]

(45) a.??? The proposition that S is the case.
b.??? The belief that S is the case.
c.??? The sentence 'S' is the case.

It has been held that the *is the case*-construction reflects the Identity Theory of truth. That is, *that* S *is the case* is true just in case that S picks out a wordly fact.[22] If S fails to pick out a worldly fact, then *that* S *is true* is false.[23] On this view, *is the case* is in fact treated as an existence predicate. An existence predicate, unlike other predicates, does not presuppose the existence of the subject referent. This account, as we will see, cannot be right. The *that*-clause in *that* S *is the case* may be evaluated relative to different situations, rather than denoting a single entity.

The most important semantic difference between *is true* and *is the case* concerns their behavior with adverbial modifiers. First, *is true* and *is the case* differ in their acceptance of location modifiers. Location modifiers are perfectly fine with *is the case*, but they are hardly acceptable with *is true*:

(46) a. In our firm, it is not the case that one gets fired without explanation.
b.?? In our firm, it is not true that one gets fired without explanation.
(47) a. In John's family, it is not the case that children respect their parents.
b.?? In John's family, it is not true that children respect their parents.

---

[20] It has been held that that *is the case* does not apply to representational objects, such as propositions, beliefs, or sentences, but only to states or affairs or situations (Mulligan 2010). But in fact explicit descriptions of situations or states of affairs are equally impossible with *is the case*:

(i) a.??? That state of affairs is the case.
b.??? The situation he described is the case.

[21] One might expect *is true* and *is the case* to differ in another respect. Whereas *true* as an adjective should have predicative status, this would not be expected for *the case* in *is the case*. Yet, *the case* in that context satisfies the same syntactic criteria for predicatehood as *true*. In particular, *true* and *the case* can be the predicate in 'small clauses', a standard linguistic criterion for predicatehood:

(i) a. I consider it true that John is a genius.
b. I consider it the case that John is a genius.

[22] The Identity Theory of truth is that of early Russell and Moore; see Candlish and Damnjanovic (2011).

[23] Wittgenstein's dictum below in (ia) appears to be an expression of the Identity Theory, given the assumption that (ia) means just what (ib) means:

(i) a. The world is everything that is the case.
b. The world is the totality of facts.

On the intended meaning, *everything that is the case* would have to stand for the totality of worldly facts that 'are the case'. The question is whether (ia) is really acceptable (and its slightly provocative sound suggests that it is not). On the present view, *everything* in (ia) would best be considered a substitutional quantifier or something close to it. But then *everything that is the case* can hardly stand for the totality of facts. Thus, (ia) comes out as unacceptable.

Whereas (46a) and (47a) are perfectly natural as statements of facts, (46b) and (47b) if not unacceptable, at least convey a somewhat particular metasemantic notion of location-relative truth.

Furthermore, *is true* can hardly go together with adverbs of quantification, which are fine with *is the case*:

(48) a. Given that she has developed Alzheimers, it will often be the case that Mary forgets something.
b.?? It will often be true that Mary forgets something.
(49) a. It was twice the case that someone was absent.
b.??? It was twice true that someone was absent.

The use of adverbs of quantification with *is the case* shows that the subject clause may be evaluated with respect to the various situations that the adverb of quantification ranges over. The *that*-clause won't denote a single entity, which means that the identity-theoretical account of *is the case* cannot be right.

The use of propositional anaphora with *again* shows the same thing:

(50) It was once the case that S. Today that is the case again.

By contrast, the *that*-clause with *is true* needs to be propositionally complete. *That* S in *that* S *is true* is understood as complete regarding context-dependent elements, such as quantifier restrictions, tense interpretation, spatial location etc., though of course the proposition expressed may involve 'unarticulated constituents'.

Is the case is not the only 'truth predicate' in English that rather than attributing truth, involves an evaluation of the *that*-clause with respect to situations. *Occur* can act that way as well, as (51), with a location modifier and an adverb of quantification illustrates:

(51) That a student in this school failed an exam has never occurred.

Occur shares also other linguistic properties with *is the case*. It allows for extraposition and subject clauses and for a replacement of a subject clause by a quantifier or pronoun:

(52) a. It is never occurred that a student in this school failed an exam.
b. Nothing of what he predicted/That has occurred.

Occur differs from *is the case,* though, in that it imposes a restriction on the *that*-clause. It accepts only *that*-causes with eventive verbs:

(53)?? In John's family, it does not occur that children respect their parents.

Another construction belonging like *is the case* and *occur* to type 2 truth predicates appears in English only in a restricted form, namely *is so*. *Is so* does not accept *that*-clauses, but only sentential anaphora:[24]

[24] The German version is not subject to the restriction:

(i) Daβ es im Winter kalt ist, war schon immer so.
'That it is cold in winter was always so.'

Below we see that the construction also allows for extraposition:

(54) a. It is perhaps so
b. Is that so?

Concerning the semantics of *is the case* (and other type 2 truth predicates), it is revealing to take a look at referential terms that can be formed with the noun *case*. Such terms are indicative as to what sorts of entities are involved in the semantics of type 2 truth predicates. They arguably are precisely the sorts of entities with respect to which the *that*-clause will be evaluated, or better that serve as truthmakers of the *that*-clause.[25]

Case occurs as the head noun of referential terms like *the case in which* S, as below:

(55) a. We discussed the case in which John might not return.
b. We cannot exclude the case in which John might be unable to do the job.

What is peculiar about such noun phrases is that they require a modal of possibility in the *that*- clause. The modal in this context serves as an indicator that the situation the NP refers to is a merely possible one. Without a modal, the *case*-NP would refer to an actual situation, and that, for some reason, is not permitted:

(56) a.??? We were relieved about the case that John returned.
b.??? We discussed the case in which John is unable to do the job.

Case-NPs thus need to refer to merely possible situations or better possible 'cases'. Note that merely possible cases are not possibilities. 'The possibility that John might return' exists just in case it is possible that John returns. For a possible case to 'exist', it has to be actual. This can be seen from the way existence predicates for cases are understood, such as *occur* and *present itself*:[26]

(57) The case that John will not return could occur/ present itself.

'Cases' are situations in a certain sense and as referents of *case*-NPs they are merely possible situations. But once they are said to 'exist', they are worldly facts or actual situations.

Cases are not true propositions or non-wordly facts. Unlike the latter, they have to be fully specific. Thus, a case in which a student fails the exam involves a particular student failing the exam, and a case in which John or Mary fails the exam involves either John's failing the exam or Mary's failing exam, but not a disjunctive condition.

(ii) Es war schon immer so, daβ es im Winter kalt ist.
'It was always so that it is cold in winter.'

(iii) illustrates that the construction does not allow for referential NPs:

(iii) * Dieser Satz/Diese Proposition/Dieser Sachverhalt was schon immer so.
'This sentence/This proposition/This state of affairs has always been so.'

That is, *is so* can mean neither 'true' nor 'obtain'.

[25] For the notion of a truthmaker see, for example, Mulligan et al. (1984), Armstrong (1997), Moltmann (2007), and Fine (2012).

[26] 'Cases' thus are also not states of affairs. States of affairs may or may not obtain. But 'cases' could not be said to 'obtain'. What exactly the ontological differences between states of affairs and cases amounts to remains to be clarified of course.

While *case*-NPs are subject to the constraint that they have to refer to merely possible 'cases', *is the case* obviously involves reference to an actual 'case'. This is not achieved by the noun phrase the *case* itself, though. *The case* in that context does not behave as a truly referential NP. It requires the simple definite determiner *the* (* *it is that case that* S, * *it is a case that* S), and does not permit any modifiers (**it is the unfortunate case that* S):

(58) * That it is raining is the case that I did not expect.

The case in *is the case,* moreover, cannot act as the antecedent of a *case*-NP:

(59) That no one came was recently the case.?? We did not like that case.

The case rather appears to be a mere referential residue. In fact, the construction *it is the case that* S generally involves quantification over cases. This is so with adverbs of quantification and also the negative sentence *it is not the case that S*, which states that there is no 'case' that supports S.

The relation between a 'case' and the *that*-clause that is involved in the semantics of the *is the case*-construction is a relation of truth-making, and it needs to be the relation of exact truth-making.[27] That is, it is the relation that holds between a case and a *that*-clause only if the case is wholly relevant to the truth of the *that*-clause. This is clear from the way quantifiers are understood:

(60) It was twice the case that John made a mistake.

Twice in (60) counts situations that are completely relevant for the truth of *John made a mistake*, that is, situations that include nothing more than John, a single mistake, and the 'making'-relation holding between the two. *Twice* does not count larger situations or sums of such situations.

The semantics of *is the case* with a location modifier or an adverb of quantification will thus be as follows, where s $\Vdash$S is the exact truth-making relation:

(61) a. *In this firm it is the case that* S is true iff for the maximal actual 'case' s such that *in this firm*(s), s $\Vdash$S.
b. *It is sometimes the case that* S is true iff for some 'cases' s, s $\Vdash$S.

The truthmaking view as a general view about truth says that if a sentence/proposition is true, it is true in virtue of something in the world that makes it true. The fact that English has constructions expressing the truth-making relation between 'cases' and the content of *that*-clauses does not imply that English is committed to the truth maker view as a view about truth itself, though. That is, it does not imply the view that the truth of any sentence must be grounded in a truthmaker. Rather, it simply means that the semantics of English involves a concept of truthmaking that relates situations or rather 'cases' to the content of *that*-clauses.

It is the case that, it seems, involves as its interpretation the sort of semantics that Austin (1950) proposed for independent sentences in general. On Austin's view, with

[27] This is the truth-making relation that is used, for example, in Moltmann (2007) and Fine (2012).

the utterance of a sentence, a speaker refers to an (actual) situation and claims that the situation referred to is of the type specified by the sentence uttered. That is, the situation referred to acts as the truthmaker of that sentence. On the present view, this is only part of the constructional meaning of *is the case*. With *is the case*, adverbs of quantification range over 'cases' and location adverbials act as predicates of cases. Austin's motivations for implicit situation reference were in fact quite different from the present ones. The situation referred to, for Austin, is responsible for contextual restrictions on quantification domains, the interpretation of tense etc. The present motivation for invoking truth-making is quite simply the semantics of the *is the case*-construction.

2.4 Conclusions

This paper has pointed out a range of linguistic facts about truth predicates in English (and German), which required a significant re-evaluation of the motivations of theories of truth that appeal to natural language. On the positive side, the paper has argued for an account of sentences of the form *that* S *is* P for a type 1 truth predicate P according to which *that* S specifies the content of a contextually given (or accommodated) attitudinal object such as a belief or claim and P acts as a predicate of that entity. Type 2 truth predicates such as *is the case*, by contrast, involve in their semantics the relation of truth-making, relating cases to the *that*-clause, and not predication of a truth property.

One obvious question the paper raises is the crosslinguistic generality of the observations made on the basis of English and German. Certainly, it is expected that the distinction between the two types of truth predicates is found across languages in general. Moreover, the observations made about type 1 truth predicates suggest that natural languages display a great variety of normative and representation-related concepts of truth. Of course, it would be highly desirable to be able to add data from other languages to this general picture and to establish further generalizations on the basis of them.

Acknowledgments I would like to thank the audiences of the conference *Truth at Work* (Paris, June 2011) and of the *New York Semantics Colloquium* (New York, November 2012), where previous versions of this paper were presented. I would also like to thank Marcel van Dikken, Hartry Field, and Paul Horwich for stimulating discussions and Wolfgang Kuenne as well as two anonymous referees for their helpful comments.

2.5 Appendix: Two Further 'Truth Predicates'

In this appendix, I will briefly discuss two further apparent truth predicates in English, *is a fact* and *is the truth*, focusing on their rather special syntax and semantics.

4. ***Is a fact***

Is a fact appears to act as a truth predicate below:

(1) That the sun is shining is a fact.

We will see, though, that the semantics of the *is a fact* -construction is fundamentally different from that of type 1 or type 2 predicates: it involves neither attribution of truth to a representational object nor the expression of the truth-making relation, but rather a specification of the 'content' of a fact.

Let us look at the linguistic properties of the *is a fact*-construction. Like *is true*-sentences, *is a fact*-sentences allow for extraposition:

(2) a. That S is a fact.
b. It is a fact that S.

Moreover, like *is true*-sentences and unlike *is the case*-sentences, *is a fact*-sentences resist location modifiers and adverbs of quantification:

(3) a.??? In our firm, it is never a fact that someone gets fired without explanation.
b.??? It was twice a fact that someone was absent.

This means that *that* S in *that S is a fact* must be propositionally complete.

There are differences, though, between *is true* and *is a fact*. Unlike *is true*, *is a fact* allows only for simple quantifiers and pronouns in subject position and not for referential NPs:[28]

(4) a. * John's belief is a fact.
b. * That sentence is a fact.
(5) a. Nothing is a fact.
b. It is raining. That is a fact.

Unlike *the case* in *is the case*, *a fact* in *is a fact* is an ordinary indefinite NP, allowing for adjectival modifiers, relative clauses, and anaphora support:

(6) a. That S is an interesting fact.
b. That S is a fact that I had never noticed.
c. That S is a fact. That fact is hardly known.

All this suggests that the *is a fact*-construction reflects the Identity Theory of truth: *That* S *is a fact* is true just in case that S picks out a fact. This would also account for the possibility of negation:

(7) That S is certainly not a fact.

However, the identity-theoretic analysis is implausible. *That* S by itself cannot stand for a fact, not only because of the lack of referential independence of *that*-clauses

[28] Note that *is a fact* does not allow for free relative clauses with attitude verbs, unlike *is the case* and *is true* (Austin 1961b):

(i) a. What John said/believes is true.
b. ?? What John said/believes is a fact.
c . What John said/believes is the case.

I do not have an explanation of that difference.

discussed earlier. Let us look at the sentences below, which display the very same construction:

(8) a. That S is a possibility.
b. That S is a common belief.

If *that* S could by itself stand for a fact, then (8a) and (8b) could have a reading on which they are false just because S is true, since a fact is neither a possibility nor a belief. A fact is not a possibility since the possibility that S exists in circumstances in which S is not true. Moreover, a belief obviously is not a fact.

More plausibly, the *that*-clause in (1) occurs nonreferentially and serves to specify the 'content' of a fact. That is, (1) expresses a relation of specification that holds between the content of the *that*-clause in subject position and a fact (and not predication of the property of being a fact of the referent of the *that*-clause). This will be the very same semantic relation that obtains between *fact* and the *that*-clause in *the fact that the sun is shining*. The same holds of course for (8a) and (8b).[29]

5. *Is the truth*

The truth predicate *is the truth* is a very puzzling one both from a semantic and a syntactic point of view:

(9) That John is guilty is the truth.

Obviously, what the subject clause in (9) denotes cannot literally be 'the truth'. It could not make up the one and only 'truth'; there are lots of 'truths'.

In what follows, I will identify a range of semantic and syntactic properties of the construction and point at the kind of syntactic and semantic analysis that it most plausibly has.

It is tempting to take *the truth* in this context to stand for the unique contextually relevant 'truth' (that is, true proposition) or to act as a predicatively used contextually restricted definite description. But a contextual restriction driving the interpretation of *the truth* in (9) is implausible. For (9) to be acceptable, no particular context is required that would restrict the denotation of *the truth*. *The truth* does not behave like predicatively used contextually restricted NPs as in the examples below:

(10) a. This chair is the yellow chair.
b. This piece of furniture is the yellow chair.

Unlike (9), (10a) and (10b) do require a particular previous discourse context that was about a unique yellow chair.

In fact, *is the truth* belongs to a different construction than that of a subject-predicate sentence, as well as that of *that S is a fact. The truth* has neither the status of a predicatively used NP nor of a referential or quantificational NP (whatever the view of definite NPs may be).

[29] For such an analysis of *That is is a fact* and *the fact that* S see Moltmann (2013, Chaps. 2 and 6).

Three properties distinguish *is the truth* from ordinary predicates. First, *is the truth* allows subject-predicate inversion, as seen in (11), unlike ordinary predicates, such as type 1 and type 2 truth predicates, as in (12a) and (12b):

(11) a. That John will not return is the truth.
b. The truth is that John will not return.
(12) a. * True/ Correct/ Right is that John will not return.
b. * The case is that John will not return.

Second, unlike predicates taking clausal subjects, *is the truth* does not allow for extraposition:[30]

(13) * It is the truth that John will not return.

Third, *is the truth* requires a definite determiner in *the truth*, unlike ordinary predicates such as *is a chair*:

(14) a. * A truth is that he will not return.
b. * That John will not return is a truth.

Given these three properties, we can conclude that *is the truth* is not a syntactic predicate taking clausal subjects. Moreover, *the truth* in that construction does not act as an ordinary definite NP used predicatively. Otherwise the restriction to definiteness would be unexpected.

The construction more plausibly is a type of specificational sentence (Higgins 1979). Specificational sentences come in two sorts: with a free relative clause in subject position, as in (15a), and with a definite NP in subject position, as in (16a) (Higgins 1979):

(15) a. What John did was kiss Mary.
b. Kiss Mary is what John did.
(16) a. The best player is John.
b. John is the best player.

(15b) and (16b) illustrate, inversion is possible in both cases.

Semantically, specificational sentences have been analysed in one of two ways: [1] as expressing a question-answer relationship, with the subject acting as a concealed question and the postcopula expression partially specifying an answer (the Question-Answer Analysis) and [2] as expressing an identity among possibly higher-level semantic values (the Identity Analysis).[31] It turns out that neither analysis can be right for specificational sentences with *that*-clauses in general. To see this, let us look at some more familiar kinds of specificational sentences involving *that*-clauses:

(17) a. John's claim is that it is raining.
b. That it is raining is John's claim.

[30] This poses difficulties for Hinzen's (2003) view, who takes *the truth* in *is the truth* to have the status of a predicate.
[31] For the Question-Answer Analysis of specificational sentences, see, for example, Schlenker (2003) and references therein. For the Identity Analysis see, for example, Sharvit (1999) and references therein.

(18) a. The idea is that there will be a party.
b. That there will be a party is the idea.

It is not entirely obvious how the Question-Answer Analysis applies to specificational sentences with *that*-clauses. On that analysis, the subject of (18a) might stand for a question of the sort 'what idea is there?' or perhaps of the sort 'what is the idea?'. The postcopula NP presumably will partially specify an answer of the sort 'the idea that there will be a party'.

This kind of analysis is not generally applicable, however. In particular, it is not applicable to sentences with *the truth* as subject. On that analysis, the postcopula *that*-clause in (9) would partially specify an answer of the sort 'the truth that John is guilty'. But *the truth that John is guilty* is ungrammatical: *truth* does not accept *that*-clauses. The difficulty arises with other specificational sentences as well. Higgins (1979) already observed that in specificational sentences, the subject and a postcopula *that*-clause need not be able to form an NP. Thus, (19a) is a specificational sentence as well, but (19b) is ungrammatical:

(19) a. The proof that John is guilty is that his fingerprints are on the knife.
b. * the proof that John is guilty that his fingerprints are on the knife

The Identity Analysis does not straightforwardly apply to specificational sentences with *that*-clauses either. What a *that*-clause generally is taken to stand for is not a claim, an idea, or a proof. A *that*-clause specifies the content of such entities, but is not identical to them. The relation expressed by a specificational sentence with a *that*-clause could only be that of content specification, not that of identity.

A further question that the *is the truth*-construction raises is, why is the subject a definite NP when there need not be a unique entity it stands for? The Question-Answer Analysis would say that an NP of the sort *the fact that* S or *the claim that* S is obligatorily definite, since indefinites such as *a fact that* S or *a claim that* S are unacceptable. But we have seen that that analysis was not generally applicable to specificational sentences with *that*-clauses. There is a more plausible way of explaining the obligatory definiteness. It is a general requirement that specificational subjects always be definite, illustrated below (Heycock and Kroch 1999):[32]

(20) a.??? A good player is John.
b.??? A problem is that it is raining.

Given that (9) is in fact the inverted structure, this means that the obligatory definiteness of *the truth* in *is the truth* would be an instance of a more general condition on the

[32] *The truth* occurs in yet a different construction, as a concealed question below:

(i) a. John told the truth.
b. John knows/found out the truth.

The concealed-question use is not available with nominalizations of other type 1 truth predicates:

(ii) a. * John told the falsehood.
b. * John told the correctness.

subject of specificational sentences. Of course, the definiteness condition on specificational subjects needs to be explained itself (perhaps by associating the subject of a specificational sentence with a particular semantic role, for example by attributing it an anaphoric status relating to what is at least implicitly under discussion).

Let us then summarize this rather inconclusive discussion by stating that *is the truth* is a pseudo-truth predicate involving a complex syntactic structure whose semantics is far from well understood.

References

Armstrong, D. (1997). *A world of states of affairs*. Cambridge: Cambridge University Press.

Aune, B. (1967). Statements and propositions. *Noûs, 1*(3), 215–229.

Austin, J. L. (1950). Truth. *Aristotelian Society, Suppl 24,* 111–129. (Reprinted in Austin (1961a)).

Austin, J. L. (1961a). *Philosophical papers*. J. O. Urmson & G. J. Warnock (Eds.). Clarendon Press: Oxford.

Austin, J. L. (1961b). *'Unfair to facts'* (pp. 102–222). Oxford: Clarendon Press.

Boghossian, P. (2003). The normativity of content. *Philosophical Issues, 13,*(1) 31–45.

Boghossian, P. (2010). Our grasp of the concept of truth: Reflections on Künne. *Dialectica, 64,* 553–563.

Candlish, S., & Damnjanovic, N. (2011). The identity theory of truth. *Stanford encyclopedia of 850 philosophy* (Online).

Fine, K. (2012). Counterfactuals without possible worlds. *Journal of Philosophy*, *109*(3), 221–246.

Gibbard, A. (2005). Truth and correct belief. *Philosophical Issues, 15,* 338–350.

Grover, D. L., Camp J. L., & Belnap N. D. (1975). A prosentential theory of truth. *Philosophical Studies, 27,* 73–125.

Heycock, C., & Kroch, A. (1999). Pseudocleft connectivity: Implications for the LF interface level. *Linguistic Inquiry, 30*(3), 365–397.

Higgins, R. (1979). *The pseudo-cleft construction in English.* Indiana University Linguistics Club.

Hinzen, W. (2003). Truth's fabric. *Mind and Language, 18*(2), 194–219.

Horwich, P. (1990). *Truth*. Oxford: Blackwell.

Kölbel, M. (2008). 'True' as ambiguous. *Philosophy and Phenomenological Research, 77*(2), 359–384.

Koster, J. (1978). Why subject sentences don't exist. In J.S. Kayser (ed.), *Recent transformational studies in European languages*, pp. 53–64. Cambridge: MIT Press.

Künne, W. (2003). *Conceptions of truth*. Oxford: Clarendon Press.

Moltmann, F. (2003a). Nominalizing quantifiers. *Journal of Philosophical Logic, 32,* 445–481.

Moltmann, F. (2003b). Propositional attitudes without propositions. *Synthese, 135,* 70–118.

Moltmann, F. (2004). Nonreferential complements, derived objects, and nominalizations. *Journal of Semantics, 13,* 1–43.

Moltmann, F. (2007). Events, tropes and truthmaking. *Philosophical Studies, 134,* 363–403.

Moltmann, F. (2013). *Abstract objects and the semantics of natural language*. Oxford: Oxford University Press.

Mulligan, K. (2010). The truth predicate vs the truth connective. On taking connectives seriously. *Dialectica, 64,* 565–584.

Mulligan, K., Simons, P., Smith B. (1984). Truthmakers. *Philosophy and Phenomenological Research, 44,* 287–321.

Portner, P. (1997). The semantics of mood, complementation, and conversational force. *Natural Language Semantics, 5,* 167–212.

Ramsey, F. P. (1927). Facts and propositions. *Aristotelian Society Supplement, 7,* 153–170. (Repinted in D. H. Mellor (ed.): *F. P. Ramsey, Philosophical papers*, Cambridge UP, Cambridge 1990).

Schlenker, P. (2003). Clausal equations (A note on the connectivity problem). *Natural Language and Linguistic Theory, 21,* 157–214.

Sharvit, Y. (1999). Connectivity in specificational sentences. *Natural Language Semantics, 7,* 299–304.

Yablo, S. (2014). *Aboutness*. Cambridge: MIT Press.

Chapter 3
Truth and Language, Natural and Formal

John Collins

Abstract The article seeks to support a version of Tarski's view of natural language truth, what is now often referred to as the 'inconsistency view'. Expressed naively, this view claims that natural languages are inconsistent because they support paradoxical reasoning. The view is mislabelled; it wasn't even Tarski's considered position that natural languages are inconsistent. I shall argue that the attribution of inconsistency to natural language is a kind of category error that reflects the fundamental difference between natural and formal languages: the former do not transparently encode semantic relations in their structure whereas the latter do. Still, the paradoxes of natural truth are insoluble, just as Tarski suggested. This is because, I shall suggest, truth, by its very semantic role, is an inherently risky notion in that its natural expression does not come with any *necessary* indication of exactly what is being claimed to be true; thus, one may accidentally fall into paradox. In a phrase, truth is an opaque metarepresentational notion. If that is so, then there can never be security against paradox, at least if truth is to retain its metarepresentational freedom. This result is expected on the view that natural language is not semantically transparent.

3.1 Introduction

There are numerous ways of distinguishing between philosophical approaches to the concept of truth. If we focus on the past 40 years or so, one clear division is that between informal and formal conceptions. The former include general characterisations of truth in terms of correspondence, deflation, coherence, pragmatic utility, and so on. The endeavour here is to capture our colloquial notion of truth as expressed by the truth predicates of natural languages. The latter formal conceptions are not necessarily at odds with the informal approaches, but seek to construct various analogues of natural truth in the context of a formal theory, mostly Peano arithmetic. Depending on the particular account, these analogues are intended to correspond more or less to the colloquial notion. Still, the chief desideratum is that

J. Collins
University of East Anglia, Norwich, England
e-mail: john.collins@uea.ac.uk

T. Achourioti et al. (Eds.), *Unifying the Philosophy of Truth,* Logic, Epistemology, and the Unity of Science 36, DOI 10.1007/978-94-017-9673-6_3

such analogues evade the paradoxes that appear to beset the common concept. The construction or axiomatisation of a consistent truth concept may thus lead one to think that natural truth is after all consistent, once suitably tided up. That has been the standard view, in opposition to Tarski's (1936/1983) view, which was that formal analogues of truth that meet certain conditions are not exactly the same notion as the colloquial one but are serviceable as truth concepts in the right formal setting. In distinction, informal approaches to truth have largely neglected the paradoxes; indeed, often the paradoxes are explicitly eschewed as orthogonal to the informal endeavour (e.g., Horwich 1990; Vision 2004).[1] My aim here is to support a version of Tarski's view, what is now often referred to as the 'inconsistency view'. Expressed naively, this view claims that natural languages are inconsistent precisely because they support paradoxical reasoning. A chief point of the following will be that the view is mislabelled; it wasn't even Tarski's considered position that natural languages are inconsistent. I shall argue that the attribution of inconsistency to natural language is a kind of category error that reflects the fundamental difference between natural and formal languages: the former do not transparently encode semantic relations in their structure whereas the latter do. Still, the paradoxes of natural truth are insoluble, just as Tarski suggested. This is because, I shall suggest, truth, by its very semantic role, is an inherently risky notion in that its natural expression does not come with any *necessary* indication of exactly what is being claimed to be true; thus, one may accidentally fall into paradox. In a phrase, truth is an opaque metarepresentational notion. If that is so, then there can never be security against paradox, at least if truth is to retain its metarepresentational freedom. From the perspective of the general lack of semantic transparency in natural language, this result is expected.

So, according to the position to be expounded below, the informal approaches' characteristic eschewal of 'solutions' to the paradoxes is correct insofar as natural language pathology is incurable, but, by the same token, the attitude is also mistaken, if we are supposed to think that pathology is not intrinsic to natural truth. On the contrary, an account of natural truth should explain why the paradoxes arise. In this sense, therefore, being heir to inconsistency is precisely what an informal approach should preserve. It is a fault of the informal approaches, therefore, that they have not properly tackled the paradoxes, or, as my position has it, faced up to the inherent pathology of natural language.[2]

[1] Such eschewal of the paradoxes is not universal, of course. There is currently much discussion of the paradoxes in relation to a general deflationism about truth. Still, this new interest was sparked by a general neglect of the paradoxes by the various deflationist proposals, even though their minimal resources appeared to be precisely the resources Tarski had identified to be paradox-inducing. Furthermore, the other variants of general accounts of truth remain relatively silent about the paradoxes.

[2] For the purposes of the following, I shall assume that the *concept* of truth is perfectly expressed by the English truth predicate and that whatever there is to the *property* of truth is likewise expressed by the predicate. It might be a fault of some deflationists that they seek metaphysical conclusions from semantic premises, and it might further be that such a mode of inference is invalid in numerous instances, but I am happy to let semantics lead the metaphysics, at least when it comes to truth.

In the paper's first part, I shall distinguish natural from formal languages such that only in regard to the latter may notions of consistency be properly applied. The paper's second part will then draw out the lessons for our understanding of truth as a concept heir to paradox in its natural setting.

3.2 Natural and Formal Languages

Before turning to truth itself, in this section I shall argue that natural language is quite distinct from formal language; so much so that the use of 'language' to describe the latter is a kind of pun. This conclusion will provide the basis for my more narrow differentiation of natural truth from formal truth. Of course, no-one labours under the impression that natural and formal languages are exactly alike; after all, the philosophical rationale for the development of formal languages originating with Frege and Russell, and many who followed them, was the perceived failing of natural language explicitly to mark and differentiate truth-relevant categories, such that, for example, proper names and quantifier phrases may be properly distinguished instead of being lumped together as subjects of predicates, as if they both named objects. That kind of complaint, however, will not be my focus, even if sound.[3] My concern is not to show that natural language disguises the thoughts it expresses, thoughts that may be displayed in their true garb via a formal language. That picture presupposes that there are such crystalline thoughts for natural language to express, which it does so, albeit in its bungled way. We have little reason, if any, to think such an arrangement obtains. My claim, rather, is that natural language does not share the design features of formal languages; in particular, there is no transparent map from syntax to semantics, but, for all that, natural languages are highly constrained structures save not of the kind exhibited by formal systems. This conclusion, to be sure, is in sympathy with the thoughts of Frege and Russell, but the reasoning remains distinct. Crucially, as advertised, the upshot will be a cleavage between natural and formal truth, not a way of rescuing the former by way of the latter.

Let us begin with formal language, for its character is much less controversial than its natural namesake; furthermore, since I am concerned just with the distinction between natural and formal language, not the precise nature of either, it will be useful to have formal language clearly in view as a template, as it were, in order to see how natural language fails to fit it independently of many of the controversies that beset the very notion of a natural language.

A formal language is a set of symbol types that satisfy a class of conditions that amount to an explicit stipulation or definition as to what is to count as a well-formed formula ('sentence') of the language. Syntactically, the formula can be viewed as

[3] For what it is worth, I think the advertised shibboleth has some truth to it, but the resources of traditional grammar, to say nothing of contemporary syntactic theory, suffice to distinguish between, say, proper names and quantifier phrases. See Oliver (1999) for an amusing discussion of this issue.

concatenations of the primitive symbols ('the alphabet'), where the definition provides a decision procedure for telling what concatenations from the alphabet are and are not formulae. Semantically, the language is designed so that the definition and associated principles of manipulation (a proof theory) applied to the formulae respect independently specified properties and relations, which comprise the interpretation or model of the language, so that each symbol (complex and simple) is assigned an interpretation. In effect, therefore, a formal language is an artefact designed to allow its users to reason mechanically about the relevant domain. Let us call this fundamental feature of formal language *full transparency*: every syntactic condition on formulae (well-formedness and proof) expresses a semantic condition. So, there are no semantic properties expressible by the symbols that are not encoded in the formulae, and there are no syntactic properties that lack determinate semantic properties. This arrangement allows the semantic properties to be read-off the syntactic ones. Thus, the semantic properties of the formulae are transparent in their syntax; formulae wear their semantics on their sleeves. Of course, the choice of syntax remains a matter of convention or choice, and there are numerous familiar syntactic systems for the one underlying system, such as first-order logic. Still, whatever syntax is chosen, it must transparently encode the relevant semantic properties.

I take the above picture of formal language to be uncontroversial. Before moving onto natural language, though, a few points should be made explicit, which, while I hope they are uncontroversial too, are not often explicitly noted.

Firstly, although a formal language, as specified by a definition, is independent of any given proof theory, the language should not be understood as independent of *every* proof theory. The language is designed to allow for mechanical reasoning, so the language must be able to support *some* proof procedure, even if one has a choice in what proofs are permitted. For example, a single language of first-order logic (with some choice of constants) supports both classical and intuitionistic proof, with the two theories producing different sets of theorems in the single language. A formal language that didn't allow the implementation of some proof procedure would be pointless, a symbol salad generator.

Secondly and similarly, a formal language is distinct from its semantics in two senses. First, understood just from its definition, a formal language is a formal object, which has no inherent interpretation at all, and about which theorems can be proved independently of any notion of interpretation. Moreover, there will be no unique interpretation of the language, in the sense that we may choose different kinds of model by which to interpret the language. This holds, of course, even if there is an 'intended interpretation', for non-standard models can always be found save for the simplest of languages. Second, one may, quite trivially, know the definition of a formal language without understanding the relevant content (intended interpretation) at all. For example, an introductory text on set theory will introduce the reader to the symbolism and language of, say, ZF set theory, but the reader does not therefore know any set theory (they have only just read the first chapter), even though they may rightly be said to know the language.

All that said, as with proof, a formal language is designed to encode some range of properties and relations. So, even though the definition of formula for the language

will not determine a unique interpretation, the language must be interpretable in the appropriate sense, i.e., formulae have determinate interpretations, given a choice of a model, and the relevant semantic properties are preserved across formulae that make up proofs in the language. In other words, an interpretation that does not track the syntax of the formulae (in accordance with *full transparency*) is not an interpretation. If a language is not interpretable, then it is mere symbol salad.

Thirdly, some consequences of what I am calling full transparency are worth spelling out explicitly. (i) Given a model, every symbol of a formula will be interpreted, or be merely typographic, such as brackets and other signs of punctuation. (ii) The interpretation will be determinate, admitting no ambiguity or polysemy. (iii) The interpretation of a formula will be strictly determined by the interpretation of its constituent symbols in the sense that it will not have an interpretation that goes beyond what the symbols provide. For example, a 2-place relational expression will not be interpreted as a 3-place relation. (iv) The position in which a symbol occurs in its host formula determines its semantic relations to the remaining symbols of the formula. Scope provides an obvious example, where, say, the scope of a quantifier is the smallest sub-formula of which it is a part. More obviously still, a variable belonging to a sub-formula will be interpreted as taking values as arguments of the given formula, not values as arguments of some other sub-formula of the host formula. All these properties make the language optimally user-friendly for encoding the relevant relations and properties to permit mechanical reasoning, and since it is *we* who invent the formal language, then we obviously make it so that it is so friendly; if the language fails in this regard, then we simply junk it or change it (consider, for example, the unwieldy diagrammatic notation of Frege's *Begriffschrift* next to the modern notation of logic).

In contrast to formal language, the status of natural language remains highly contentious. There are three broad conceptions of it. The cognitive conception, mostly associated with Chomsky (1986), views language as an internally constituted cognitive capacity, ultimately a biological phenomenon analogous to vision or insect navigation. What we may call the social conception views language as an external symbolic medium, which is not independent of human cognition, but is, nonetheless, not an internally constituted state of speaker/hearers (e.g., Soames 1984; Devitt 2006). Lastly, there is the Platonic conception, associated with Katz (1981) and Katz and Postal (1991). According to this perspective, languages are abstract objects, which humans access, but which are not constituted by our psychology or our social exchanges.

Space precludes an adjudication of these different conceptions, but I think it is clear that the cognitive conception is the most minimal and so should be the default view, for *every* conception of human language must presuppose a special human design for language acquisition and competence, given that no other species possesses the capacity or can acquire it.[4] We require arguments to move beyond

[4] To be sure, we talk of many non-human species and non-biological systems as possessing language, but this is essentially punning. Hereon, by 'natural language' I mean the peculiar cluster of features

what any remotely plausible view would presuppose. Such arguments have been put forward, of course, but they are far from compelling.[5] What all these conceptions do share explicitly, however, is the notion that natural language is a phenomenon, in the sense that we cannot stipulate or agree on its properties, but must discover them. Correlatively, there are bounds on what properties natural languages may realise. Someone may, indeed did, decide to use *bad* to mean good, but no-one may successfully decide to form the interrogative of a declarative by saying the sentence backwards, or decide to drop subjects of sentences trusting the determination of the understood subject to context (at least not in English and similar languages). Since, therefore, natural language is not a matter of decision or stipulation or agreement, the question arises of how user-friendly it is, for we cannot simply make it so. Natural language appears to be not at all user-friendly, at least not when assessed in contrast to formal language. The crucial factor here is that natural language is not fully transparent; indeed, it is as if designed to confuse and mislead. The reason for this, it seems, is that natural language is an interaction effect between distinct systems, which produces certain compromises, the net product of which is that the system is massively sub-optimal, if construed as a symbol set to support mechanical reasoning the way formal languages are designed to do.

1. Ambiguity

In natural language, ambiguity is ubiquitous at the level of the word, the phrase, and the sentence. By this feature alone, we know that natural language radically departs from full transparency in that the structure of the symbol type is not a sure guide to its interpretation. There is, of course, an awful lot that could be said about the significance of ambiguity, and I shall turn to two general thoughts that might be taken to ameliorate its import for the divide between natural and formal languages for which I am arguing. First, though, let me emphasise an obvious point that is too rarely expressed. Structural (as opposed to lexical) ambiguity is highly constrained; we do not, that is to say, find sentences to be indefinitely ambiguous. Consider (1):
(1) The man spied on the woman with binoculars (1) is ambiguous between a reading where the man uses binoculars to spy on the woman and a reading where the man spies on the woman who possesses binoculars. The adjunct, *with binoculars*, may therefore modify *the woman* or the verb phrase it contains, *spy on the woman*. What is not available is a reading where the adjunct modifies *the man*. The relevant thought or meaning, though, is perfectly coherent and is expressed in the reformulation *The*

that characterises human language. That there are such features is the overarching assumption of modern linguistics and the relevant branches of psychology.

[5] It bears emphasis that our common notion of language *simpliciter* is not deserving of any unifying conception at all. I take the point of the divergent conceptions not to be conceptual analyses of what we mean by 'language', but different attempts to say what does or should animate theoretical inquiry into linguistic phenomena. So, no-one need deny, say, that language allows for communication, but it hardly follows that communication should be explanatorily basic in any sense.

man, with binoculars, spied on the woman. Here, the man might have not used the binoculars to do his spying; he is merely identified as possessing binoculars. One finds similar constraints wherever ambiguity arises: available thoughts are inexpressible by the given structure, even though a variety of other thoughts are expressible.

From the perspective of a formal language, a language designed for the transparent expression or encoding of a determinate set of thoughts, this situation is a double flaw. Firstly, a given structure does not express a unique thought, so extra non-encoded information is required for one to move from the structure to a determinate thought.[6] Secondly, the existence of constraints on the available readings is unpredictable from such readings. In some sense, (1) is talking about a man, a woman, spying, and binoculars. It is determinate in (1) that the man spied on the woman. One might expect, therefore, that where indeterminacy does arise, all options are possible, which would involve the adjunct expression being able to modify all remaining constituents of the sentence (the woman, the spying, and the man), but this is not so, a situation that appears to lack any semantic rationale.

2. Empty words

Natural languages typically contain words that make no contribution to the interpretation of their host sentences.[7] Pleonastic *it* and *there* in weather reports (*It's raining/sunny*) and existentials (*There is a fly in my soup*), respectively, are obvious examples.[8] One may also include infinitive *to* and optional complementizer *that*. The point is somewhat more general, though, than there merely being words that appear optional or obviously empty. Many prepositions are necessary, but look to have no semantic rationale. For instance, the preposition is required in *Bill is proud of Mary*, but one may wonder why on earth it is there; why shouldn't we be able to say *Bill prides Mary*? Similarly, the preposition is required in *Bill gave flowers to Mary*, but we know that *to* isn't required for the interpretation of the dative given the double-object construction: *Bill gave Mary flowers*.

As with ambiguity, from the perspective of a formal language, the presence of empty words appears to be a terrible design flaw, for the words are redundant to the purpose of the design of the language. Imagine, for instance, formulating a notation for the expression of a range of concepts, some new deontic logic, say. Included in the notation is a symbol '$' that systematically occurs in the language, but which does not contribute to the interpretation of the language or its associated proof theory. '$' would have a very short life expectancy.

---

[6] Such information is highly variable, including previous discourse, what is most plausible, intonation, salient features of the context of utterance, and so on. The crucial point is that none of this needs to be *linguistically* encoded within a given token of a sentence.

[7] Not all languages are so free with empty words as English. Curiously, Polish does not contain pleonastic elements such as *it*. Tarski's famous *It is snowing* translates (via the German) the original Polish *Śnieg pada*, whose literal translation is *The snow is falling*. Thanks to Monika Gruber for this information.

[8] It is a mistake to think that *there* expresses existence, for *A fly is in my soup* serves equally well.

One might protest that empty words serve as a kind of optional punctuation, much as brackets do in various notations. There is something to this, for the words do earn their keep by contributing to the identification of various grammatical or functional roles, but the analogy is far from exact. Whether explicitly present or not, a system of punctuation must be understood, in terms, say, of ordering symbols in a notation, for otherwise formulae become ambiguous. In the case of empty words, this is not so. The dative, for example, is introduced by *to*, but it doesn't need to be as the double-object construction reveals; *mutatis mutandis* for complementizer *that*.[9] There also appears to be no punctuational purpose at all for pleonastic subjects, such as *it* and *there*. Independent of particular cases, though, the general point is that punctuation in a formal language must possess a semantic rationale, but that is precisely what is lacking in the case of empty words.

3. Missing words

Just as there are words that appear to be semantically empty, so there are gaps in sentences where a word should be relative to the thought expressed. There are many varied examples of this phenomenon, but, for the purposes of exposition, I shall discuss just a few cases of so-called *object deletion*.

Consider:
(2)a Bill ate an apple
b Bill ate
c Bill didn't eat

Clearly, the transitive and intransitive forms of *eat* express the same concept (cf., *read*, *write*, *drink*, *bake*, *weed*, etc.). The problem, though, is that equivalent concepts surely must have the same adicity. From the perspective of formal language, it should be, contrary to fact, that (2a) and (2b) express different concepts. The plot thickens. One might argue that the transitive and intransitive forms do have the same adicity, for (2a) is naturally read as *Bill ate something*, which suggests that the deletion of an object (e.g., *an apple*) simply triggers existential generalisation. The matter is not so simple, however. Firstly, (2b) doesn't have a mere existential reading, for what is eaten must be foodstuff. Secondly, (2b) has an *activity* reading, whereas (2a) has *accomplishment* reading; that is, (2a) implies that the apple was eaten, finished, but (2b) carries no such implication: whatever the something was, Bill needs merely to have eaten of it for (2b) to be true.[10] Thirdly, on the assumption that object deletion cases do feature existential generalisation, they are scopally ambiguous or, at any rate, formally indeterminate. This is revealed once we explicitly introduce

[9] The complementizer that is not always optional. Compare:
(i) Bill thought (that) Man United would win
(ii) That Man United would win is what Bill thought
(iii) *Man United would win is what Bill thought
Still, there is no obvious semantic rationale for this pattern.

[10] Of course, the condition on (2a) doesn't mean that every bit of the apple, including pips, stork, and core, need be eaten, only that the apple needs to be finished in the appropriate sense. Appropriateness is relative to the nature of the object.

a scope-taking element, such as negation, as in (2c). The sentence is ambiguous between Bill not eating at all (existentially narrow scope relative to negation) or Bill not eating something in particular (existentially wider scope). Not only, therefore, are semantically relevant words missing, but there is also no simple compensating mechanism, such as mere existential generalisation that will resolve the structure into a determinately interpretable form.

The matter is actually worse than is suggested, for in other cases of object deletion, the interpretation of the sentences undergoes a radical change.

(3)a The vase broke the window
b The vase broke
c Bill is too heavy to lift Mary
d Bill is too heavy to lift

(3a) entails that the window broke, not the vase, whereas (3b) does not entail that the vase broke something; on the contrary, it was the vase that broke. The other pair is more complicated. (3d) does have a reading analogous to the existential generalisation over the position of *Mary* in (3c) (imagine Bill, due to his obesity, being excluded from a job involving lifting), but (3d) has a further more natural reading where the heaviness of Bill precludes his being lifted. Notice, though, that (3d) doesn't have a reflexive reading, where Bill is too heavy to lift himself, although *Bill is too stubborn to wash* does have such a reading.

Needless to say, the cases exemplified are the tip of a huge iceberg.[11] The general point the cases demonstrate is that the meaning of a perfectly well-formed non-elliptical sentence cannot be systematically pinned down onto its parts, which again tells against viewing natural language as transparent.

4. Misalignment of positions

Under the last two headings, we saw how sentences often contain semantically redundant words and lack words that are required for the (transparent) interpretation of the sentence. More than these two design oddities, we also find words in semantically uninterpretable positions. Consider this familiar pair:

(4)a Bill is easy to please
b Bill is eager to please

(4a) has the paraphrase *It is easy to please Bill*, with *it* pleonastic. This suggests that the position of *Bill* as the subject is misaligned with its would-be transparent interpretation, i.e., Bill is not easy himself, but his being pleased is what is easy to bring about. In contrast, (4b) lacks the corresponding paraphrase: *It is eager to please Bill* can only mean that some non-human (e.g., a dog) is eager to please Bill; if *it* is construed pleonastically, then the structure is uninterpretable. Thus, the subject position is semantically empty in (4a), but not in (4b), even though the two cases look exactly parallel. Again, from the perspective of a formal language, this is a

[11] For some sense of the empirical complexity of argument deletion and alternation, see Levin's (1993) survey of English.

design flaw, for a position within a structure is not a predictor of the interpretation of its occupier, a gross contravention of full transparency.

5. Intuitions, psychology, and languages

Perhaps the most overarching thing to say about natural language is that it does not support a decidable notion of being a well-formed formula. Principally, this is due to the properties of natural language being determined by the psychological states of users of the particular language at issue. Of course, this doesn't mean that speaker-hearers simply decide, Humpty Dumpty-like, what words mean from occasion to occasion; rather, the point is that there is no arbitration about what words and sentences do mean beyond the evidence gleaned from speaker-hearers. Semantic questions, therefore, along with all other linguistic matters, cannot be decided by stipulation in a way that subsequently governs what a speaker-hearer should or must understand. The properties of particular formal systems, then, no matter how obvious or basic they may appear to the philosopher or logician, cannot be imposed upon speaker-hearers; it must be demonstrated that such properties in fact hold of the speaker-hearers understanding of their language. So, while there is a clear difference between the most simple examples of acceptability (*The man is fat*) and word salad (*The is man fat*), there are innumerable cases of structures of questionable status (*Mary looks like the Battle of Hastings was in 1066*). Narrowing the idea of a decidable natural language to that of an idiolect does not help, even if otherwise acceptable, for variation through time is as ruinous to decidability as is variation across persons.

It is no threat to this position that it undermines the idea of a public language; on the contrary, that is a welcome consequence, for, again, it must be demonstrated what explanatory contribution such a posit may have to understanding the properties of natural language in a way that goes beyond the states of speaker-hearers. Doing without the posit is the most minimal and natural assumption in the absence of contrary evidence. Speaker-hearers work with on-going, developing idiolects, where it is undecidable whether a given string is to count as part of the language or not. Of course, syntactic theory hypothesizes a core syntactic engine, if you will, but such a system massively underdetermines what speaker-hearers will unreflectively (and reflectively, for that matter) sanction as part of their language. Equally, the system will generate structures beyond the means of recognition of speaker-hearers, such as those featuring a million clauses. Thus, we may speak of language as a constantly shifting, occasion-bound snap shot of a speaker-hearer's competence, or instead settle on a hypothesised underlying core system, which is radically misaligned with what the speaker-hearer acknowledges.

So, when one goes to attribute properties to natural language appropriate to formal languages, such as consistency, completeness, and so on, it needs to be established that such properties are supported by speaker-hearers' understanding of their language in either of the senses alluded to. The matter is not simple to decide, for it turns on the proper separation of what is determined by linguistic competence itself and broader psychology. The linguistic and philosophical literature is replete with disputes of just this character, as to whether or not, for example, 'logical constants'

have their formal interpretation for speaker-hearers, and, consequently, whether the formal inferences are sanctioned as part of natural semantic competence.

Before moving onto the concept of truth, let me respond to two likely objections to my reasoning so far.

The Reformulation Objection The first objection might go as follows. 'It is perfectly true that natural language is *opaque* in the way you suggest, in the way formal languages are designed not to be. Still, we can always reformulate natural language sentences in a formally perspicuous way. So, in effect, natural language is transparent'.[12] This objection is mistaken on two counts.

First off, the possibility of reformulation, even if always available, is irrelevant. Let it be the case that, for any sentence *S* that exhibits one or more of the features described under (i)–(iv) above, there is a distinct sentence *S′*, such that *S′* lacks all of the relevant features, but *S* and *S′* are still semantically equivalent. So what? It remains the case that the troublesome features are an aspect of the design of natural language, and so natural language remains opaque. The objection is of the same order as the thought that ambiguity is not a real phenomenon because of disambiguation. Reformulation exhibits the plasticity of language, but not that there is a class of primitive sentences somehow revelatory of the true transparent character of natural language.

Secondly, even though reformulation is often available, as with the resolution of ambiguity, other design 'defects' are recalcitrant. Position within a sentence, for instance, does not correlate with a definite semantic construal (see above). Moreover, not every required reformulation could be carried out in the relevant sense precisely because the reformulations would be in natural language that exhibits the 'defects' as design features. itself. Either way, the reformulation position needs to be argued for rather than assumed.

The Logical Form Objection The second objection might take the following form. 'Of course, semantics is not transparently encoded in natural language *as such*, but it is encoded in the logical form of language. So, natural language is transparent, but only at a particular level of analysis'.[13] If by 'logical form' one has in mind the structure imposed by a theoretical elaboration of natural language meaning, then the objection is not to the point. Granting the legitimacy of such approaches, we also grant a level of logical form, but the level is not thereby a property of natural language; it is a property of the theories. Natural language is a real phenomenon; it does not have design by fiat. The proper way to state the objection, therefore, is to claim that there is a real level of linguistic structure that is fully transparent, and so, at that level, natural language is like a formal language. A range of complex

[12] In different ways, this position has been articulated by Quine (1960), Davidson (1984), and numerous others. For an independent criticism of this kind of response to the apparent pathology of natural language truth, see Azzouni (2003).

[13] This position is mostly associated with Montague (1974) and much of natural language semantics in both the generative and non-generative traditions. See Larson and Segal (1995) and Hein and Kratzer (1998) for textbooks animated by the assumption.

issues arise with this objection (properly stated), which I have partly spoken about elsewhere, so let me be sparring (Collins 2007a).

Any level of structure of natural language will not be essentially answerable to a non-linguistic realm; evidence is required to establish such transparency, if, say, the intuitive truth conditions of a sentence are recorded in the logical form. It doesn't suffice to posit the level and stipulate that whatever is relevant to truth conditions is therein encoded. As regards the features adumbrated above, therefore, the bare idea of logical form as a syntactic structure that interfaces with semantic interpretation does not immediately remove the relevance of the features. Consider ambiguity. It might seem obvious that logical form is precisely where ambiguity is resolved, but this need not be so. It is common, for instance, to depict the two readings of *Every philosopher thinks he is a genius* as follows, where joint indexes indicate the referential dependence of *he* on the quantifier phrase (the bound reading) and disjoint indexes indicate a deictic construal of the pronoun:

(5)a [Every philosopher]$_i$ thinks he$_i$ is a genius
b [Every philosopher]$_i$ thinks he$_j$ is a genius

It is, however, far from obvious, to say the least, that indexes are part of natural language at all, logical form or not. For all the world they appear to be *our* way of marking the relevant relations of (in)dependence. If so, language, in the guise of logical form, might just be indifferent to the resolution of some ambiguities, perhaps because the resolution itself would involve extra-linguistic matters, such as a contextually salient person to support the (5b) reading.

3.3 The Concept of Truth in Formal and Natural Language

If I am right about the fundamental design differences between formal and natural language, then one might well expect a concomitant divergence between how to understand the concept of truth in natural and formal settings. The expectation is sound enough, I think, but just what the divergence should amount to is not obvious. First, let us go back to Tarski.

3.3.1 Tarski on Truth

Tarski (1936/1983, p. 155) introduces the idea of a semantic definition of truth in the following terms:

> [A] semantic definition [of truth is one]... we can express in the following words: a true sentence is one which says that the state of affairs is so, and the state of affairs indeed is so and so. From the point of view of formal correctness, clarity, and freedom from ambiguity ... [this] leaves much to be desired. Nevertheless its intuitive meaning and general intention seem to be quite clear and intelligible. To make this intention more definite, and to give it a correct form, is precisely the task of a semantical definition.

In a 'popular' paper some years later, Tarski (1944, p. 32) expresses the same thought: 'I happen to believe that the semantic conception does conform to a considerable extent with the common-sense usage'. It seems, therefore, that Tarski seeks to show that a formally precise conception of truth is available and that this conception is one that captures the familiar notion of truth (the 'correspondence' or 'classical' conception) as expressed in the natural vernacular. Familiarly, Tarski's position was not so naïve. He writes:

> A thorough analysis of the meaning current in everyday life of the term 'true' is not intended here... [Regarding] colloquial language... [t]he conclusion is totally negative. In that language it seems to be impossible to define the notion of truth or even to use this notion in a consistent manner and in agreement with the laws of logic. (Tarski 1936/1983, p. 152)

To pursue an account of truth in natural language will give rise to 'insuperable difficulties' (ibid, p. 164) and 'confusions and contradictions' (ibid., p. 267). The chief problem here is that natural languages are universal in that (i) the laws of logic hold; (ii) they contain their own truth predicate; and (iii) they possess the means of naming/describing sentences within the language. Such properties 'seem to provide a proof' that natural languages 'must be inconsistent' (ibid., p. 164–165). The 'proof' simply amounts to a 'liar sentence' being constructable, which when an instance of the T-scheme, gives rise to a contradiction ('α is true iff α is not true').

Roughly, there have been two kinds of response to Tarski's position, with numerous variations in each camp. The optimistic response is to think that Tarski was too pessimistic about the prospects of a formally precise and consistent account of natural language truth. The resulting brief here is to show that a contradiction-free notion of truth is specifiable in a more or less 'universal language'. The pessimistic response is to agree with Tarski. This latter position is often dubbed the 'inconsistency view', its claim being that natural languages are indeed inconsistent, as Tarski apparently claimed, but formal truth concepts may be adequately defined that serve the mathematical ends that were Tarski's sole concern.

I think both of these approaches are mistaken, at least as so baldly stated, but each finds succour in the faults of the other and so gains sustenance. Indeed, their shared flaw is the thought that natural language is, in some sense, a formal object, or at least should be conceived as a set of sentences that be assessed as consistent or not.[14] In

[14] Horsten (2011, p. 23) presents an argument for Peano arithmetic (PA) being the appropriate general background theory for inquiry into truth that might be taken to generalise to natural language beyond the formal languages Horsten has in mind. The argument is essentially that Gödel numbering demonstrates that PA is the least rich system that can encode the syntax of a language:

> Formulae of formal systems are just finite strings constructed from a finite alphabet. So in the final analysis, a syntactic theory concerning a language is just a bit of finite combinatorics. It is no real surprise that Peano arithmetic can do finite combinatorics!

This reasoning is sound, but does not apply to natural language precisely because its syntax is not a matter of mere finite combinatorics. Natural language sentences, *qua* syntactic objects, are not finite strings, but hierarchically structured objects, organised according to principles to be discovered. Gödel numbering, therefore, will be able to represent a set of English strings (morphophonemic forms), say, but it will not be able to represent the syntax of English; for instance, morphophonemically identical English strings might be syntactically divergent.

effect, the problem lies in the second condition on universalism, i.e., that the laws of logic hold in natural language. Let us first consider the inconsistency view, which, as we shall presently see, does not properly express Tarski's view.

3.3.2 *A Kind of Category Mistake: An Inconsistent Natural Language*

Tarski's claim, as expressed in the last quotation above, appears to be that natural languages are inherently inconsistent *qua* universal. Let us for the purposes of argument accept that any endeavour to make natural languages non-universal would simply be a distortion of them; one would, that is, no longer be targeting natural truth but some substitute notion, or a family of such notions spread over a hierarchy of sub-languages. There is an oddity here, though, for part of Tarski's reasoning against the prospects of an account of natural truth is precisely that natural language resists formal treatment. How, then, it might be wondered, can we even think that natural language is inconsistent? We say that a language/theory is formally inconsistent if every formulae is a theorem; the language is consistent if at least one formulae isn't a theorem. Such is why, *inter alia*, inconsistency is to be avoided, for it makes proof empty. Yet, without further ado, it makes precious little sense to say that every sentence of English is a theorem, or, for that matter, to say that any sentence is or isn't a theorem. To say such things, one needs to have in mind a proof theory that applies to the language, but there is no proof theory for English, and nor can one be proposed until English is suitably regimented. This is one of Tarski's points: natural languages are not so regimented, and that is why formal truth definitions do not apply. In short, Tarski's reasoning appears to be confused: he is pessimistic about defining truth for natural language, because it is inconsistent, but the concepts of consistency and inconsistency only apply to suitably formalised languages. Tarski appears to have made a kind of category error (cf. Burge 1979, p. 84, n. 2). This is not so, however. Note that Tarski, as quoted above, only speaks of natural languages *seeming* to be inconsistent. In the '44 paper, Tarski (1944, p. 21) has the following to say:

> At first blush it would seem that [natural] language satisfies... [the conditions sufficient for the generation of paradox], and that therefore it must be inconsistent. But actually the case is not so simple. Our everyday language is certainly not one with an exactly specified structure. We do not know precisely which expressions are sentences, and we know even to a smaller degree which sentences are to be taken as assertible. Thus the problem of consistency has no exact meaning with respect to this language. We may at best only risk the guess that a language whose structure has been exactly specified and which resembles our everyday language as closely as possible would be inconsistent.

Nothing has happened in the decades since Tarski wrote these words for us to question his judgement; indeed, the claims should now be obvious. Two points bear emphasis. Firstly, the great advances made in linguistics since the 1950s have not led to a conception of natural language as akin to a formal language that supports properties of

consistency or inconsistency. On the contrary, they have led to elaborations of general principles of design and interpretation that reveal the peculiar properties (from a formal perspective) exemplified above. Formal techniques are often employed, but the object of inquiry does not thereby become formal. Secondly, Tarski's 'guess' has come to fruition in that 'fragments' of natural language are formalised and these may be consistent or not, but exactly as Tarski suggests, this does not amount to natural language itself being consistent or not. A 'fragment' is not a proper part of a natural language deductively related to the rest of the language as a whole, but an abstraction made over a restricted class of constructions, so any properties the fragment has, once formalised, are not inherited by the whole. Still, might there be other, perhaps more relaxed senses, in which natural language might be inconsistent?[15]

Eklund (2002, 2007) endorses an 'inconsistency view' according to which the meaning of the truth predicate is constituted by its inferential role, i.e., the rules that license inferences to, from, and between truth-involving claims. Of course, it is a phenomenon of such inferences that we fall prey to the liar, whether consciously or unconsciously. Thus, if our inferences are inherently heir to the possibility of liar-type reasoning, and such inferences are constitutive of truth, then truth is, indeed, a paradoxical, inconsistent notion. Let me be all too brief.

First off, one might justly be leery of 'inferential role' accounts of the meaning of any word at all, including 'logical constants' such as *and* and *or*. The meaning of a word should *explain* why certain inferences featuring the word hold, not the other way around. The operative inferences are not ones anyone need actually make, but those competent agents should make, given the attributed content. There is, therefore, no way of isolating the set of relevant inferences independently of specifying an invariant meaning for the relevant word. On the other hand, one may specify the content of a word without presupposing any inference at all, let alone a specified set of inferences. What holds for *and* holds for *true* too (cf., Collins 2002). The point is simple. The relevant inferences cannot be those we do make, or even those we are disposed to make, because, *inter alia*, both classes involve all kinds of mistakes. What individuates the relevant instances, therefore, is some specification of that which is invariant over 'correct' inferences. Such invariance, though, is not constituted by any inferential behaviour at all, which is very messy, but our conception of what counts as 'correct'. Secondly and correlatively, I don't see any principled way of isolating the inferential role of the truth predicate to just the meaning-constituting inferences (ditto every other word, for that matter). One might follow Horwich (1998) and claim that an explanatorily basic use pattern corresponding to a disquotational schema constitutes the meaning of *true*. That won't do, though, for many truth predications just do not submit to any such schema, e.g., *Most/many/few/several things Bob said are true* (see Collins 2010, for detail).

[15] There are a range of 'inconsistency views', ranging from Tarski, as discussed above, and more recently Chihara (1979), Azzouni (2003, 2007), Grover (2005), and Burgess and Burgess (2011). Due to the demands of space, below I shall only consider versions of the view due to Eklund and Patterson, but I intend my remarks here to apply generally, where applicable.

An intriguing alternative inconsistency view is offered by Patterson (2007, 2009). Patterson takes it as given that liar-type reasoning is a real phenomenon, an aspect of our truth-competence, one might say. If we further assume that semantic competence is based on the cognition of a truth theory—most minimally, that our knowledge of meaning is knowledge of truth conditions—then it would appear to follow that our semantic competence is based upon an inconsistent theory, i.e., a false theory about the truth conditions of our sentences. Patterson commendably bites the bullet and argues that this entails that our sentences as recognised by ourselves and others do not have consistent truth conditions; they merely appear to, but since the illusion is mutual, we communicatively get along. The illusion is not shattered in our day-to-day business simply because of 'logical indolence' or indifference, i.e., we do not often face liar-type situations or reflect on their possibility.

Two features of this position are very attractive. Firstly, it takes liar-type reasoning to be a real phenomenon that must be accounted for, as opposed to diagnosed as essentially avoidable faulty reasoning. From a naturalistic perspective, such inconsistency is a feature of human cognition, not a rectifiable slip such as a common mistake in arithmetic. Secondly, although admitting liar-type reasoning, Patterson's position does not render our very thought incoherent; on the contrary, we simply employ a false (inconsistent) theory about our language construed as a public realm of sentences. In some sense, therefore, our thoughts are incoherent, but only in light of a metaphysically overblown conception of language. That said, I am doubtful in the first place of the idea that language is best modelled as a set of sentences about which we might have a consistent or inconsistent theory. Again, unless natural language sufficiently resembles a formal language, then no issue of consistency or inconsistency arises. Still, if all that is involved in the 'inconsistency view' is the idea that liar-type reasoning is a natural feature of our truth-competence, then it seems undoubtedly correct (see below). That notion, though, can be separated from the idea that languages/theories constitutive of human linguistic competence are consistent or not; that is, acknowledgement of liar-type reasoning as an unavoidable trait of our thinking neither entails nor presupposes the idea of a natural language as a set of symbol strings that may possess the properties of a formal language. We can further see the virtues of an inconsistency view, at least if read in the minimal way just suggested, by looking at what I called the 'optimistic' approach to the apparent dilemma Tarski bequeathed us.

3.3.3 Grounds for Optimism?

Recall that the optimistic response to Tarski takes his negative conclusions about the prospects of an account of natural truth to be a challenge to devise more sophisticated or flexible means to formalise natural truth. The rationale for this approach is not hard to see. The conclusion that our colloquial truth concept is intrinsically inconsistent threatens to wreck any notion of sound inference within natural language. Indeed, if Patterson's reasoning is right, then seriously accepting Tarski's position would appear

to lead to semantic nihilism. There are roughly two ways of being 'optimistic'. Firstly, following the lead of Kripke (1975), one may hope for a genuinely universal account of truth for a formal language that evades inconsistency, where the universalism would justify the formal account being an apt analogue for natural truth.[16] Halbach (2010, p. 333) expresses this dominant view as follows:

> Like many other philosophers, I see the theory of truth for the language of arithmetic as the starting point for developing a theory of truth for other, usually more comprehensive languages as base languages, and perhaps eventually for natural languages.

So, this view does not target natural truth directly, but attempts to offer accounts that have the promise of being scaled up. It is the norm here to assume that the relevant (axiomatic) theory will render truth a consistent notion whatever the language. It is worth noting, though, that this is by no means mandatory. One could think, for instance, that truth is consistently definable for a range of increasingly comprehensive languages, but not for the embedding natural language.[17]

---

[16] Kripke (1975, pp. 80–81), of course, acknowledges that 'we still cannot avoid the need for a metalanguage. . . the goal of a universal language seems elusive. . . [The truth predicate in the bivalent language of the theory] expresses the genuine concept of truth' for the natural language of the theory as a whole. The problem arises from the fact that Kripke's central claim that neither the liar nor its negation is true (the liar is 'ungrounded') is itself neither true nor false (it is 'ungrounded' because the liar is). Kripke (ibid., p. 80) does suggest that the consistent trivalent model in which truth claims are grounded might depict 'natural language at a stage before we reflect on the generation process associated with the concept of truth, the stage which continues in the daily life of nonphilosophical speakers'. There are at least two problems with this thought. Firstly, as McGee (1991, p. 91–92) points out, it is most 'likely that the order of development is just the opposite of what Kripke describes', i.e., the semantic competence of the 'nonphilosophical' is based on a bivalent logic, with the trivalent logic of the model being a feature of our competence, if at all, only 'as a highly sophisticated response to problems that are only visible upon philosophical reflection'. Secondly, even if Kripke's model does serve as an account of prereflective competence, once reflective competence is achieved, we would appear to be landed with a hierarchy of truth concepts, which belies Kripke's very endeavour of consistently capturing the universalism of natural truth.

[17] Horsten (2011, p. 24), who shares Halbach's general position, insists that '[a]xiomatic theories of truth should be *sound*. They should prove only sentences that we instinctively and immediately accept or, after reflective consideration, can come to see to be correct'. To be sure, one hardly wants a theory to fail to predict the relevant phenomena, but it is here assumed that the phenomena do not include contradictions. Horsten motivates the general axiomatic approach by an analogy with classical mechanics, where the notions of mass, force, acceleration, etc. are treated functionally, in the sense that the mechanics does not specify essences, but defines concepts that track the relevant phenomena in a way that is predictable from the mathematical interaction of the defined concepts. Thus: 'Tarski developed a theory of truth that describes the functioning of the concept of truth, which puts truth to use. Traditional theories of truth, by contrast, entail little about the use of the concept of truth, just as Aristotle's theory of motion does not make precise predictions about the velocity of falling bodies' (ibid., p. 16). There is a crucial difference, though, between the two theories. Classical mechanics is not intended to recover *any* intuitive conception of bodies and forces; nor is it answerable to any such conception. The job of the theory is to capture the actual phenomena, much of it peculiar and unpredicted (this is so regardless of the mechanics being an 'extension of common sense'). So, to retain Horsten's analogy and insist that a theory of truth should be sound appears to express a commitment to a certain consistent idealisation of natural truth that may serve mathematical purposes, which is not to seek an empirical theory that captures how we can in fact

An alternative, but still optimistic view does not seek to eliminate inconsistency but to quarantine it.[18] Such approaches agree with Tarski that paradoxical reasoning is a genuine phenomenon intrinsic to colloquial truth, but additionally claim that paradox is avoidable in the sense that we can gain a clear diagnostic of how inconsistency arises and, crucially, when it does not arise. Thus, we may rightly acknowledge the naturalness of paradox without it threatening the very coherence of our colloquial truth concept and the reasoning employing it.

The chief problem facing the first brand of optimism is precisely what appears to be correct about the inconsistency views and the second kind of optimism. Paradoxical reasoning does indeed occur and the most simple and natural position on such phenomena is to hold our natural truth concept responsible for it. The kind of scaling up Halbach suggests would appear to create an artificial surrogate of natural truth precisely because it avoids what, for all the world, looks to be an inescapable feature of the concept. The best reason for viewing natural truth as unavoidably paradoxical is its *inherent riskiness*.

I have elsewhere characterised truth as a device of *opaque metarepresentation* (Collins 2007b). By this I mean that truth predications state a property of representations (represent representations) but not in a way that necessarily entails an understanding of the representation (hence opaque).[19] In simple terms, this means that one can predicate truth of representations without having any essential cognitive access to the content of those representations, i.e., it is not part of one's semantic competence with truth that one's general competence covers that claimed to be true. So, trivially, one can think a sentence or assertion is true without understanding the sentence or assertion itself; for instance, one may simply trust the source. Similarly, one can generalise and claim that all theorems of ZF set theory are true, without knowing any set theory at all. To be sure, if one does predicate truth to some collection (a class of theorems, everything Bob says, etc.), then one is rationally obliged to *assent* to anything identified as falling within that class, but one cannot be rationally

employ truth, often inconsistently. In other words, in the case of truth, unlike the case of mechanics, there is no phenomenon beyond our conception, so if the analogy is to hold, one cannot pre-empt the character of that conception.

[18] The approaches I have in mind here include the 'contexualist' accounts of Parsons (1974) and Burge (1979) and the 'Austian' account of Barwise and Etchemendy (1987), which all admit circularity. The position of the 'rule of revision' account (Gupta and Belnap 1993) probably should belong in the same camp, but it is most often associated with the kind of model presented by Kripke (1975).

[19] An intimately related point is that there is no proper distinction in natural language between direct and indirect reporting, for speech marks or the like are part of sophisticated orthography, not language. Thus, I read (i) as being perfectly acceptable:

(i) It is true that arithmetic is incomplete. It is what Gödel proved, but I haven't a clue what it means.

In contrast, I find first-person ascriptions of a similar kind decidedly odd:

(ii) I believe that arithmetic is incomplete. It is what Gödel proved, but I haven't a clue what it means.

This suggests that self-ascription of this kind is 'transparently metarepresentational'.

obliged to *assert* it, for one need not be in a position to understand it, and assertion depends upon some understanding.[20]

If all that is so, then truth is inherently risky exactly because one need not be able to track or recognise what one is claiming to be true, and so it is not a feature of being competent with truth that one can, let alone must, guarantee one's freedom from inconsistency. As Kripke (1975) points out, one can fall into paradox by accident. I presume that happens regularly unbeknownst to the parties to the contradiction. So, if truth is a device of opaque metarepresentation, then the possibility of paradox immediately follows.

Such inherent riskiness of truth also tells against the second brand of optimism that acknowledges the intrinsic paradoxicality of truth, but seeks to quarantine the strictly pathological. Any such quarantine looks impossible exactly because one can fall into paradox through no rational fault of one's own, merely by generalising with the truth predicate. Of course, it is possible to spell out the conditions under which paradoxes arise even if truth is opaquely metarepresentational, and to that extent, the paradoxes can be isolated. My point, however, is that any such spelling out is orthogonal to competence with truth and hardly amounts to a rational burden on a speaker's competence with truth predications. If it were, then accidental paradoxes would signal some form of semantic deviance, but the competence called upon in such cases is perfectly normal and in order; after all, the paradox is an *accident*.

3.4 Conclusion

If the above is on the right lines, we may conclude that Tarski was right about the inherent pathology of natural truth. This conclusion should not be shocking, however, for it does not amount to the claim that natural languages are inconsistent. *That* claim makes little sense, unless it merely means that paradox is a genuine phenomenon of natural language. Natural languages lack the transparency of formal languages, so it is incorrect to say that they are either consistent or not. What can be said is that humans in their reasoning are all too often contradictory, and the attempt to save them is forlorn, especially given the inherent riskiness of truth predications. Truth serves us well as, *inter alia*, a metarepresentational device of generalisation, but that comes at a price of potential paradox.[21]

References

Azzouni, J. (2003). The strengthened liar, the expressive strength of natural languages, and regimentation. *Philosophical Forum, 34,* 329–350.

[20] See Ziff (1972) for an unduly neglected account of understanding in line with these remarks.

[21] Thanks. My thanks go to Monika Gruber, Friederike Moltmann, Hartry Field, and Doug Patterson.

Azzouni, J. (2007). The inconsistency of natural languages: How we live with it. *Inquiry, 50,* 590–605.
Barwise, J., & Etchemendy, J. (1987). *The liar: An essay on truth and circularity*. Stanford: CSLI Publications.
Burge, T. (1979). Semantical paradox. *Journal of Philosophy, 76,* 169–198.
Burgess, A., & Burgess, J. (2011). *Truth*. Princeton: Princeton University Press.
Chihara, C. (1979). The semantic paradoxes: A diagnostic investigation. *Philosophical Review, 88,* 590–618.
Chomsky, N. (1986). *Knowledge of language: Its nature, origin, and use*. Westport: Praeger.
Collins, J. (2002). Truth or meaning? A question of priority. *Philosophy and Phenomenological Research, 65,* 497–536.
Collins, J. (2007a). Syntax, more or less. *Mind, 116,* 805–850.
Collins, J. (2007b). Declarative thought, deflationary truth and metarepresentation. In D. Greiman & G. Siegwart (Eds.), *Truth and speech acts*: *Studies in the philosophy of language* (pp. 157–177). London: Routledge.
Collins, J. (2010). Compendious assertion and natural language (generalized) quantification: A problem for deflationary truth. In C. Wright & N. Pederson (Eds.), *New waves in truth* (pp.81–96). London: Palgrave Macmillan.
Davidson, D. (1984). *Inquiries into truth and interpretation*. Oxford: Clarendon Press.
Devitt, M. (2006). *Ignorance of language*. Oxford: Oxford University Press.
Eklund, M. (2002). Inconsistent languages. *Philosophy and Phenomenological Research, 64,* 251–275.
Eklund, M. (2007). Meaning-constitutivity. *Inquiry, 50,* 559–574.
Grover, D. (2005). How significant is the Liar? In J. C. Beall & B. Amour-Garb (Eds.), *Deflationism and paradox* (pp. 177–202). Oxford: Oxford University Press.
Gupta, A., & Belnap, N. (1993). *The revision theory of truth*. Cambridge: MIT Press.
Halbach, V. (2010). *Axiomatic theories of truth*. Cambridge: Cambridge University Press.
Heim, I., & Kratzer, A. (1998). *Semantics in generative grammar*. Oxford: Blackwell.
Hornsten, L. (2011). *The Tarskian turn: Deflationism and axiomatic truth*. Cambridge: MIT Press.
Horwich, P. (1990). *Truth*. Oxford: Blackwell.
Horwich, P. (1998). *Meaning*. Oxford: Clarendon Press.
Katz, J. (1981). *Language and other abstract objects*. Oxford: Blackwell.
Katz, J., & Postal, P. (1991). Realism vs. conceptualism in linguistics. *Linguistics and Philosophy, 14,* 515–554.
Kripke, S. (1975). Outline of a theory. *Journal of Philosophy, 72,* 690–716. (References to the reprint in R. Martin (Ed.) (1984) Recent Essays on Truth and the Liar (pp. 53–81). Oxford: Oxford University Press).
Larson, R., & Segal, G. (1995). *Knowledge of meaning: Introduction to semantic theory*. Cambridge: MIT Press.
Levin, B. (1993). *English verb classes and alternations: A preliminary investigation*. Chicago: Chicago University Press.
McGee, V. (1991). *Truth, vagueness, and paradox: An essay on the logic of truth*. Indianapolis: Hackett.
Montague, R. (1974). *Formal philosophy: Selected papers*. R. Thomason (Ed.). New Haven: Yale University Press.
Oliver, A. (1999). A few remarks on logical form. *Proceedings of the Aristotelian Society, 99,* 247–272.
Parsons, C. (1974). The liar paradox. *Journal of Philosophical Logic, 3,* 381–412.
Patterson, D. (2007). Understanding the liar. In JC Beall (Ed.), *Revenge of the liar: New essays on the paradox* (pp. 197–223). Oxford: Oxford University Press.
Patterson, D. (2009). Inconsistency theories of semantic paradox. *Philosophy and Phenomenological Research, 79,* 387–422.
Quine, W. V. O. (1960). *Word and object*. Cambridge: MIT Press.

Soames, S. (1984). Linguistics and psychology. *Linguistics and Philosophy, 7,* 155–179.
Tarski, A. (1936/1983). *The concept of truth in formalized languages*. In J. Corcoran (Ed.), Logic, semantics, metamathematics: Papers from 1923 to 1938 (pp. 152–277). 2nd Edn. (trans: J. H. Woodger). Indianapolis: Hackett.
Tarski, A. (1944). The semantic conception of truth. *Philosophy and Phenomenological Research, 4,* 341–376. (References to the reprint in B. Linsky (1952), Semantics and the Philosophy of Language (pp. 13–47). Chicago: University of Illinois Press).
Vision, G. (2004). *Veritas: The correspondence theory and its critics*. Cambridge: MIT Press.
Ziff, P. (1972). *Understanding understanding*. Ithaca: Cornell University Press.

Chapter 4
Truth and Trustworthiness

Michael Sheard

Abstract In the course of ordinary communication, people transmit messages (i.e., say things) which may involve the application of a truth predicate. The receiver of such a message needs to have a method which allows the extraction of non-truth-theoretic information from uses of the truth predicate; such a method can be modeled with an axiomatic system. On close examination, the choice of which axiomatic system to employ can be seen to depend on whether or not the source of the message is considered trustworthy—that is, whether the information in the message can simply be accepted, or if it must first be examined for consistency with previously known information and, on the basis of that determination, possibly be rejected. This paper explores some of the consequences involved in this framework.

4.1 Setting the Problem

Three philosophers—a deflationist, an advocate of the correspondence theory of truth, and a believer in the coherence theory—go out to dinner. They have a riotously good time debating politics, sports, gossip about other employees of their university—anything except philosophy. Throughout the evening, their conversation is peppered with phrases like "That's true!", "That can't be true," and "Nothing that man has ever said is true." Remarkably, they all understand each other completely at those moments, even though they do not agree in the least on what it means for something to be "true". How is it that they are able to communicate so effectively using a concept about which they so thoroughly disagree?

To a deflationist, of course, there is nothing surprising here. Since truth (for a deflationist) is a logical or linguistic concept that operates at a surface level, of course it can be employed in ordinary conversation without need of deep analysis or occasion for disagreement. But while the deflationist may not be surprised *that* the process works, such confidence is not the same as having an explanation of *how* the process works. Meanwhile, a substantivist may maintain very emphatically that at its core, truth is a much deeper phenomenon that the deflationist mistakenly dismisses.

M. Sheard
Rhodes College, Memphis, TN, USA
e-mail: sheardm@rhodes.edu

T. Achourioti et al. (Eds.), *Unifying the Philosophy of Truth,* Logic, Epistemology, and the Unity of Science 36, DOI 10.1007/978-94-017-9673-6_4

Nonetheless, all but the most extreme substantivist will have to admit that, whatever the deep issues may be, those issues are not engaged in a meaningful way when people use words like "true" in the course of ordinary conversation. While truth may be deep, it must have some shallow features that allow it to serve as a mechanism of day-to-day speech. In either case, we still have work to do to uncover how truth functions in ordinary conversation. We need a logical system that can model the way the concept of truth is used to convey information.

If we were to stipulate that a truth predicate can only be applied to sentences that do not involve truth itself, then there would be few problems. In formal terms, we could adopt as an axiom the Tarski T-sentence T('A') ↔ A for each sentence A which does not contain the truth predicate (e.g., " 'snow is white' is true if and only if snow is white"), and modulo a simple ability to unpack sentences, we would be done. This restriction is unrealistic, however. I ought to be able to say that everything Bob said in his lecture today was true, even if one of the things Bob said was that Emily's statement was true. Applying the truth predicate to sentences that themselves involve the truth predicate is completely natural. We need an analysis which is robust enough to account for an untyped truth predicate.[1]

Obviously, however, insistence on an untyped truth predicate raises other problems. Given our ability to formulate Liar-like sentences, both in ordinary language and in its formal analogues, we know that there are instances of the unrestricted T-sentences which lead in short order to contradictions. One response (one which I will admit I find somewhat attractive) is to grant that people actually work day-to-day on a basis of the unrestricted T-sentences, and simply do not carry their reasoning with inconsistent hypotheses far enough to derive explicit contradictions.[2] As a practical matter, this is a reasonable position to maintain about actual people, as part of a much more general phenomenon—probably most people hold inconsistent beliefs of one sort or another, yet rarely get into trouble simply because they do not reason through to an explicit contradiction. While there may be merit in this hypothesis, it is unsatisfactory as an exercise in logical modeling. If we adopt this attitude, there appear to be only two places to go next. One is to announce that humans are irrational creatures and walk away, which does nothing useful to address the original question about how people are able to communicate effectively using the concept of truth. The other option is to begin an empirical study of how real people manage, maintain, and apply actual inconsistent assertions about truth, but such a study is more suitable for psychology than for philosophy. Moreover, there is nothing in such a project that is specific to the study of the logic of truth *per se*, since the same questions could be asked any time someone holds inconsistent beliefs in any context. Neither of these approaches advances our understanding of how a truth predicate can be used in communication.

Instead, then, let us assume that each person has some logically consistent framework for processing information that is transmitted by means of a truth predicate.

[1] Kripke (1975) makes this point in much more detail.

[2] Horwich (1990) speaks about our *inclination* to accept the T-sentences.

Becoming more formal, we can capture this framework as a set of axioms and rules of inference concerning truth which are overlaid on whatever base of factual information the person has and whatever ordinary rules of logical inference the person employs. When someone speaks, we can think of what he says as a message which is transmitted with the intent that it be added to the listener's base of factual information. The listener applies her own truth-specific axioms and rules to extract the non-truth-theoretic content of the message. Our goal will be to explore what axioms and rules are needed for this process, and how they are to be applied.

4.2 The Form of a Message

Ideally, a successful analysis would account for all possible uses of a truth predicate in ordinary communication. There are good reasons, however, for believing that such a far-reaching goal may simply be impossible. Our aspirations will be more modest at the outset, so let us focus on three very common uses of truth in communication:

1. Direct attribution of truth
2. Denial
3. Generalization

Let us briefly consider each of these in turn.

A direct attribution of truth states that a specified sentence is true, and can take several forms. Direct attribution can occur in quotational (or equivalent) form, in which the sentence to which truth is being attributed is immediately displayed: "The sentence 'Amsterdam is in the Netherlands' is true". While there are grammatical, linguistic, and perhaps logical differences between this example and "It is true that Amsterdam is in the Netherlands", it is hard to argue that they convey any different information, either implicitly or explicitly.[3] Note that in this regard attributions of truth differ from some other kinds of linguistic communication. There is a meaningful distinction between "Catherine said 'Amsterdam is in the Netherlands' " and "Catherine said that Amsterdam is in the Netherlands"—are we repeating her exact words, or paraphrasing?—but there is no difference in the context of our analysis of the uses of truth in communication.

Direct attribution need not be quotational or nearly-quotational. It can proceed by indexical reference: "That is true", where the referent of "that" is unambiguous in context. It can proceed by some sort of definite description: "The first sentence in Susan's essay is true." It can also proceed—perhaps a bit artificially—via description of the construction of the sentence to which truth is attributed; such an approach is valuable in creating unassailably self-contained examples of self-reference, like the Liar sentence.

[3] The explicitly quotational version as written does imply, or at least seems to assume, that sentences are appropriate bearers of truth. That discussion can be left for a different occasion.

One may wonder why direct attribution of truth would pose a problem at all in the context of communication. At some level, of course, it does not. The redundancy theory, the disquotational theory, the prosentential theory, and the minimal theory of truth will all tell you that to communicate a direct attribution of truth is to communicate the underlying sentence itself. If we could stop there, the answer would be fully sufficient. When we build a theory robust enough to handle denials and generalizations, however, the mechanisms we put in place may not be sufficient to achieve what we would hope to achieve in our analysis of direct attribution.

It is tempting to regard a denial of truth as an immediate variant of a direct attribution of truth—no more and no less problematic. Certainly attributing falsity to a statement has much the same feel as attributing truth—the same action with just a negation operator inserted at some appropriate point. In fact, though, the nature of that negation operation, and the question of exactly what is the "appropriate" point for its insertion, turn out to make a huge difference in the logical analysis; this observation will come into sharper focus later. For now, let us just look at what happens when we entertain the possibility that a sentence might—for whatever reason—lack a truth value, that is, might be neither determinately true nor determinately false. Certainly there are theories of truth that include the possibility of indeterminate truth values in their structure. A direct attribution of truth unquestionably asserts that the sentence in question has a determinate truth value. But if someone says that a sentence is "not true", does that mean "false or possibly indeterminate"? Or does it mean "determinately false"? The former seems a more reasonable reading on purely logical grounds (given that we already accept that there is a category of sentence which is neither true nor false), but possibly not in the spirit of the way the expression might be used in ordinary communication.[4] If someone chooses to say that a sentence is "not true" —rather than "meaningless", or "impossible to determine", for example—might we not be justified in the inference that the speaker believes that the sentence in question does indeed have a determinate truth value? There are many other ways to signal that a statement fails to have a truth value for some intrinsic reason, which are more specific and perhaps more sensible than simply saying it is "not true". Alternatively, if we decide that we are not justified in assuming that "not true" excludes the possibility of indeterminacy, then can we run the train in the other direction, and also interpret "false" not as "determinately false" but simply "not true"? After all, if asked, most people would define the word "false" as "not true".

We may choose to dismiss these questions as a linguistic muddle arising from the ambiguity of ordinary language, but we will have no such luxury when we try to formalize the logic which is used in the process of communication. In any case, what is apparent is that the question of denial in the use of a truth predicate is *not* merely the mirror image of the question of direct attribution of truth. For these reasons, it is appropriate to keep denial as a separate prototypical example.

[4] Similarly, in English, "I don't think it will work" really means "I think it will not work."

Last, we come to generalization. Some philosophers have claimed that generalization is the whole reason that there is a problem about the theory of truth at all; if all we had available in our language were direct attributions and denials, most of the problems would disappear via some sort of redundancy interpretation, quibbles around the margins notwithstanding. At the very least, the problems to be solved without generalization would be substantially smaller in number and magnitude. Generalization applies the truth predicate to a defined list of sentences, where membership on the list is given by specification of a shared property rather than enumeration: "Every sentence in the book is true." It is worth exploring different features that the specification can have. If the specification unambiguously specifies a finite list, then the generalization can be read as a mere stand-in for the conjunction of direct attributions of truth to each of the sentences individually: "The first sentence in the book is true and the second sentence in the book is true and . . ." If the list is actually or potentially infinite, then this interpretation cannot be sustained: the claim "Every theorem of Peano Arithmetic is true" attributes truth to a list of sentences which is not only infinite but also provably undecidable. Finally, there is the situation in which the list, while presumably finite, is not known or not yet established. "Everything Bob has ever said about chemistry is true" specifies a large finite list of which neither the speaker nor the listener is likely to know the exact membership. "Everything Bob will ever say about chemistry will be true" specifies a list which does not (yet) even have an exact membership. Why the speaker would make such a reckless claim may be an interesting question, but it in no way impinges on the use of the truth predicate itself as a way of conveying information. The content the speaker aims to convey is clear enough.

4.3 Logical Systems

Now we need a logical system for our idealized listener to apply to decode incoming messages. There are three standout candidates for the role: the systems known in the literature as FS, KF, and VF.[5] (To be precise, these designations all denote systems of arithmetic augmented with a truth predicate, in which the underlying axiomatization is Peano Arithmetic. Working informally, I will use the same designations for the corresponding truth-theoretic apparatus overlaid on any system with sufficient expressive power to permit discussion of linguistic elements like sentences, which is a necessary precondition for applying a truth predicate.) While all three have been studied thoroughly for their truth theoretic properties, and KF in particular has been the focus of some discussion concerning its suitability as a theory *about* truth (most notably by Reinhardt (Reinhardt 1986)), there has not been much discussion of their

[5] For a comprehensive presentation of all of these systems, see Halbach (2011).

relative merits as a tools for the kind of communication described here. Let us take a quick look at the principal features of each.[6]

FS is Halbach's axiomatization (Halbach 1994) of a theory which replaces the T-sentences with corresponding rules of inference:

From A, deduce T(‘A’)
From T(‘A’), deduce A
From ¬A, deduce T(‘¬A’)
From T(‘¬A’), deduce ¬A

In the analysis that follows, it is important to remember that these rules can only be applied to sentences that are given as axioms or have already been proved; they may not be used in conditional subproofs. Thus in general there is no deduction theorem for FS.

KF is Feferman's axiomatization (Feferman 1991) of Kripke's basic fixed-point model. It is most smoothly axiomatized with both a truth predicate and a falsity predicate. The salient axioms are compositional; for example, here are the axioms for the truth and falsity of negations, conjunctions, and truth-attributions:

T(‘¬A’) ↔ F(‘A’)
F(‘¬A’) ↔ T(‘A’)
T(‘A&B’) ↔ T(‘A’) & T(‘B’)
F(‘A&B’) ↔ F(‘A’) ∨ F(‘B’)
T(‘T(A)’) ↔ T(‘A’)
F(‘T(A)’) ↔ F(‘A’)

There are similar sets of axioms for disjunctions, material conditionals, quantifiers, and falsity-attributions. In addition, KF contains the T-Consistency axiom: ¬(T(A) & T(¬A)). One consequence of the T-Consistency axiom is that by induction on the build-up of formulas, one can prove T(‘A’) → A (that is, one direction of the biconditional T-sentence) for each sentence A. Critically, KF does not assert the truth of validities of first-order logic: for one significant example, not every sentence of the form T(‘A∨¬A’) is provable.

VF is Cantini's axiomatization (Cantini 1990) of the Kripke/ van Fraassen supervaluation model. The central axiom of VF is again the one direction of the T-sentences: T(‘A’) → A, for all sentences A. Unlike KF, it also contains axioms guaranteeing that the set of true sentences is closed under logical implication, although in exchange it gives up some of the principles of compositionality, such as T(‘A∨B’) ↔ T(‘A’) ∨ T(‘B’).

[6] Each system has additional axioms, including ones that establish which basic sentences are to be declared true. For simplicity, I will suppress mention of most of those here. Moreover, my notation is intentionally over-simplified in some regards to improve readability.

4.4 Trustworthiness

If a speaker conveys a message with intent that it be added to a listener's base of factual knowledge, the listener faces a decision. If (in our idealized model) it can be assumed with certainty that the message will not be in conflict with information already known to the listener, then the listener can decode the message and add the new message directly to the knowledge base. I will call the source of the message in this case *trustworthy*. Alternatively, if there is no assumption that the message will not be in conflict with preexisting knowledge, then the user may need to evaluate the information as an additional step in the process, and perhaps may even draw inferences from the outcome of that evaluation (such as perhaps concluding that the speaker is a liar). I will call such a source *untrustworthy*. The distinction will matter in selecting an appropriate logical system for the task.

4.5 Decoding the Messages

Let us look first at the situation of a trustworthy source, and consider in turn each of the three principal kinds of messages. Imagine that a trustworthy source sends a message which makes a direct attribution of truth, which we can represent as T(‘A’). All three systems are strong enough extract the information conveyed. In FS, one applies the rule of semantic descent to T(‘A’) to derive A. In VF or KF, one grabs the axiom/theorem T(‘A’)→A and applies *modus ponens*.

As I have suggested already, the situation for a denial is a little more delicate. If we represent a denial as ¬T(‘A’), then FS is fully equipped to handle it: there is a rule of inference to derive ¬A. The systems VF and KF, however, pose more of a stumbling block. Both systems prove all instances of T(‘A’)→A, but not in general A→T(‘A’), so that there is no generally valid way to deduce ¬A from ¬T(‘A’). Here, perhaps the best solution is to fall back on the suggestion to render a denial as T(‘¬A’) rather than ¬T(‘A’), which solves the problem immediately. It also can be applied uniformly to include FS, since the inference from T(‘¬A’) to ¬A is also valid there.

Finally, generalization turns out to pose little problem. The formal representation of the use of truth for generalization as we have defined it has the form ∀x(R(x)→ T(x)). If ‘A’ is any sentence which falls into the list defined by R(x), then all three systems will allow one to move directly from R(‘A’) and ∀x(R(x)→ T(x)) to T(‘A’). One can then apply the system's method for handling direct attribution of truth to extract A.

If a source is not trustworthy, then any message received from it needs to be checked for the possibility of inconsistency with existing knowledge (or internal self-contradiction) before it can be added to the listener's base of knowledge. For systems like KF and VF which are purely axiomatic—i.e., having no additional auxiliary rules of inference—this is simple, since they are closed under *reductio ad absurdum*: if sentence B (whether truth-theoretic or not) is inconsistent with existing

knowledge, then already ¬B is a logical consequence of existing knowledge. In principle, then, under these systems screening information from an untrustworthy source is logically straightforward. Any sentence, whether containing a truth predicate or not, can be accepted if it is not a direct contradiction of a known consequence of existing knowledge.

For FS, with its auxiliary rules of inference that do not admit a deduction theorem, such a *reductio* is not available in general. To take an extreme example, let L be the Liar sentence. If one tries to add L to an existing knowledge base, then the rule of semantic assent applied to L produces T('L'), and an immediate contradiction. Nonetheless, ¬L is not a theorem of FS, nor can it even be a member of any consistent set of sentences closed under the rules of FS, since it leads to a contradiction of its own in much the same manner. In the end, however, the difference is of minor import for the purposes of evaluating messages. As a model, one can envision a person who uses the rules of FS as a truth mechanism provisionally accepting a message from an untrustworthy source, and then testing to see if a contradiction emerges. If one does, then the new information and anything that followed from it is rejected and the *status quo ante* is restored. If no contradiction emerges, then the new information remains. For a simple decision about accepting or rejecting a message, the full force of a *reductio* argument is not needed.

4.6 Next Steps

So far, the distinctions proposed here may seem to be much ado about little. One does not have to go much further beyond the restricted realm of direct attribution, denial, and generalization, however, to reach a point where the complications begin to mount up. Consider the case of a corporate rumor claiming that Smith lied to the president of the company, to which the wise and trustworthy old-timer announces, "If the rumor is true, then Smith will be fired." The wise and trustworthy old-timer's statement has the form T('A') → B, but the obvious inference to A → B is beyond the scope of the basic inferential mechanism of *any* of our three systems. The system FS fares slightly better than the others, in that if it turns out to be the case that Smith did indeed lie to the president of the company, then FS can deduce T('A') from A by semantic ascent, and then draw the accurate conclusion that Smith is on his way out. Without semantic ascent, KF and VF cannot do even that, unless someone chooses to announce specifically that the rumor was true.

Before reading too much into the preceding example, however, note that FS tends to underperform in situations where a message from an untrustworthy source can be examined logically not just for acceptance or rejection, but as a basis of logical inferences to acquire new additional information. Elsewhere (Sheard 2008) I have offered as an example a variation of a problem of Smullyan (Smullyan 1978):

> There is an island on which every inhabitant either always tells the truth or always lies, but the two types are otherwise indistinguishable. You encounter two of them; one says "At least one of us always lies." Which type is each of them?

In either VF or KF, some easy conditional reasoning allows one to answer the question. (The speaker is a truth-teller.) In FS, however, without *reductio* available for conditional inferences involving the truth predicate, the speaker's statement, while consistent and therefore not to be rejected out of hand, cannot be followed to its apparent logical conclusion. In this example FS fails a simple test for suitability as a logical system for reasoning about information from untrustworthy sources.

As these examples suggest, it is likely that in the end no axiomatic system will prove ideal for handling all reasonable uses of a truth predicate in the communication process. By carving out a large and productive fragment of instances where the tools available to us *can* be applied, and by uncovering issues like trustworthiness that shape the context in which these systems operate, we should be in a better position to assess the merits of axiomatic systems for truth.

References

Cantini, A. (1990). A theory of formal truth arithmetically equivalent to ID_1. *Journal of Symbolic Logic, 55,* 244–259.

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

Halbach, V. (1994). A system of complete and consistent truth. *Notre Dame Journal of Formal Logic, 35,* 311–327.

Halbach, V. (2011). Axiomatic theories of truth. Cambridge: Cambridge University Press.

Horwich, P. (1990). Truth. Oxford: Basil Blackwell.

Kripke, S. (1975). Outline of a theory of truth. *Journal of Philosophy, 72,* 690–716.

Reinhardt, W. N. (1986). Some remarks on extending and interpreting theories with a partial predicate for truth. *Journal of Philosophical Logic, 15,* 219–251.

Sheard, M. (2008). A transactional approach to the logic of truth. In C. Dimitracopoulos, L. Newelski, D. Normann, & J. Steel (Eds.), Logic colloquium 2005 (pp. 202–220). Cambridge: Cambridge University Press.

Smullyan, R. (1978) What is the name of this book? The riddle of dracula and other logic puzzles. Englewood Cliffs: Prentice-Hall.

Part II
Uses of Truth

Chapter 5
Putting Davidson's Semantics to Work to Solve Frege's Paradox on Concept and Object

Philippe de Rouilhan

Abstract What Frege's paradox on concept and object (FP) consists in and the manner in which Frege coped with it (the ladder strategy) are briefly reviewed (§ 5.1). An idea for solving FP inspired by Husserl's semantics is presented; it results in failure, for it leads to a version of Russell's paradox, the usual solution of which implies something like a resurgence of FP (§ 5.2). A generalized version of Frege's paradox (GFP) and an idea for solving it inspired by Davidson's semantics are presented; three theorems about recursive definability of truth are put forward and used to determine whether this idea can be successfully applied to certain putative forms of the Language of Science (§ 5.3). Proofs of these three theorems, in particular of the third, which answers a question that does not seem to have drawn logicians' attention, are then given (§ 5.4). Finally, it turns out that there is a tension between the proposed solution of GFP and the idea of Language of Science assumed so far in this paper, and a way of solving it is proposed (§ 5.5).

5.1 Frege's Paradox on Concept and Object (FP) and How Frege Put up with it

5.1.1. We all remember Frege's famous letter to Husserl dated May 24, 1891 (Frege 1976, Brief XIX/1; 1980, letter VII/1), in which the former objects to the latter's semantic analysis of concept-words and sums up the main points of his own new semantics in a chart, reproduced below.[1]

The ideas presented in this article were first presented at international colloquia held in Paris and Mexico City in 2003, then in Paris and Nancy, France in 2011. I am particularly grateful to Serge Bozon, as usual. Thank you to Max Fernandez and Arnaud Plagnol for their comments on one or another of the earlier formulations of my ideas. Thank you also to Claire O. Hill, who translated the umpteenth, not quite definitive, French version of this paper into English.

[1] The capital letters that I have bestowed upon the translation of certain terms in this paper ("Proper Name" for "*Eigenname*", "Concept-word" for "*Begriffswort*", "Sense" for "*Sinn*", "Meaning" for "*Bedeutung"*, "Truth-Value" for "*Warheitswert*", "Object" for "*Gegenstand*", "*Begriff*" for "Concept", etc.) are only there to remind readers that the terms are to be understood in the technical

Ph. de Rouilhan
IHPST (Paris 1–Panthéon-Sorbonne/ENS/CNRS), Paris, France
e-mail: rouilhan@orange.fr

T. Achourioti et al. (Eds.), *Unifying the Philosophy of Truth,* Logic, Epistemology, and the Unity of Science 36, DOI 10.1007/978-94-017-9673-6_5

Sentence	Proper Name	Concept-word
↓	↓	↓
Sense of the Sentence (Thought)	Sense of the Proper Name	Sense of the Concept-word
↓	↓	↓
Meaning of the Sentence (Truth-value)	Meaning of the Proper Name (Object)	Meaning of the Concept-word (Concept)[a]

[a]At the bottom to the right of this box of the chart, Frege added: "→ Object that falls under the Concept", thus showing where he parted ways with Husserl with regard to Concept-words. For the former, the relation of Sense to Object was mediated by a Concept, while for the second, let it be said in Fregean terms, Senses referred directly to Object, which usurped the place of the Concept

The different columns may be understood as corresponding to different categories of entity. To give an idea of the difference between the different categories, Frege resorted, from that time on, to the metaphor of the completeness vs. the incompleteness, or the saturatedness vs. the unsaturatedness, of the entities under consideration, which paradoxically led him to consider Sentences as Proper Names, and the first column as a particular instance of the second. In the letter to Husserl, however, Frege did not feel the need to take that step expressly and, in my own presentation, I shall not do so either.

To each line corresponds one of the three levels—Expression, Sense, Meaning—of Fregean semantics. Frege defends the thesis of what I shall call the *categorial parallelism* of the *three* levels: Just as a Proper Name (a complete Expression) may complement a Concept-word (an incomplete Expression) to combine with it to make a Sentence (a complete Expression), so the Sense (complete) of a Proper Name may complement the Sense (incomplete) of a Concept-word to combine with it to make a Thought (a complete Sense), and so again an Object (a complete Meaning) may complement a Concept (an incomplete Meaning) to combine with it to make a Truth-Value (*sic,* a complete Meaning).

Thus, Concepts are not Objects, for example, the Concept *horse* is not an Object. In his article "Concept and Object" (Frege 1892b), Frege sought to refute Benno Kerry's objection. Kerry used the example: "The concept *horse*[2] is easily attained". In this sentence, he argued, the words "the concept *horse*" designate an object. Therefore, the concept *horse* is an object, some objects are concepts, and concepts are objects.

and more or less deviant, depending on the case, sense that Frege gave to the original. "*Eigenname*", in Frege's sense, is what is usually called a "singular term"; "*Bedeutung"* and "*Begriff*" are to be understood in a deviant sense, of which the chart provides an initial idea.

[2] Actually Kerry put quotation-marks around "horse" instead of italicizing it, as I do in Frege's wake to the same end.

Frege's reaction to Kerry's objection is extremely surprising. On the one hand, he gives in. Yes, he acknowledges, the words "the Concept *horse*" designate an Object, and the Concept *horse* is that Object. But on the other hand, he resists and persists to the point of paradox, what I call "Frege's paradox" (FP). If Concepts are not Objects and the Concept *horse* is an Object, then the Concept *horse* is not a Concept. This paradox is the price to be paid for holding on to the controversial thesis. Frege pays the price and holds on to the thesis that Concepts are not Objects.

Long ago, I proposed an in-depth analysis of Frege's paradox (Rouilhan 1988), but here (§§ 5.1–5.2) I shall restrict myself to showing readers the shortest path leading from FP to what I shall call the "generalized Frege's paradox" (GFP).

5.1.2. The situation, from a pragmatic point of view, is the following, according to Frege. By using the words "the Concept *horse*", one does not succeed in speaking about what they would like to speak, namely about a Concept, which is an incomplete entity; they only speak about a complete entity, more precisely, about an Object. Or, to speak from now on in a more *suggestive* manner than Frege did, when one uses the Proper Name "the Concept *horse*" to speak of the Concept *horse itself* (*in itself*, *as it is in itself*), they do not succeed in speaking about it, because they are trying to speak about it as they would speak of an Object, and they are indeed speaking only of an Object. In the following chart, which partially sums up the situation, the schematic Expression "Φ(ξ)" is replaceable by a Concept-word, and the letter "ξ" but marks the empty place of that Concept-word, liable to be occupied by a Proper Name in order to obtain a Sentence. The Expression "the Concept *horse*" must be construed as a variant of the Expression "the Concept *horse* (ξ)", of schema "the Concept Φ(ξ)".

Expression	"Φ(ξ)"	"The Concept Φ(ξ)"
Sense	Sense	Sense
Meaning	Φ(ξ) (it is the Concept Φ(ξ) *itself* and it is not an Object)	The Concept Φ(ξ) (it is *not* the Concept Φ(ξ) *itself*; it is an Object)

If one uses the Sentence "the Concept *horse* is not an Object" to illustrate the thesis that Concepts are not Objects, they do not say what they wanted to say, because what they are saying is literally false. And if one just states the succession of words "*horse* is not an Object" as if it were a Sentence, that does not work either, because the Concept-word "*horse*" is an incomplete Expression that cannot complete the incomplete Expression "is not an Object" so as to form a Sentence. The alleged Sentence "*horse* is not an Object" is the result of a category mistake and does not mean anything at all. In both cases, what one wanted to say was that the Concept *horse itself* is not an Object, but that, strictly speaking, cannot be said. It can only be suggested. And the same is so for Frege's thesis. If he uses the words "Concepts are not Objects" in the sense of "for every *f*, the Concept *f* is not an Object" (where "*f*" is a variable of the category of Concept-words), he is saying something that is literally false; and if it is in the sense of "for every *f*, *f* is not an Object", he is making

a category mistake and is not saying anything at all. Of course, what he wanted to say was that the Concepts *themselves* are not Objects, but that, strictly speaking, cannot be said. It can only be suggested.

Thus, instead of seeing in the paradox of the Concept *horse* the symptom of an error to be spotted and corrected, Frege simply takes note of it and holds on obstinately to the thesis of the categorial difference between Concepts (*in themselves*) and Objects, which is at the origin of the paradox and implies its own ineffability. Towards the end of "Über Begriff und Gegenstand" (Frege 1892b), he lucidly makes the point:

> I admit that there is a quite peculiar obstacle in the way of an understanding with my reader. By a kind of necessity of language, my expressions, taken literally, sometimes miss my thought; I mention an Object, when what I intend is a Concept. I fully realize that in such cases I was relying upon a reader who would be ready to meet me half-way—who does not begrudge me a pinch of salt. (p. 196).

Frege's strategy for overcoming the obstacle is not essentially different from the one generally attributed, rightly or wrongly[3], to the Wittgenstein of the *Tractatus*, the *ladder strategy*. If one cannot *say* what they would like to *say* (for example that Concepts *themselves* are not Objects), at least they can *suggest* it (as I am used to saying) and count on the *good will* of the reader or the interlocutor (as Frege more or less said), at least they can *show* it (as Wittgenstein would say). One can do so by temporarily diverting language from its normal function as a means of expressing Sense in order to use it as a means for suggesting or showing what, strictly speaking, is an inexpressible non-Sense. When this unconventional usage of language has had its effect, when it has made it possible to see what needed to be seen, one will be able to go back to conventional usage and remain there. To say this in terms akin to those of the early Wittgenstein: what cannot be said, can be shown to those who have not seen it yet by setting up the ladder of non-Sense for them to climb. Once they have seen what needed to be seen, they will have to throw away the ladder and they will finally see the world aright.

5.2 An Idea for Solving FP Inspired by Husserl's Semantics and its Failure

5.2.1. A simple way of putting an end, at least temporarily, to the dispute with Kerry and of solving FP would have been to admit that Kerry was right and to acknowledge with him that Concepts (*themselves*) are definitely Objects. In fact, Frege's best

[3] In an article of 1991 (Conant 1991), James Conant argued that, despite appearances, the Wittgenstein of the *Tractatus* (1921) did *not* take up Frege's lesson. I shall retain only the following from Conant's long, subtle analysis: the ladder strategy in the *Tractatus* is not designed to make people see what cannot be said and can only be shown, for what someone who has climbed the ladder of non-sense is supposed to see is that *there is nothing to see*. The first colloquium mentioned above (n1), at which Conant was present, focused precisely on this article.

adversary in such a dispute would have been Husserl, Husserl of *Logische Untersuchungen* (Husserl 1900–1901)[4], so close to Frege in many respects. Like Frege, he distinguished between three levels: expression, meaning (*Bedeutung*) and object (*Gegenstand*) or objectivity (*Gegenstandlichkeit*)[5]. Like Frege, he recognized the categorial parallelism of the first two levels (those of expression and of meaning), which he explained in terms of dependence and independence. However, unlike Frege, he denied, with reasons to back it up, the existence of any parallelism between these first two levels and the third (that of objectivity, see n5). Evoking the idea that "categorematic expressions represent independent objects, and syncategorematic expressions dependent objects", Husserl objects that the expression of a *dependent moment* immediately provides a decisive counter-example (Investigation IV, § 8). For, Husserl thinks, as a common noun, this expression has an independent meaning, and that in no way keeps it from representing those dependent objects that are the said dependent moments. Admittedly, as his letter of May 24, 1891 precisely shows, Frege did not agree with Husserl about the semantics of concept-words. Husserl could, nonetheless, have made an analogous objection to the Fregean idea that complete Expressions (for example, Proper Names) Mean complete entities (that is Objects), and incomplete Expressions (for example, Concept-words), incomplete entities (in this case, Concepts *as they are in themselves*). He could have objected that an Expression of schema "the Concept $\Phi(\xi)$" immediately provides a decisive counter-example.

This, therefore, is what Frege should have acknowledged, he too, to his own advantage: that the categorial parallelism of the levels of Expression and of Sense did not extend to that of Meaning. One can definitely say that a Proper Name and a Concept-word are the constituents of a Sentence and explain that the Proper Name is precisely the sort of complement that the Concept-word needs to constitute with it the unity of the Sentence. One can definitely also say that the Sense of the Proper Name and the Sense (*in itself*) of the Concept-word are constituents of the Thought and explain that the Proper Sense is precisely the sort of complement that the Conceptual Sense (*in itself*, therefore incomplete) needs to constitute with it the unity of the Thought. But, unless possessed by the demon of analogy, one can certainly not say that the Object and the Concept (*in itself*) are constituents of the Truth-value and explain that the Object is precisely the sort of complement that the Concept (*in itself*, therefore incomplete) needs to constitute with it the unity of the Truth-value. One can do this no more than, more generally, they can say that an Object and a Function (*in itself*) are constituents of the Value of this Function (*in itself*) at this Object as argument and explain that the Object is precisely the sort of complement that the Function (*in itself*, therefore incomplete) needs to constitute with it the unity of the Value. This is, moreover, what Frege would end up understanding, as seen in his

[4] See more specifically Logical Investigation IV (in 1st ed., vol. II, 1901; 2[d] ed., vol. II.1, 1913).
[5] Within the context of his analysis, Husserl used "*Gegenstandlichkeit*" as a technical term having a certain extension greater than the ordinary term "*Gegenstand*".

1919 notes for Ludwig Darmstaedter (Frege 1919), where he would write that "[o]ne cannot say that Sweden is part of the Capital of Sweden"[6].

If he had done this, nothing, in the discussion with Kerry, would have kept him from saying that Concepts (*themselves*) *are* Objects, that moreover the same is so for entities of any category, including those whose incomplete nature would not have come into question, like for example the Sense (*itself*) of a Concept-word, and finally that one can say *everything*, that one can talk *about everything*. And in his famous letter to Husserl, he would have been able to draw up the following chart[7]:

Expression	"Φ(ξ)"	"The Concept Φ(ξ)"
Sense	Sense	Sense
Meaning	The Concept Φ(ξ) (it is Concept Φ(ξ) *itself* and it is an Object)	

The Fregean critique of the Husserlian semantics of concept-words would not for all that have lost its *raison d'être*. It would have only gained in simplicity and in credibility.

The identity of the Meanings of the two Expressions schematized by "Φ(ξ)" and "the Concept Φ(ξ)" would not have prevented neither of these Expressions from playing the role corresponding to its category in the formation of a Sentence or prevented its Sense from playing the role corresponding to its category in the formation of a Thought. Nothing would have changed in this regard with respect to Frege. There would just no longer have been a way back enabling one to find again the category of an Expression and that of its Sense—and thus the role of that Expression and its Sense in the formation of a Sentence and of a Thought—from the category of its Meaning.[8]

5.2.2. The solution to one paradox may hide another, and the solution of this other paradox may involve the return of the same.

---

[6] Frege was already aware of the difficulty when he wrote, as early as 1892: "One might also say that judgments (*Urteilen*) are distinctions of parts (*Teilen*) within Truth-values. [. . .] However, I have here used the word 'part' in a special sense. [. . .] This way of speaking can certainly be attacked [. . .]. A special term would need to be invented" (Frege 1892a, pp. 35–36, 1984, p. 165).

[7] Without neglecting to add to it at the bottom to the right of the chart: "→ Object falling under the Concept" (compare with the chart of Sect. 5.1.1).

[8] For this corrected version of Frege's semantics under consideration, I am prepared to describe the role of a Proper Name, (schematized by) "Δ", and that of a Concept-word, (schematized by) "Φ(ξ)", in the Sentence (schematized by) "Φ(Δ)" nearly as C. Wright did in 1998 (Wright 1998, cf. p. 260) (this is not a quotation): the Sense of "Φ(ξ)" so relates it to the Concept Φ(ξ) that it may be used in concatenation with "Δ" in order to *ascribe* the Concept Φ(ξ) *to* Δ; and the Sense of "Δ" so relates it to Δ that it may be used in concatenation with "Φ(ξ)" in order to *subsume* Δ *under* the Concept Φ(ξ). Indeed, the solution of FP inspired by Husserl outlined in Sect. 5.2.1 could be so presented as to be clearly, essentially equivalent to Wright's solution. Unfortunately, as we shall see in Sect. 5.2.2, the theory of Concepts upon which these solutions are based falls prey to a certain version of Russell's paradox. More will then be said about Wright.

For Frege, Concepts obeyed an extensional criterion of identity. Now, if all Concepts are Objects, then nothing any longer safeguards them from a certain version of Russell's paradox. It suffices to choose for "Φ(ξ)" the Concept-word:

$$\exists f(\xi = \text{the Concept } f(\zeta) \,\&\, \neg f(\xi))$$

(where "f" is a variable of the category of Concept-words) and, using "w(ξ)" as an abbreviation of this Concept-word, to ask the fateful question whether, yes or no,

$$\text{w(the Concept w}(\xi)).$$

If yes, then

$$\exists f(\text{the Concept w}(\xi) = \text{the Concept } f(\zeta) \,\&\, \neg f(\text{the Concept w}(\xi))),$$

whence, readily,

$$\neg\text{w(the Concept w}(\xi));$$

and if no, then

$$\forall f(\text{the Concept w}(\xi) = \text{the Concept } f(\zeta) \Rightarrow f(\text{the Concept w}(\xi))),$$

whence, readily,

$$\text{w(the Concept w}(\xi)).$$

From each answer follows the opposite answer.[9]

One is thus led to acknowledge that all Concepts *in themselves* are not Objects, that there are exceptions to the principle that all are. There are, perhaps, Concepts that *in themselves* are Objects, for example—let us admit it—, the Concept *horse*, but there are surely ones that are not, for example the Concept w(ξ). The Concept w(ξ) *itself* is not an Object. One cannot argue any longer, as Frege did, that the Concept *horse* is an Object and *thus* not a Concept, for the very same entity now is both an Object and a Concept. Nor can one argue that the Concept w(ξ) is an Object and *thus* not a Concept, for an Object *may* now be a Concept. Let me dwell on that point.

If, in spite of the version of Russell's paradox presented above, the Proper Name "the Concept w(ξ)" is to have a Meaning, as Frege required of all expressions of the

[9] This paradox was notably pointed out by T. Parsons in 1986 (Parsons 1986, pp. 454–455). Wright mentions it at the end of his article, but deals with it in a somewhat offhand manner: "This, like the recent resurgence of tuberculosis in the Western world, is a disappointment. But I do not think it is really an objection—too many of the family of paradoxes that exercised Russell survive the imposition of Frege's hierarchy to allow us to think that it gets to the root of that particular one" (Wright 1998, p. 263). Wright may be right in the second part of the last sentence, but not in the first one. The first lesson to be learnt from the paradox in question (see the next paragraph in the text) immediately gives rise to a sort of resurgence of FP itself. So the paradox in question *is* an objection.

Language of Science, this can only be an Object arbitrarily chosen to play this part, an Object *ad hoc*. As to whether the use of the Proper Name "the Concept w(ξ)" does or does not give rise to FP, I mean to the paradox according to which the Concept w(ξ) is not a Concept (*as it is in itself*), this depends on the Object chosen to play the part of the Meaning of the Proper Name "the Concept w(ξ)". If it is an Object that is not a Concept (*as it is in itself*), like the Moon—let us admit it—, for example, that is chosen to play this part, then we are entitled to claim that the Concept w(ξ) is not a Concept (*as it is in itself*), and thus FP is back. But if it is an Object like the Concept *horse*—which is nothing other than the Concept *horse itself*, as admitted at the beginning of the preceding paragraph—that is chosen for this part, then there is no reason to claim that the Concept w(ξ) is not a Concept (*as it is in itself*).

However, whether the use of the Proper Name "the Concept w(ξ)" gives rise or not to FP (in the sense specified above), the situation is still paradoxical. Since the Concept w(ξ) is an Object and the Concept w(ξ) *itself* is not one, the Concept w(ξ) is not the Concept w(ξ) *itself*, or, as Frege would have simply said, the Concept w(ξ) is not the Concept w(ξ). By using the Proper Name "the Concept w(ξ)", we do not therefore succeed in speaking of the Concept w(ξ) *itself*, we are speaking of an Object, and even of an Object that has nothing to do with the Concept w(ξ) *itself* at all. There are things that one would like to say, but cannot, etc.

The semantics of Concept-words (expressions schematized by "Φ(ξ)") and Proper Names obtained by prefixing them with the operator of nominalization "the Concept" (and thus schematized by "the Concept Φ(ξ)") is summed up below in terms of whether the Concept Φ(ξ) *itself* is or is not an Object.

1st case: the Concept Φ(ξ) *itself* is an Object (the case, for example—as we have admitted—, of the Concept *horse*)

Expression	"Φ(ξ)"	"The Concept Φ(ξ)"
Sense	Sense	Sense
Meaning	The Concept Φ(ξ) (this is the Concept Φ(ξ) *itself* and it is an Object)	

2nd case: the Concept Φ(ξ) *itself* is not an Object (the case, for example, of the Concept w(ξ)

Expression	"Φ(ξ)"	"The Concept Φ(ξ)"
Sense	Sense	Sense
Meaning	Φ(ξ) (this is the Concept Φ(ξ) *itself* and it is not an Object)	the Concept Φ(ξ) (this is not the Concept Φ(ξ) *itself*, it is an *ad hoc* Object)

5.3 Generalized Frege's Paradox (GFP); an Idea for Solving it Inspired by Davidson's Semantics; Putting this Idea to the Test

5.3.1. What was at stake for Frege in his paradox was the possibility, for an author writing for readers, or a teacher speaking to students, of explicating the content of the Expressions of the Language of Science. The teacher was supposed to explain that there were different categories of Meaning, notably that of Concept and that of Object, that these categories were pairwise disjoint and that they were not to be confused, in particular, that no Concept was an Object any more than any Object was a Concept, etc. However, it transpired that, in saying this kind of thing, the teacher was ineluctably failing to say what he or she wanted to say, that what he or she wanted to say involved a category mistake and was therefore impossible to say.

The fact that the categories of Meaning were pairwise disjoint in character was not essential to FP. Let me here leave Frege and his terminology, but for the phrase "Language of Science". Generally, for a language taken to be the Language of Science to be open to a paradox of the same sort as FP, it suffices for this language to contain different categories of reference, or denotation, are not all included in a single category. In terms of variables: it suffices for this language to contain variables of different types whose domains of variation are not all included in a single domain of variation. The same reasons, *mutatis mutandis*, that hold in Frege's case lead to the same conclusion, namely, that it is impossible to explicate the content of the expressions of such a language without making a category mistake (relative to this supposed Language of Science; the resurgence of FP in Sect. 5.2.2 is a good example). This is what I call *generalized Frege's paradox* (GFP).

More precisely, the category mistake would have the form of surreptitious introduction of a new category of variable irreducible to those available in the supposed Language of Science. Let us call it *the Mistake*. If the impossibility in question were established, there would be no other solution for solving GFP than to require of any language taken to be the Language of Science that its variables range over domains that are all included in one of them (as it happens in particular and in the simplest way when all the variables range over one and the same domain). Then it would only remain to ascertain that complying with this requirement made it effectively possible to explicate, without making the Mistake, what expressions of such a language mean. But has the impossibility in question been established? Is it true that, when the requirement in question has not been met, a teacher wishing to explicate the content of the expressions of the language under consideration to a student is doomed to make the Mistake? I used to believe this, but I have not believed it for a long time (see Rouilhan 1988, pp. 186–187, and 2002, pp. 198–199). It is possible to solve GFP without having to shoulder the requirement in question.

My solution will be grounded on the basic idea of Davidson's semantics (Davidson 1984). Explicating what the expressions of a language, $\mathcal{L}$, mean is, as Davidson puts it, (not to *translate*, but) to *interpret* them. The interpretation of component expressions of statement (closed sentence) of $\mathcal{L}$ is determined by their contribution

to the interpretation of the statements of $\mathcal{L}$ in which they occur. As for the statements of $\mathcal{L}$, according to an idea Frege himself had, which was taken up successively by Wittgenstein, Carnap and Davidson, their interpretation is determined by their *truth-conditions*. Davidson more specifically requires that those truth-conditions be stated in the form of what he calls a *recursive* ***theory*** *of truth à la Tarski* for $\mathcal{L}$. This is precisely the basic idea of Davidson's semantics, and the only one I want to exploit to solve GFP.

Whence the following idea of solution to GFP for a language, $\mathcal{L}$, taken to be the Language of Science: Either it is possible to construct a recursive theory of truth *à la* Tarski for $\mathcal{L}$ without making the Mistake, and GFP is solved; or this is impossible and GFP is an indirect proof that $\mathcal{L}$ cannot be the Language of Science, and again, at least indirectly, GFP is solved. In the latter case, the impression of paradox is liable to linger until further, direct reasons are found for not mistaking $\mathcal{L}$ for the Language of Science.

5.3.2. Now, let me speak about Tarski and truth. In his famous 1935 paper (Wb) on the concept of truth (Tarski 1935), Tarski reasoned within the framework, taken to be universal, of the extensional, simple theory of types, and examined the possibility of *explicitly defining* the concept of truth for object-languages *grounded on* this theory, that is to say, obtained from a segment (possibly the totality) of its language by adding finitely many constants of certain categories. We all remember the results obtained: (1) Tarski proposed a method for explicitly defining truth for languages of finite order through an explicit definition of the relation of satisfaction, itself obtained by a conversion of a recursive definition of this relation; (2) He demonstrated the impossibility of an explicit definition of truth for languages of infinite order; (3) He indicated that the nowadays so-called *minimal* axiomatic theory of truth for an infinite-order language, whose axioms are the so-called *T-sentences* for that language, is too weak for one to be able to prove the semantic version of the fundamental laws of logic there, and that the same is so of the extensions obtained from this theory by adding one or another of these laws as new axiom. [In his post-scriptum of 1936 (Tarski 1935), Tarski was to take into consideration languages other than those to which he had limited himself up to that point, in particular to languages grounded on some set theory or other of Zermelo and his successors.]

The path that led Tarski to an explicit definition of truth for a language, $\mathcal{L}$, when such a definition is possible, goes by way of the conversion of a recursive definition of satisfaction for $\mathcal{L}$ into an explicit definition. If one skips this step to go directly from a recursive definition of satisfaction to the explicit definition of truth in terms of satisfaction, the two definitions together constitute what I am calling here *a recursive* ***definition*** *of truth à la Tarski* for $\mathcal{L}$. In Wb, whenever Tarski constructed an *explicit* definition of truth for $\mathcal{L}$, a *recursive* definition of truth was available, but Tarski did not turn his attention to this point. If he did not do it, it is because he was seeking an explicit definition and that, if a recursive definition of truth for $\mathcal{L}$ is possible at all within the chosen framework, that of the extensional, simple theory of types, it can always be converted into an explicit definition. If $\mathcal{L}$ is of finite order, a recursive definition is quite possible and so is its conversion, why therefore would he have turned his attention to? And if $\mathcal{L}$ is of infinite order, a recursive definition

is impossible. Otherwise, by conversion, an explicit definition would be possible as well, which is impossible [see above, result (2)]. Therefore, the question did not come up.

Yet, recursive definitions have their own advantages, an advantage over minimal theories, of course, whose essential weakness they do not share, but indeed an advantage also over explicit definitions. Tarski's *negative* theorem mentioned above is known to hold for many languages. Sometimes, the corresponding *positive* theorem for recursive definition of truth holds, but sometimes not.

In the following examples (theorems A-C), ZFC is Zermelo-Fraenkel set theory with axiom of choice [and without excluding individuals (in the sense of *Urelemente*)]. I note SSTT^{α} the initial (maybe total) segment of order $\alpha \leq \omega$ of the *monadic*, extensional, simple theory of types (with axiom of infinite and axiom of choice). $\mathrm{SSTT} = \mathrm{SSTT}^{\omega}$ is the simplest version of the simple theory of types, to which, as is well known, STT, the full, extensional, simple theory of types is reducible thanks to, e.g., Kuratowski's definition of ordered pairs. Let us say that a language, $\mathcal{L}$, is an *admissible* extension of the language of ZFC (SSTT^{α}, respectively) if, and only if, $\mathcal{L}$ is obtained from the latter language by adding finitely many constants each one of which is either a singular term or a predicate or functor of such a category that its addition is possible without adding new variables. If $\mathcal{L}$ is *such* an extension and the same is so of a certain extension, $\mathcal{M}$, of $\mathcal{L}$, let us say, naturally, that $\mathcal{M}$ is an *admissible* extension of $\mathcal{L}$. ZFC and SSTT^{α} for $\alpha \geq 4$ (α must be ≥ 4 for SSTT^{α} to contain Russell arithmetic) are of interest for us insofar as, *prima facie* (but see below), the Language of Science could plausibly be given the form of an admissible extension of any one of them. We know from Tarski that, *if a language, $\mathcal{L}$, is an admissible extension of the language of* ZFC (SSTT^{α} *with* $\alpha \geq 4$, *respectively*), *then an* ***explicit*** *definition of truth for $\mathcal{L}$ is* ***impossible*** *in any admissible extension of $\mathcal{L}$.* On the other hand, the corresponding results concerning *recursive* definitions of truth are the following.

Theorem A.—*Let $\mathcal{L}$ be an admissible extension of the language of* ZFC. *A* ***recursive*** *definition of truth for $\mathcal{L}$ is* ***possible*** *in some admissible extension of $\mathcal{L}$.*

This *positive* result is well known and very easy to prove, see Sect. 5.4.1.

Theorem B.—*Let $\mathcal{L}$ be an admissible extension of the language of* SSTT^{ω}. *A* ***recursive*** *definition of truth for $\mathcal{L}$ is* ***impossible*** *in any admissible extension of $\mathcal{L}$.*

This *negative* result is also well known, and hardly less easy to prove than theorem A, see Sect. 5.4.2.

Theorem C.—*Let $\mathcal{L}$ be an admissible extension of the language of* SSTT^{n} *with n natural number* ≥ 4. *A* ***recursive*** *definition of truth for $\mathcal{L}$ is* ***possible*** *in some admissible extension of $\mathcal{L}$.*

This result is the *positive* answer to a question that does not seem to have attracted logicians' attention. It is not that easy to prove, see the proof I propose in Sect. 5.4.3.

5.3.3 In a more general way, Tarski supposed a *translation* of an object-language, $\mathcal{L}$, into a metalanguage, $\mathcal{M}$, to be given, and sought the conditions of possibility of an *explicit definition* of truth for $\mathcal{L}$ in $\mathcal{M}$ relative to this translation—retrospectively, one

can say that, in Wb, it went without saying that the translation was homographic[10]. He could just as well have taken interest in the less restrictive conditions of possibility of a *recursive definition* of truth for $\mathcal{L}$ in $\mathcal{M}$ relative to this translation (comp. above, § 5.3.2). Davidson starts, inversely, from a language, $\mathcal{L}$, whose meaning may be unknown to us, and asks for what form an *interpretation* of $\mathcal{L}$ in our used language, $\mathcal{M}$, supposed to give us this meaning should take. His answer is that such an interpretation should take the form of a *recursive* ***theory*** *of truth à la Tarski* for $\mathcal{L}$ in $\mathcal{M}$.

Actually, if such a recursive theory of truth is available, then it is possible recursively to define a (unique up to alphabetical change of bound variables) translation of $\mathcal{L}$ into $\mathcal{M}$ by following the clauses of the recursive theory of truth under consideration. Say that this translation *canonically* corresponds to that theory of truth, or that it is the *canonical* translation corresponding to that theory. The idea for a solution of GFP envisioned in the present paper can now be described in the following two ways. To solve this paradox for a language, $\mathcal{L}$, taken to be the Language of Science, it would suffice to show that it is possible, *without making the Mistake*, to construct a recursive *theory* of truth *à la* Tarski for $\mathcal{L}$ in some extension, $\mathcal{M}$, of $\mathcal{L}$, such that the corresponding canonical translation is homographic—or, equivalently, to construct a recursive *definition* of truth *à la* Tarski, corresponding to the homographic translation, for $\mathcal{L}$ in some extension, $\mathcal{M}$, of $\mathcal{L}$. If the latter construction is worked out for an *admissible* extension, $\mathcal{L}$, of ZFC (SSTT^{α}, with $\alpha \geq 4$, respectively) in an *admissible* extension, $\mathcal{M}$, of $\mathcal{L}$, then the italicized condition above, relative to the Mistake, is obviously fulfilled.

It thus follows from theorems A-C that GFP *is* solvable in this way for any admissible extension of the language of ZFC (th. A) or SSTT^{n} for $n \geq 4$ (th. C)—but *not* for any admissible extension of the language of SSTT^{ω} (th. B). I am not prepared here to enter into a discussion about the very notion of Language of Science, but I think that there are some *direct* reasons, independent of GFP, why such infinite-order languages as admissible extensions of the language of SSTT^{ω} could *not* play the part of the Language of Science (see above, § 5.3.1, last paragraph).

5.4 Proofs of Theorems About Recursive Definition of Truth Stated in the Preceding Section[11]

Let me rest content with giving proof of theorem A (B, C respectively) for a simple, exemplary, admissible extension of the language of ZFC (SSTT^{ω}, SSTT^{n} for $n = 6$ respectively), for it will become self-evident that the same method of proof could have been applied to any other explicitly given admissible extension of ZFC (SSTT^{ω}, SSTT^{n} for $n \geq 4$ respectively).

---

[10] It goes without saying that I am here borrowing this qualifier not from geometry, but from linguistics.

[11] The non-mathematically-minded reader may skip this section and go directly to Sect. 5.5.

5.4.1 Proof of Theorem A

The intended universe of ZFC is the class of what I shall call *objects*, viz. individuals (*Urelemente*) and sets. In one possible version, the signs of the language of ZFC are the variables, viz., the terms of a certain sequence (indexed by the set of non-null natural numbers) of objects, $\mathbf{v} = (\mathbf{v}_k)_{k\geq 1}$; the constants "¬", "∨", "∃", "=", "Set" (monadic predicate of sethood) and "∈" (dyadic predicate of membership); punctuation marks "(" and ")". Let $\mathcal{L}$ be the admissible extension obtained from that language by adding, for example, the constant "a" of the category of singular terms, and the constant "P" of the category of triadic predicates. The rules of formation are the usual ones. In the definitions below, "x", "y", and "z" are (primitive) variables of $\mathcal{L}$, and non-primitive symbols are contextually definable: "σ"and "τ" are sequence (of objects) variables; "$\mathbf{t}_1$", "$\mathbf{t}_2$ ", and "$\mathbf{t}_3$", term (of $\mathcal{L}$) variables; "**A**" and "**B**", (open or closed) sentence (of $\mathcal{L}$) variables; "i", "j" and "k", non-null natural number variables; "⌜" and "⌝", Quine's quasi-quotation marks; "$\Leftrightarrow_{ij}$", the operator of formal equivalence relative to variables "i" and "j"; etc.

We shall begin with an explicit definition of a predicate of relative denotation, "$\text{Den}_{\mathcal{L}}$", for $\mathcal{L}$. Then we shall recursively define "$\text{Sat}_{\mathcal{L}}$" in terms of the eliminable "$\text{Den}_{\mathcal{L}}$". Finally, we shall explicitly define "$\text{Tr}_{\mathcal{L}}$" in terms of "$\text{Sat}_{\mathcal{L}}$". Thus, "$\text{Tr}_{\mathcal{L}}$" will have been recursively defined in an admissible extension obtained from $\mathcal{L}$ by adding constants of the elementary syntax of $\mathcal{L}$ and the primitive predicate "$\text{Sat}_{\mathcal{L}}$". There is no need to worry about coding the objects of this syntax. They are simply supposed to be themselves already there, somewhere in the intended universe of ZFC.

Denotation of a Term Relative to a Sequence The terms of a language are variables and constants of the same category as some variables; those of $\mathcal{L}$ are $\mathbf{v}_k$ for $k \geq 1$ and "a".

$x\ \text{Den}_{\mathcal{L}}\ y, z \Leftrightarrow_{\text{df}} x$ and z are a term, $\mathbf{t}$, and a sequence, $\sigma = (\sigma_k)_{k\geq 1}$, respectively, such that [($\mathbf{t}$ is of the form $\mathbf{v}_k$ & $\sigma_k = y$) ∨ ($\mathbf{t}$ = "a" & a = y)].

Thus, $\mathbf{v}_k\ \text{Den}_{\mathcal{L}}\ y, \sigma \Leftrightarrow_k \sigma_k = y$ and "a" $\text{Den}_{\mathcal{L}}\ y, \sigma \Leftrightarrow \text{a} = y$.

Satisfaction of a Sentence by a Sequence A first clause insures that the dyadic relation $\text{Sat}_{\mathcal{L}}$ can only hold between a sequence of objects and a sentence of $\mathcal{L}$. Four clauses then fix the conditions of satisfaction of an atomic sentence of $\mathcal{L}$ by a sequence:

- $\sigma\ \text{Sat}_{\mathcal{L}}\ \ulcorner \mathbf{t}_1 = \mathbf{t}_2 \urcorner \Leftrightarrow \exists x \exists y(\mathbf{t}_1\ \text{Den}_{\mathcal{L}}\ x, \sigma\ \ \&\ \ \mathbf{t}_2\ \text{Den}_{\mathcal{L}}\ y, \sigma\ \ \&\ \ x = y)$;
- $\sigma\ \text{Sat}_{\mathcal{L}}\ \ulcorner \text{Set}\mathbf{t}_1 \urcorner \Leftrightarrow \exists x(\mathbf{t}_1\ \text{Den}_{\mathcal{L}}\ x, \sigma\ \ \&\ \ \text{Set}x)$;
- $\sigma\ \text{Sat}_{\mathcal{L}}\ \ulcorner \mathbf{t}_1 \in \mathbf{t}_2 \urcorner \Leftrightarrow \exists x \exists y\ (\mathbf{t}_1\ \text{Den}_{\mathcal{L}}\ x, \sigma\ \ \&\ \ \mathbf{t}_2\ \text{Den}_{\mathcal{L}}\ y, \sigma\ \ \&\ \ x \in y)$;
- $\sigma\ \text{Sat}_{\mathcal{L}}\ \ulcorner \text{P}\mathbf{t}_1\mathbf{t}_2\mathbf{t}_3 \urcorner \Leftrightarrow \exists x \exists y \exists z(\mathbf{t}_1\ \text{Den}_{\mathcal{L}}\ x, \sigma\ \&\ \mathbf{t}_2\ \text{Den}_{\mathcal{L}}\ y, \sigma\ \&\ \mathbf{t}_3\ \text{Den}_{\mathcal{L}}\ z, \sigma\ \&\ \text{P}xyz)$.

Three clauses finally fix the conditions of satisfaction of a non-atomic sentence by σ according to the satisfaction of shorter sentences by this same sequence or by others, $\tau = (\tau_k)_{k\geq 1}$, connected to it:

- $\sigma\ \text{Sat}_{\mathcal{L}}\ \ulcorner(\neg \text{A})\urcorner \Leftrightarrow \neg(\sigma\ \text{Sat}_{\mathcal{L}}\ \mathbf{A})$;
- $\sigma\ \text{Sat}_{\mathcal{L}}\ \ulcorner(\ \mathbf{A} \vee \mathbf{B})\urcorner \Leftrightarrow (\sigma\ \text{Sat}_{\mathcal{L}}\ \mathbf{A} \vee \sigma\ \text{Sat}_{\mathcal{L}}\ \mathbf{B})$;
- $\sigma\ \text{Sat}_{\mathcal{L}}\ \ulcorner \exists \mathbf{v}_i\ (\mathbf{A})\urcorner \Leftrightarrow_i \exists \tau((j \neq i \Rightarrow_j \tau_j = \sigma_j)\ \&\ \ \tau\ \text{Sat}_{\mathcal{L}}\ \mathbf{A})$.

Truth of a Statement It is easily shown that a statement (closed sentence) of $\mathcal{L}$ is satisfied by any sequence or by none, whence the definition sought for the truth predicate, "$\mathrm{Tr}_{\mathcal{L}}$", for $\mathcal{L}$ in an admissible extension of $\mathcal{L}$:

$$\mathrm{Tr}_{\mathcal{L}}\mathbf{A} \Leftrightarrow_{\mathrm{df}} \mathbf{A} \text{ is a statement of } \mathcal{L} \;\&\; \forall\sigma(\sigma\ \mathrm{Sat}_{\mathcal{L}}\ \mathbf{A}).$$

5.4.2 *Proof of Theorem B*

The intended universe of SSTT$^{\omega}$ is composed of individuals and classes corresponding to a simply infinite hierarchy of types (or orders), viz., the type (or order) 1 for individuals, 2 for classes of individuals, 3 for classes of classes of individuals, etc. In one possible version, the signs of the language of SSTT$^{\omega}$ are, but for differences to be presently explained, the same as those of the language of ZFC. Variables are typed: for any explicitly given $i \geq 1$, the variables of order i, ranging over the domain of the entities of order i, are the terms of a certain sequence, or, more precisely, K-sequence, noted $\mathbf{v}_{\mathrm{K}}{}^{(i)} = (\mathbf{v}_{\mathrm{K}}{}^{(i)}{}_k)_{k\geq 1}$, where a K-sequence is a class of K-ordered pairs of a certain sort (see below), and a K-ordered pair (of entities of the same order) is an ordered pair as coded, or defined, by Kuratowski. "Set" and "$\in$" have been eliminated, and "=" maintained as corresponding to identity between individuals.[12] Natural numbers are assumed to be defined *à la* Russell and, except for any duly marked exception, to be of the lowest possible order, viz., 3, their definition is of order 4, and so is Russell arithmetic. Now, let $\mathcal{L}^{\omega}$ be the admissible extension of the language of SSTT$^{\omega}$ obtained by adding, for example, the constants "a" of the category of singular terms, denoting the individual a, and "P" of the category of dyadic predicates whose first argument values are entities of order 3 and second argument values entities of order 5. Further notions and notations are progressively introduced when needed.

It is impossible to construct a recursive definition of truth for $\mathcal{L}^{\omega}$ in any admissible extension of $\mathcal{L}^{\omega}$, for such a definition would be convertible in this extension into an explicit definition of truth for $\mathcal{L}^{\omega}$, which since Tarski we know is impossible. However it is easy to find a *method* for recursively defining truth for any explicitly given, *finite*-order, initial segment of $\mathcal{L}^{\omega}$, in some admissible extension of $\mathcal{L}^{\omega}$.

Let us present this method by means of an example, by constructing a recursive definition of truth for the initial fragment, $\mathcal{L}^{6}$, of $\mathcal{L}^{\omega}$ obtained by eliminating all the

[12] It is well known that "$=$" is definable in the language of SSTT$^{\omega}$ in terms of other primitives, but the same is not so for the language of any explicitly given, finite, initial segment, SSTT^{n}, of SSTT$^{\omega}$. Whence my maintaining of "$=$" in the language of SSTT$^{\omega}$, for the sake of the overall simplicity of the present Sect. 5.4. I maintain "=" as primitive only for individuals, because identity for two entities of any explicitly given order > 1 is definable in terms of that primitive.

variables of order >6. (This exercise will also prove useful in the next Sect. 5.4.3.) But for a few differences, this definition resembles that of Sect. 5.4.1.

Denotation of a Term Relative to six K-sequences of Entities of Order $1, \ldots, 6$ *Respectively* The terms of $\mathcal{L}^6$ are, for each explicitly given i, ${\mathbf{v}_\mathrm{K}}^{(i)}{}_k$ for $i \geq 1$, and "a". The relation in question, can only hold between a term of $\mathcal{L}^6$ of explicitly given order i such that $1 \leq i \leq 6$, an entity of order i, and six K-sequences, ${\sigma_\mathrm{K}}^{(1)}, \ldots, {\sigma_\mathrm{K}}^{(6)}$, of entities of order 1,. . ., 6 respectively. Indeed, there are six relations at stake here, which, by abuse of language, I shall uniformly note $\mathrm{Den}_{\mathcal{L}^6}$. Noting $\langle a, b\rangle_\mathrm{K}$ the K-ordered pair whose terms are a and b (in this order),[13] for the relation $\mathrm{Den}_{\mathcal{L}^6}$ to hold between its eight arguments, they must more precisely be as follows:

1. a term, $\mathbf{t}^i$, of explicitly given order i such that $1 \leq i \leq 6$, and so of order 2 as an entity, if we hold with Tarski that an expression is a certain class of inscriptions and that inscriptions are individuals;[14]
2. an entity, y^i, of order i if the aforesaid term is of the form ${\mathbf{v}_\mathrm{K}}^{(i)}{}_k$, and of order 1 if it is "a";
3. six sequences, $\sigma^{(1)} = (\sigma^{(1)}{}_k)_{k\geq 1}, \ldots,\ \sigma^{(6)} = (\sigma^{(6)}{}_k)_{k\geq 1}$, of entities of order $1, \ldots, 6$ respectively, which are classes of K-ordered pairs of certain form, from which the orders of the K-sequences are computable. See the chart below.

Sequence	${\sigma_\mathrm{K}}^{(1)}$	${\sigma_\mathrm{K}}^{(2)}$	${\sigma_\mathrm{K}}^{(3)}$	${\sigma_\mathrm{K}}^{(4)}$	${\sigma_\mathrm{K}}^{(5)}$	${\sigma_\mathrm{K}}^{(6)}$
Terms of the sequence	${\sigma_\mathrm{K}}^{(1)}{}_k$ for $k \geq 1$	${\sigma_\mathrm{K}}^{(2)}{}_k$ for $k \geq 1$	${\sigma_\mathrm{K}}^{(3)}{}_k$ for $k \geq 1$	${\sigma_\mathrm{K}}^{(4)}{}_k$ for $k \geq 1$	${\sigma_\mathrm{K}}^{(5)}{}_k$ for $k \geq 1$	${\sigma_\mathrm{K}}^{(6)}{}_k$ for $k \geq 1$
Order of these terms	1	2	3	4	5	6
Members of the sequence	$\langle k, \{\{{\sigma_\mathrm{K}}^{(1)}{}_k\}\}\rangle_\mathrm{K}$ for $k \geq 1$	$\langle k, \{{\sigma_\mathrm{K}}^{(2)}{}_k\}\rangle_\mathrm{K}$ for $k \geq 1$	$\langle k, {\sigma_\mathrm{K}}^{(3)}{}_k\rangle_\mathrm{K}$ for $k \geq 1$	$\langle \{k\}, {\sigma_\mathrm{K}}^{(4)}{}_k\rangle_\mathrm{K}$ for $k \geq 1$	$\langle \{\{k\}\}, {\sigma_\mathrm{K}}^{(5)}{}_k\rangle_\mathrm{K}$ for $k \geq 1$	$\langle \{\{\{k\}\}\}, {\sigma_\mathrm{K}}^{(6)}{}_k\rangle_\mathrm{K}$ for $k \geq 1$
Order of these members	5	5	5	6	7	8
Order of the sequence	6	6	6	7	8	9

[13] In Sect. 5.4.3 other ways of coding, or defining, ordered pairs will be put to work.

[14] Tarski's view in Wb was roughly as follows: *inscriptions* are concrete individuals with a form and a size; *expressions* of a language are equivalence classes of certain inscriptions with respect to the relation of having the same form and size; some of these expressions are *simple*, others are *complex*; complex ones result from simple ones through a finite process of *concatenation*, internal to order 2. For there to be, as needed, infinitely many expressions in the language under consideration, there should be infinitely many inscriptions. Tarski was perfectly lucid about the formidable problems surrounding such a requirement (see Tarski 1956 or 1983, p. 174).

If $i = 1$, then the relation $\mathrm{Den}_{\mathcal{L}^6}$ holds between such arguments if, and only if, $\mathbf{t}^1$ is of the form $\mathbf{v}^{(1)}_{\mathrm{K}\,k}$ and $\sigma^{(1)}_{\mathrm{K}\,k} = y^1$, or $(\mathbf{t}^1 = \mathrm{a} \wedge \mathrm{a} = y^1)$; and if $2 \leq i \leq 6$, then it holds if, and only if, $\mathbf{t}^i$ is of the form $\mathbf{v}_{\mathrm{K}}{}^{(i)}{}_k$ and $\sigma_{\mathrm{K}}{}^{(i)}{}_k = y^i$.

It would be easy, but space-consuming, to give to these considerations, for every explicitly given i such that $1 \leq i \leq 6$, the rigorous form of an explicit definition of "$x^2\ \mathrm{Den}_{\mathcal{L}^6}\ y^i, z^6{}_1, z^6{}_2, z^6{}_3, z^7{}_4, z^8{}_5, z^9{}_6$", and then to deduce that

$$\mathbf{v}_{\mathrm{K}}{}^{(i)}{}_k\ \mathrm{Den}_{\mathcal{L}^6}\ y^i, \sigma_{\mathrm{K}}{}^{(1)}, \ldots, \sigma_{\mathrm{K}}{}^{(6)} \Leftrightarrow_k \sigma_{\mathrm{K}}{}^{(\mathrm{i})}{}_k = y^i;$$

$$\text{“}a\text{”}\ \mathrm{Den}_{\mathcal{L}^6}\ y^1, \sigma_{\mathrm{K}}{}^{(1)}, \ldots, \sigma_{\mathrm{K}}{}^{(6)} \Leftrightarrow \mathrm{a} = y^1.$$

Satisfaction of a Sentence by six K-sequences of Entities of Order 1, ..., 6 *Respectively* A first clause insures that the heptadic relation $\mathrm{Sat}_{\mathcal{L}^6}$ can only hold between $\sigma_{\mathrm{K}}{}^{(1)}, \ldots, \sigma_{\mathrm{K}}{}^{(6)}$ and a sentence. Twelve (five plus six plus one) clauses then fix the conditions of satisfaction of an atomic sentence of $\mathcal{L}^6$ by $\sigma_{\mathrm{K}}{}^{(1)}, \ldots, \sigma_{\mathrm{K}}{}^{(6)}$:

- for any explicitly given i such that $1 \leq i \leq 5$:

$$\sigma_{\mathrm{K}}{}^{(1)}, \ldots, \sigma_{\mathrm{K}}{}^{(6)}\ \mathrm{Sat}_{\mathcal{L}^6}\ \ulcorner \mathbf{t}^{i+1}\mathbf{t}^i \urcorner \Leftrightarrow \exists x^i \exists x^{i+1} (\mathbf{t}^i\ \mathrm{Den}_{\mathcal{L}^6}\ x^i, \sigma_{\mathrm{K}}{}^{(1)}, \ldots, \sigma_{\mathrm{K}}{}^{(6)}\ \ \&$$
$$\mathbf{t}^{i+1}\ \mathrm{Den}_{\mathcal{L}^6}\ x^{i+1}, \sigma_{\mathrm{K}}{}^{(1)}, \ldots, \sigma_{\mathrm{K}}{}^{(6)}\ \ \&\ \ x^{i+1} x^i);$$

- for any explicitly given i such that $1 \leq i \leq 6$:

$$\sigma_{\mathrm{K}}{}^{(1)}, \ldots, \sigma_{\mathrm{K}}{}^{(6)}\ \mathrm{Sat}_{\mathcal{L}^6}\ \ulcorner \mathbf{t}^i{}_1 = \mathbf{t}^i{}_2 \urcorner \Leftrightarrow \exists x^i \exists y^i (\mathbf{t}^i{}_1\ \mathrm{Den}_{\mathcal{L}^6}\ x^i, \sigma_{\mathrm{K}}{}^{(1)}, \ldots, \sigma_{\mathrm{K}}{}^{(6)}\ \ \&$$
$$\mathbf{t}^i{}_2\ \mathrm{Den}_{\mathcal{L}^6}\ y^i, \sigma_{\mathrm{K}}{}^{(1)}, \ldots, \sigma_{\mathrm{K}}{}^{(6)}\ \ \&\ \ x^i = y^i);$$

- $\sigma_{\mathrm{K}}{}^{(1)}, \ldots, \sigma_{\mathrm{K}}{}^{(6)}\ \mathrm{Sat}_{\mathcal{L}^6}\ \ulcorner \mathrm{P}\mathbf{t}^3\mathbf{t}^5 \urcorner \Leftrightarrow \exists x^3 \exists x^5 (\mathbf{t}^3\ \mathrm{Den}_{\mathcal{L}^6}\ x^3, \sigma_{\mathrm{K}}{}^{(1)}, \ldots, \sigma_{\mathrm{K}}{}^{(6)}\ \&$
$\mathbf{t}^5\ \mathrm{Den}_{\mathcal{L}^6}\ x^5, \sigma_{\mathrm{K}}{}^{(1)}, \ldots, \sigma_{\mathrm{K}}{}^{(6)}\ \&\ \mathrm{P}x^3x^5).$

Eight (one plus one plus six) clauses then fix the conditions of satisfaction of a non-atomic sentence by $\sigma_{\mathrm{K}}{}^{(1)}, \ldots, \sigma_{\mathrm{K}}{}^{(6)}$ according to the satisfaction of shorter sentences by these same sequences or by others connected with them; $\tau^{(1)}_{\mathrm{K}} = (\tau^{(1)}_{\mathrm{K}} k)_{k \geq 1}$, ..., $\tau^{(6)}_{\mathrm{K}} = (\tau^{(6)}_{\mathrm{K}} k)_{k \geq 1}$ are supposed to answer to the same constraints as $\sigma^{(1)}_{\mathrm{K}}$, ..., $\sigma^{(6)}_{\mathrm{K}}$ respectively.

- $\sigma_{\mathrm{K}}{}^{(1)}, \ldots, \sigma_{\mathrm{K}}{}^{(6)}\ \mathrm{Sat}_{\mathcal{L}^6}\ \ulcorner (\neg \mathbf{A}) \urcorner \Leftrightarrow \neg (\sigma_{\mathrm{K}}{}^{(1)}, \ldots, \sigma_{\mathrm{K}}{}^{(6)}\ \mathrm{Sat}_{\mathcal{L}^6}\ \mathbf{A})$;
- $\sigma_{\mathrm{K}}{}^{(1)}, \ldots, \sigma_{\mathrm{K}}{}^{(6)}\ \mathrm{Sat}_{\mathcal{L}^6}\ \ulcorner (\mathbf{A} \vee \mathbf{B}) \urcorner \Leftrightarrow$
$(\sigma_{\mathrm{K}}{}^{(1)}, \ldots, \sigma_{\mathrm{K}}{}^{(6)}\ \mathrm{Sat}_{\mathcal{L}^6}\ \mathbf{A} \vee \sigma_{\mathrm{K}}{}^{(1)}, \ldots, \sigma_{\mathrm{K}}{}^{(6)}\ \mathrm{Sat}_{\mathcal{L}^6}\ \mathbf{B})$;
- $\sigma_{\mathrm{K}}{}^{(1)}, \ldots, \sigma_{\mathrm{K}}{}^{(6)}\ \mathrm{Sat}_{\mathcal{L}^6}\ \ulcorner \exists \mathbf{v}_{\mathrm{K}}{}^{(1)}{}_k (\mathbf{A}) \urcorner \Leftrightarrow_k \exists \tau_{\mathrm{K}}{}^{(1)} ((j \neq k \Rightarrow_j \tau_{\mathrm{K}}{}^{(1)}{}_j = \sigma_{\mathrm{K}}{}^{(1)}{}_j)$
$\&\ \ \tau_{\mathrm{K}}{}^{(1)}, \sigma_{\mathrm{K}}{}^{(2)}, \sigma_{\mathrm{K}}{}^{(3)}, \sigma_{\mathrm{K}}{}^{(4)}, \sigma_{\mathrm{K}}{}^{(5)}, \sigma_{\mathrm{K}}{}^{(6)}\ \mathrm{Sat}_{\mathcal{L}^6}\ \mathbf{A})$;
- $\sigma_{\mathrm{K}}{}^{(1)}, \ldots, \sigma_{\mathrm{K}}{}^{(6)}\ \mathrm{Sat}_{\mathcal{L}^6}\ \ulcorner \exists \mathbf{v}_{\mathrm{K}}{}^{(2)}{}_k (\mathbf{A}) \urcorner \Leftrightarrow_k \exists \tau_{\mathrm{K}}{}^{(2)} ((j \neq k \Rightarrow_j \tau_{\mathrm{K}}{}^{(2)}{}_j = \sigma_{\mathrm{K}}{}^{(2)}{}_j)$
$\&\ \ \sigma_{\mathrm{K}}{}^{(1)}, \tau_{\mathrm{K}}{}^{(2)}, \sigma_{\mathrm{K}}{}^{(3)}, \sigma_{\mathrm{K}}{}^{(4)}, \sigma_{\mathrm{K}}{}^{(5)}, \sigma_{\mathrm{K}}{}^{(6)}\ \mathrm{Sat}_{\mathcal{L}^6}\ \mathbf{A})$;

.....

- $\sigma_{\mathrm{K}}{}^{(1)}, \ldots, \sigma_{\mathrm{K}}{}^{(6)}\ \mathrm{Sat}_{\mathcal{L}^6}\ \ulcorner \exists \mathbf{v}_{\mathrm{K}}{}^{(6)}{}_k (\mathbf{A}) \urcorner \Leftrightarrow_k \exists \tau_{\mathrm{K}}{}^{(6)} ((j \neq k \Rightarrow_j \tau_{\mathrm{K}}{}^{(6)}{}_j = \sigma_{\mathrm{K}}{}^{(6)}{}_j)$
$\&\ \ \sigma_{\mathrm{K}}{}^{(1)}, \sigma_{\mathrm{K}}{}^{(2)}, \sigma_{\mathrm{K}}{}^{(3)}, \sigma_{\mathrm{K}}{}^{(4)}, \sigma_{\mathrm{K}}{}^{(5)}, \tau_{\mathrm{K}}{}^{(6)}\ \mathrm{Sat}_{\mathcal{L}^6}\ \mathbf{A})$.

Truth of a Sentence Here is the definition of truth predicate in terms of satisfaction for $\mathcal{L}^6$ sought after in the admissible extension of $\mathcal{L}^6$ obtained by adding constants of the elementary syntax of $\mathcal{L}^6$ and the predicate "$\mathrm{Sat}_{\mathcal{L}^6}$":

$$\mathrm{Tr}_{\mathcal{L}^6}\mathbf{A} \Leftrightarrow_{\mathrm{df}} (\mathbf{A} \text{ is a statement of } \mathcal{L}^6 \,\&\, \forall \sigma_{\mathrm{K}}{}^{(1)} \ldots \forall \sigma_{\mathrm{K}}{}^{(6)} (\sigma_{\mathrm{K}}{}^{(1)}, \ldots, \sigma_{\mathrm{K}}{}^{(6)} Sat_{\mathcal{L}^6}\ \mathbf{A})).$$

The method illustrated is obviously applicable for any explicitly given, finite-order, initial fragment of $\mathcal{L}^\omega$ in order recursively to define a truth predicate for it in some admissible extension of $\mathcal{L}^\omega$. All these fragments form a naturally increasing sequence, whose limit, in a sense, is their union, and extensions of truth predicates for these fragments do so as well. The union of these fragments is $\mathcal{L}^\omega$, and the union of these extensions is, from some (if any) transcendent point of view, the class of true statements of $\mathcal{L}^\omega$, but this class cannot be the extension of any predicate of any admissible extension of $\mathcal{L}^\omega$. What is possible for the terms of a sequence need not be so for the limit. I repeat: A recursive definition of truth for $\mathcal{L}^\omega$ itself is impossible in any admissible extension of $\mathcal{L}^\omega$.

5.4.3 *Proof of Theorem C*

SSTT^n is the initial segment of SSTT^ω of order n for any explicitly given $n \geq 4$, and $\mathcal{L}^n$ is the corresponding initial segment of $\mathcal{L}^\omega$ as in Sect. 5.4.2. In Sect. 5.4.3.1, we prove that a recursive definition of truth is possible for $n \geq 4$ in the admissible extension obtained from $\mathcal{L}^n$ by adding primitive pairing functors, primitive constants of elementary syntax of $\mathcal{L}^n$, and primitive predicates of satisfaction for $\mathcal{L}^n$. The proof is based on a stratagem devised by Quine and Boolos (henceforth QB's trick). In Sect. 5.4.3.2, we try to prove that one can dispense with adding primitive pairing functors. It turns out that, with the best definitions of ordered pair available and QB's trick again, one obtains the result hoped for only for $n \geq 5$, not for $n = 4$. The latter case would merit further study, something remaining to be undertaken.

5.4.3.1. The K-sequences $\sigma_{\mathrm{K}}^{(1)}, \ldots, \sigma_{\mathrm{K}}^{(6)}$ (and the same holds for $\tau_{\mathrm{K}}^{(1)}, \ldots, \tau_{\mathrm{K}}^{(6)}$) involved in the recursive definition of truth for $\mathcal{L}^6$ given in Sect. 5.4.2 are of order 6, 6, 6, 7, 8, 9 respectively, so that the last three fall outside the intended universe of $\mathcal{L}^6$. Note that Kuratowski's definition of ordered pair, according to which an ordered pair is two orders higher than its terms, is to a large extent responsible for such an overflow. Let us abandon Kuratowski's definition and in place of it, for every explicitly given i such that $1 \leq i \leq 6$, use a primitive pairing functor,[15] say (by abuse of language, as if there were only one functor instead of six) "C", that can be attached

[15] Bourbaki did it, *mutatis mutandis*, in the first two editions, dated 1954 and 1960, of his fascicule containing the chapter on set theory, see Bourbaki 1954, chap. 2, § 1, n° 1 and § 2, n°1, but abandoned it in the third edition, dated 1966, and naturally in the one volume edition of book I, dated 1970. Curiously, the English translation of book I, dated 1968, goes back to the French, first edition of this fascicule instead of the second one.

to two terms, $\mathbf{t}^i{}_1$, $\mathbf{t}^i{}_2$, of order i to form a term of the same order, $\ulcorner \mathrm{C}\mathbf{t}^i{}_1\mathbf{t}^i{}_2 \urcorner$, or rather, more suggestively, $\ulcorner \langle \mathbf{t}^i{}_1, \mathbf{t}^i{}_2 \rangle_{\mathrm{C}} \urcorner$, and such that

$$\left\langle x^i, y^i{}_2 \right\rangle_{\mathrm{C}} = \left\langle u^i, v^i \right\rangle_{\mathrm{C}} \Rightarrow (x^i = u^i \ \ \& \ \ y^i = v^i).$$

The K-sequences $\sigma_{\mathrm{K}}^{(1)}, \ldots, \sigma_{\mathrm{K}}^{(6)}$ can now be replaced by C-sequences, say $\sigma_{\mathrm{C}}{}^{(1)}, \ldots, \sigma_{\mathrm{C}}{}^{(6)}$, of order 4, 4, 4, 5, 6, 7 respectively, and only the last one falls outside the intended universe of $\mathcal{L}^6$.

It is possible to rid ourselves of this last C-sequence $\sigma_{\mathrm{C}}{}^{(6)}$ by the means of QB's trick.[16] Generally speaking, and using an outdated terminology dating back to Euler, the gist of QB's trick consists in coding a *single-valued function*, *f*, whose value at every argument, *x*, is a class, *fx*, by the *multiple-valued function* whose values at *x* are the members of *fx*. More specifically, here, the C-sequence $\sigma_{\mathrm{C}}{}^{(6)}$ can be coded by the relation, R, holding exactly between any natural number $k \geq 1$ and each element of $\sigma_{\mathrm{C}}{}^{(6)}{}_k$, this relation being itself coded by the class, Σ^6, of ordered pairs of the form $\langle \{\{k\}\}, x^5 \rangle_{\mathrm{C}}$ such that $\mathrm{R}kx^5$. The sequence $\sigma_{\mathrm{C}}{}^{(6)}$, of order 7, is thus coded by the class Σ^6, of order 6, of the $\langle \{\{k\}\}, x^5 \rangle_{\mathrm{C}}$ such that $\sigma_{\mathrm{C}}{}^{(6)}{}_k x^5$. So, for any j and any x^5, $\sigma_{\mathrm{C}}{}^{(6)}{}_k x^5$ if, and only if, $\Sigma^6 \langle \{\{k\}\}, x^5 \rangle_{\mathrm{C}}$; whence the two lemmas that will be used below (in the second one, T^6 is supposed to code $\tau_{\mathrm{C}}{}^{(6)}$ as Σ^6 does $\sigma_{\mathrm{C}}{}^{(6)}$):

$$\text{Lemma 1. } \sigma_{\mathrm{C}}{}^{(6)}{}_k = x^6 \Leftrightarrow_k \left(x^6 x^5 \Leftrightarrow_{x^5} \Sigma^6 \left\langle \{\{k\}\}, x^5 \right\rangle_{\mathrm{C}}\right);$$

$$\text{Lemma 2. } \tau_{\mathrm{C}}{}^{(6)}{}_k = \sigma_{\mathrm{C}}{}^{(6)}{}_k \Leftrightarrow_k \left(\mathrm{T}^6 \left\langle \{\{k\}\}, x^5 \right\rangle_{\mathrm{C}} \Leftrightarrow_{x^5} \Sigma^6 \left\langle \{\{k\}\}, x^5 \right\rangle_{\mathrm{C}}\right).$$

But what about the K-sequence $\mathbf{v}_{\mathrm{K}}{}^{(i)}$ for any explicitly given i such that $1 \leq i \leq 6$? "$\mathbf{v}_{\mathrm{K}}{}^{(i)}$" is not a variable, but a (syntactic) constant, so that it is not the order of $\mathbf{v}_{\mathrm{K}}{}^{(i)}$, but that of its *members,* that matters. The *terms* of $\mathbf{v}_{\mathrm{K}}{}^{(i)}$, viz., variables of order *i*, are entities of order 2, thus its *members* are of order 5 (see the chart of Sect. 5.4.2), and there is no problem for them to be present in the intended universe of $\mathcal{L}^6$, nor would there be any problem for $\mathcal{L}^n$, for any explicitly given $n \geq 5$, with the K-sequences $\mathbf{v}_{\mathrm{K}}{}^{(i)}$, for any explicitly given i such that $1 \leq i \leq n$. However, there would *be* a problem for $\mathcal{L}^4$ with the K-sequences $\mathbf{v}_{\mathrm{K}}{}^{(i)}$, for any explicitly given i such that $1 \leq i \leq 4$. For the sake of uniformity, I shall replace the K-sequences $\mathbf{v}_{\mathrm{K}}{}^{(i)}$ of variables of $\mathcal{L}^6$ by the C-sequences of the same *terms* whose *members* are of order 3.

Now let me, as briefly as possible, present the recursive definition of truth sought for, which does not commit one to anything outside the intended universe of $\mathcal{L}^6$.

Relative Denotation The explicitly definable relation $\mathrm{Den}_{\mathcal{L}^6}$ can only hold between a term of $\mathcal{L}^6$, an entity of the same order as this term, five C-sequences, $\sigma_{\mathrm{C}}{}^{(1)}, \ldots, \sigma_{\mathrm{C}}{}^{(5)}$, and a class, Σ^6, of ordered pairs of the form $\langle \{\{k\}\}, x^5 \rangle_{\mathrm{C}}$; and it is such that, for any explicitly given i such that $1 \leq i \leq 5$,

[16] Quine (1952) and Boolos (1985) used it to construct a recursive definition of truth for the languages of ML and ZF2 respectively, in the extension obtained from that language by adding a predicate of satisfaction.

$$\mathbf{v}_C{}^{(i)}{}_k \ \mathrm{Den}_{\mathcal{L}^6}\ x^i, \sigma_C{}^{(1)}, \ldots, \sigma_C{}^{(5)}, \Sigma^6 \Leftrightarrow_k \sigma_C{}^{(i)}{}_k = x^i;$$

$$\mathbf{v}_C{}^{(6)}{}_k \ \mathrm{Den}_{\mathcal{L}^6}\ x^6, \sigma_C{}^{(1)}, \ldots, \sigma_C{}^{(5)}, \Sigma^6 \Leftrightarrow_k \left(x^6 x^5 \Leftrightarrow_{x^5} \Sigma^6 \langle\{\{k\}\}, x^5\rangle_C\right);^{17}$$

$$\text{“a”}\ \mathrm{Den}_{\mathcal{L}^6}\ x^1, \sigma_C{}^{(1)}, \ldots, \sigma_C{}^{(5)}, \Sigma^6 \Leftrightarrow \mathrm{a} = x^1.$$

Satisfaction The first clause stipulates that the relation $\mathrm{Sat}_{\mathcal{L}^6}$ can only take place between five C-sequences, $\sigma^{(1)}, \ldots, \sigma^{(5)}$, a class, Σ^6, of ordered pairs of the form $\langle\{\{k\}\}, u\rangle_C$, and a sentence of $\mathcal{L}^6$. Clauses for atomic sentences:

- for any explicitly given i such that $1 \leq i \leq 5$:

$$\sigma_C{}^{(1)}, \ldots, \sigma_C{}^{(5)}, \Sigma^6\ \mathrm{Sat}_{\mathcal{L}^6} \ulcorner \mathbf{t}^{i+1}\mathbf{t}^i \urcorner \Leftrightarrow \exists x^i \exists x^{i+1} (\mathbf{t}^i{}_1\ \mathrm{Den}_{\mathcal{L}^6}\ x^i, \sigma_C{}^{(1)}, \ldots, \sigma_C{}^{(5)}, \Sigma^6 \ \&\ \mathbf{t}^i{}_2\ \mathrm{Den}_{\mathcal{L}^6}\ x^{i+1}, \sigma_C{}^{(1)}, \ldots, \sigma_C{}^{(5)}, \Sigma^6 \ \&\ x^{i+1} x^i);$$

- for any explicitly given i such that $1 \leq i \leq 6$:

$$\sigma_C{}^{(1)}, \ldots, \sigma_C{}^{(5)}, \Sigma^6\ \mathrm{Sat}_{\mathcal{L}^6} \ulcorner \mathbf{t}^i{}_1 \mathbf{t}^i{}_2 \urcorner \Leftrightarrow \exists x^i \exists y^i (\mathbf{t}^i{}_1\ \mathrm{Den}_{\mathcal{L}^6}\ x^i, \sigma_C{}^{(1)}, \ldots, \sigma_C{}^{(5)}, \Sigma^6 \ \&\ \mathbf{t}^i{}_2 \mathrm{Den}_{\mathcal{L}^6}\ y^i, \sigma_C{}^{(1)}, \ldots, \sigma_C{}^{(5)}, \Sigma^6 \ \&\ x^i = y^i);$$

- $\sigma_C{}^{(1)}, \ldots, \sigma_C{}^{(5)}, \Sigma^6\ \mathrm{Sat}_{\mathcal{L}^6} \ulcorner \mathrm{P}\mathbf{t}^3\mathbf{t}^5 \urcorner \Leftrightarrow \exists x^3 \exists x^5 (\mathbf{t}^3\ \mathrm{Den}_{\mathcal{L}^6}\ x^3, \sigma_C{}^{(1)}, \ldots, \sigma_C{}^{(5)}, \Sigma^6 \ \&\ \mathbf{t}^5\ \mathrm{Den}_{\mathcal{L}^6}\ x^5, \sigma_C{}^{(1)}, \ldots, \sigma_C{}^{(5)}, \Sigma^6 \ \&\ \mathrm{P}x^3 x^5).$

Clauses for non-atomic sentences:

- $\sigma_C{}^{(1)}, \ldots, \sigma_C{}^{(5)}, \Sigma^6\ \mathrm{Sat}_{\mathcal{L}^6} \ulcorner (\neg \mathrm{A}) \urcorner \Leftrightarrow \neg (\sigma_C{}^{(1)}, \ldots, \sigma_C{}^{(5)}, \Sigma^6\ \mathrm{Sat}_{\mathcal{L}^6}\ \mathbf{A})$;
- $\sigma_C{}^{(1)}, \ldots, \sigma_C{}^{(5)}, \Sigma^6\ \mathrm{Sat}_{\mathcal{L}^6} \ulcorner (\mathbf{A} \vee \mathbf{B}) \urcorner \Leftrightarrow (\sigma_C{}^{(1)}, \ldots, \sigma_C{}^{(5)}, \Sigma^6\ \mathrm{Sat}_{\mathcal{L}^6}\ \mathbf{A}) \vee \sigma_C{}^{(1)}, \ldots, \sigma_C{}^{(5)}, \Sigma^6\ \mathrm{Sat}_{\mathcal{L}^6}\ \mathbf{B})$;
- $\sigma_C{}^{(1)}, \ldots, \sigma_C{}^{(5)}, \Sigma^6\ \mathrm{Sat}_{\mathcal{L}^6} \ulcorner \exists \mathrm{v}_C{}^{(1)}{}_k (\mathbf{A}) \urcorner \Leftrightarrow_k \exists \tau_C{}^{(1)} \left(\left(j \neq k \Rightarrow_j \tau_C{}^{(1)}{}_j = \sigma_C{}^{(1)}{}_j \right) \ \&\ \tau_C{}^{(1)}, \sigma_C{}^{(2)}, \sigma_C{}^{(3)}, \sigma_C{}^{(4)}, \sigma_C{}^{(5)}, \Sigma^6\ \mathrm{Sat}_{\mathcal{L}^6}\ \mathbf{A} \right)$;
- $\sigma_C{}^{(1)}, \ldots, \sigma_C{}^{(5)}, \Sigma^6\ \mathrm{Sat}_{\mathcal{L}^6} \ulcorner \exists \mathbf{v}_C{}^{(2)}{}_k (\mathbf{A}) \urcorner \Leftrightarrow_k \exists \tau_C{}^{(2)} \left(\left(j \neq k \Rightarrow_j \tau_C{}^{(2)}{}_j = \sigma_C{}^{(2)}{}_j \right) \ \&\ \sigma_C{}^{(1)}, \sigma_C{}^{(2)}, \sigma_C{}^{(3)}, \sigma_C{}^{(4)}, \tau_C{}^{(5)}, \Sigma^6\ \mathrm{Sat}_{\mathcal{L}^6}\ \mathbf{A} \right)$;

  .....

- $\sigma_C{}^{(1)}, \ldots, \sigma_C{}^{(5)}, \Sigma^6\ \mathrm{Sat}_{\mathcal{L}^6} \ulcorner \exists \mathbf{v}_C{}^{(5)}{}_k (\mathbf{A}) \urcorner \Leftrightarrow_k \exists \tau_C{}^{(5)} \left(\left(j \neq k \Rightarrow_j \tau_C{}^{(5)}{}_j = \sigma_C{}^{(5)}{}_j \right) \ \&\ \sigma_C{}^{(1)}, \sigma_C{}^{(2)}, \sigma_C{}^{(3)}, \sigma_C{}^{(4)}, \tau_C{}^{(5)}, \Sigma^6\ \mathrm{Sat}_{\mathcal{L}^6}\ \mathbf{A} \right)$;
- $\sigma_C{}^{(1)}, \ldots, \sigma_C{}^{(5)}, \Sigma^6\ \mathrm{Sat}_{\mathcal{L}^6} \ulcorner \exists \mathbf{v}_C{}^{(6)}{}_k (\mathbf{A}) \urcorner \Leftrightarrow_k \exists \mathrm{T}^6 \left(\left(j \neq k \Rightarrow_j \left(\mathrm{T}^6 \langle\{\{j\}\}, x^5\rangle_C \Leftrightarrow_{x^5} \Sigma^6 \langle\{\{j\}\}, x^5\rangle_C \right) \right) \ \&\ \sigma_C{}^{(1)}, \ldots, \sigma_C{}^{(5)}, \mathrm{T}^6\ \mathrm{Sat}_{\mathcal{L}^6}\ \mathbf{A} \right)$.

Truth $\mathrm{Tr}_{\mathcal{L}^6} \mathrm{A} \Leftrightarrow_{\mathrm{df}} (\mathbf{A}$ is a statement of $\mathcal{L}^6$ & $\forall \sigma_C{}^{(1)} \ldots \forall \sigma_C{}^{(5)} \forall \Sigma^6 (\sigma_C{}^{(1)}, \ldots, \Sigma_C{}^{(5)}, \Sigma^6\ \mathrm{Sat}_{\mathcal{L}^6}\ \mathbf{A}))$.[18]

[17] See lemma 1.

[18] See lemma 2.

The analysis of the construction of this definition of truth for $\mathcal{L}^6$ shows that it is easily transposable to $\mathcal{L}^n$ for any explicitly given $n \geq 4$.

5.4.3.2. One can do without primitive pairing functors, at least if $n \geq 5$.

To contain the overflow connected with the use of the notion of K-ordered pair, for which an ordered pair is two orders higher than its terms, one can, to begin with, replace this notion by another one that is more economical in terms of order. The first definition of the ordered pair such that an ordered pair is only one order higher than its terms is due to Quine and dates back to 1941.[19] Let us say that a Q_1-ordered pair, $\langle a, b\rangle_{Q_1}$, is, by definition, the class of singletons of member of a and of complements (with respect to the class of classes of the same order as a and b) of singletons of member of b. This definition only applies if a and b are classes, entities of order ≥ 2. It is adequate insofar as it implies that two Q_1-ordered pairs are identical only if their homologous terms are.

Subsequently (in his 1945), Quine had an idea for a second definition according to which an ordered pair is of the same order as its terms.[20] Let us say that a Q_2-ordered pair, $\langle a, b\rangle_{Q_2}$, with a and b of order m high enough for what follows to have a sense, is, by definition, $\#a \cup \flat b$, where $\#a$ results from a by simultaneously replacing in each of its members every natural number of order $m-2$ (if any) by its immediate successor (so that 0 of order $m-2$ does not belong to any of its members), and $\flat y$ results from $\#y$ by adding 0 of order $m-2$ to each of its members. $\#a$ and $\flat b$, and therefore also $\#a \cup \flat b$, are of order m. The notion of Q_2-ordered pair can only apply to classes, a, b, whose members can contain natural numbers, *i.e.* to classes of order ≥ 5. This definition is also adequate insofar as it implies that two Q_2-ordered pairs can be equal only if their homologous terms are.[21]

The following chart gives the easily computable order, whose knowledge is subsequently useful, of certain entities:

C-sequence	$\sigma_C^{(1)}$	$\sigma_C^{(2)}$	$\sigma_C^{(3)}$	$\sigma_C^{(4)}$	$\sigma_C^{(5)}$	$\sigma_C^{(6)}$
Order of a and b for any *member*, $\langle a, b\rangle_C$, of the C-sequence	3	3	3	4	5	6
Order of the K-sequence of the same entities as the C-sequence	6	6	6	7	8	9
Order of the Q_1-sequence of the same entities as the C-sequence	5	5	5	6	7	8
Order of the Q_2-sequence of the same entities as the C-sequence					6	7

[19] It was first related by Goodman (1941, p. 150, n5) and subsequently by Quine (1945).

[20] Quine's second definition introduces the notion of natural number of any order whatsoever ≥ 3, but for us this is an exception. Everywhere else in Sect. 5.4 of the present article, natural numbers are of order 3.

[21] The two Quinean definitions are mentioned in Scott and McCarty 2008, but not in Kanamori 2003, in spite of the latter's being historically much richer than former. Indeed, that is quite in order, given the respective theoretical aims of those papers.

It is possible recursively to define truth for $\mathcal{L}^6$ by exclusively using defined notions of ordered pair at our disposal. For example, we can use the sequences $\sigma_{Q_1}{}^{(1)}$, $\sigma_{Q_1}{}^{(2)}, \sigma_{Q_1}{}^{(3)}, \sigma_{Q_1}{}^{(4)}, \sigma_{Q_2}{}^{(5)}, \sigma_{Q_2}{}^{(6)}$, of orders 5, 5, 5, 6, 6, 7 respectively, and then code the one sequence of order > 6 by a class of order 6 thanks to QB's trick.[22] (And the same is so, *mutatis mutandis*, for $\mathcal{L}^n$ for any explicitly given $n \geq 6$.) In the case of $\mathcal{L}^5$, we can use the sequences $\sigma_{Q_1}{}^{(1)}$, $\sigma_{Q_1}{}^{(2)}$, $\sigma_{Q_1}{}^{(3)}$, $\sigma_{Q_1}{}^{(4)}$, $\sigma_{Q_2}{}^{(5)}$, of orders 5, 5, 5, 6, 6 respectively, and code the two sequences of order > 5 by classes of order 5 thanks to QB's trick.[23] On the other hand, the case of $\mathcal{L}^4$ is quite different. If we use the sequence $\sigma_{Q_1}{}^{(1)}$, $\sigma_{Q_1}{}^{(2)}$, $\sigma_{Q_1}{}^{(3)}$, $\sigma_{Q_1}{}^{(4)}$, of orders 5, 5, 5, 6 respectively, which is the best we can do with definitions of ordered pair at our disposal, then we can code these four sequences of order > 4 by classes of orders 4, 4, 4, 5 respectively, thanks to QB's trick, but then how could we get rid of the remaining class of order > 4? I do not know.

5.5 Tension Existing Between the Proposed Solution of GFP and the Idea of Language of Science, and How to Solve it

Among the admissible extensions of the language of ZFC and among those of the language of SSTT^n (for $n \geq 4$) respectable candidates may be found to take on the role of Language of Science. The idea is not a new one and it is what prompted me to take an interest in these extensions. I have accepted the basic idea of Davidson's semantics, and theorems A and C have provided me with a solution to GFP for all these extensions and therefore for the candidates in question. However, found among the admissible extensions of SSTT$^\omega$ are also candidates just as qualified, *prima facie*, to play the role of Language of Science, but theorem B has made my solution to GFP inapplicable to such languages. I have found this to be an indirect reason to deny those languages the right to play the role of Language of Science. As for direct reasons that could justify this prohibition, it would fall to a serious analysis of the very idea of Language of Science to produce them.

---

[22] Noting Σ^6 the sixth-order class coding the seventh-order sequence $\sigma_{Q_2}{}^{(6)}$ (and likewise with T^6 and $\tau Q_2{}^{(6)}$), the two lemmas to be proved and applied can be obtained from lemmas 1 and 2 by replacing "C" by "Q_2".

[23] Noting $\Sigma^5{}_1$ and $\Sigma^5{}_2$ the fifth-order classes coding sixth-order sequences $\sigma_{Q_1}{}^{(4)}$ and $\sigma_{Q_2}{}^{(5)}$ respectively (and likewise with T^5_1 and T^5_2, and $\tau_{Q_1}{}^{(4)}$ and $\tau_{Q_2}{}^{(5)}$), there are now four lemmas to be proved and applied:

$$\sigma_{Q_1}{}^{(4)}{}_k = x^4 \Leftrightarrow_k (x^4 x^3 \Leftrightarrow_{x^3} \Sigma^5{}_1\langle k, x^3\rangle_{Q_1});$$

$$\sigma_{Q_2}{}^{(5)}{}_k = x^5 \Leftrightarrow_k (x^5 x^4 \Leftrightarrow_{x^4} \Sigma^5{}_2\langle\{k\}, x^4\rangle_{Q_2});$$

$$\tau_{Q_1}{}^{(4)}{}_k = \sigma_{Q_1}{}^{(4)}{}_k \Leftrightarrow_k (T^5{}_1\langle k, x^3\rangle_{Q_1} \Leftrightarrow_{x^3} \Sigma^5{}_1\langle k, x^3\rangle Q_1);$$

$$\tau_{Q_2}{}^{(5)}{}_k = \sigma_{Q_2}{}^{(5)}{}_k \Leftrightarrow_k (T^5{}_1\langle\{k\}, x^4\rangle_{Q_2} \Leftrightarrow_{x^4} \Sigma^6{}_1\langle\{k\}, x^4\rangle_{Q_2}).$$

But another, more profound, difficulty arises once again involving the idea of Language of Science. If a recursive definition of truth *à la* Tarski for an admissible extension of the language of ZFC or of SSTT^n (for $n \geq 4$) taken to be the Language of Science, is possible in some admissible extension of this language, such a definition is nevertheless impossible *in* this language *itself*. Admittedly, this is but a manifestation of the Liar paradox and has nothing to do with one or another resurgence of GFP, but the fact of the matter is that the truth predicate for the so-called Language of Science under consideration is excluded from this language. Hence, one of two things. Either the recursive definition of truth *à la* Tarski—Tarski, the founder of semantics as science!—for the Language of Science is not a matter of Science, or it is a matter of Science, but then the language in question is not the Language of Science.

One will no more be able to solve this dilemma than one could solve GFP by considering the Language of Science, no longer in a static way, as I have done up to this point, although not without some ulterior motive, but in a dynamic way, as the moving, unforeseeable multiplicity of historically and geographically situated languages ever put to work in the enterprise of knowledge. For what is targeted as Language of Science in the dilemma and was so already in GFP, is obviously a language corresponding to a unified, stabilized, tame form of knowledge to which the enterprise of knowledge in general and as such ultimately aspires. It would now be required that the explication of the content of the expressions of such a language be possible, not only without making the Mistake relative to that language, but also without going beyond its limits. Which would not only imply a post-Tarskian solution to the Liar, but also... Also what? A corresponding post-Davidsonian semantics? There is no doubt that, technically and philosophically, greatest difficulties lay in store for the enterprise.

Awaiting better times, I am inclined to relax a bit the requirement of unity weighing upon the idea of Language of Science. In the best of cases, the unity of the so-called Language of Science would be not that of a single language, but that of an extensible, finite class of *suitable* (in a sense to be specified) extensions of a single language. The latter could be, for example, the language of ZFC or that of SSTT^n for some $n \geq 4$ [24], with *suitability* then specified as admissibility. The dilemma could be solved in that way, at least in those exemplary cases, and the solution to GFP proposed in the present article would pass the test unscathed.

[24] If this common sub-language and the rules governing the use of its signs were called "*logical*", and and the signs and rules proper to its admissible extensions, "*extra-logical*", I could be said to be proposing to renounce the unity of science dear to the Vienna Circle and even its linguistic unity, for its logical unity alone. And then one would find back the version of the logical universalism that I have defended in a recent article (Rouilhan 2012).

References

Boolos, G. (1985). Nominalist platonism. *Philosophical Review, 94,* 327–344.

Bourbaki, N. (1954). *Eléments de Mathématique*, livre I, chap. I et II. Paris: Hermann (Actualités Scientifiques et Industrielles, n°1212) (2nd ed. 1960, 3rd ed. 1966).

Bourbaki, N. (1968). *Elements of Mathematics. Theory of Sets*. Berlin: Springer.

Bourbaki, N. (1970). *Eléments de Mathématique. Théorie des ensembles*. Paris: Hermann.

Conant, J. (1991). The search for logically alien thought: Descartes, Kant, Frege, and the *Tractatus*. *Philosophical Topics, 20,* 115–180.

Davidson, D. (1984). *Inquiries into Truth and Interpretation*. Oxford: Clarendon Press.

Frege, G. (1892a). Ueber Sinn und Bedeutung. *Zeitschrift für Philosophie und philosophische Kritik, 100,* 25–50. (English trans. On Sense and Meaning in Frege 1984, pp. 157–177).

Frege, G. (1892b). Ueber Begriff und Gegenstand. *Vierteljahrsschrift für wissenschaftliche Philosophie, 19,* 192–205. (English trans. On Concept and Object in Frege 1984, pp. 182–194).

Frege, G (1919). Aufzeichnungen für Ludwig Darmstaedter. (In Frege 1969, pp. 273–277; English trans. Notes for Ludwig Darmstaedter in Frege 1979, pp. 253–257).

Frege, G. (1969). H. Hermes, et al. (Eds.), *Nachgelassene Schriften*. Hamburg: Felix Meiner. (English ed. Frege 1979).

Frege, G. (1976). G. Gabriel, et al. (Eds.), *Wissenschaftlicher Briefwechsel.* Hamburg: Felix Meiner. (English ed. Frege 1980, XIX/1; 1980, VII/1).

Frege, G. (1979). H. Hermes, et al. (Eds), *Posthumous writings*. Oxford: Blacwell. (trans: P. Long, et al.).

Frege, G. (1980). G. Gabriel, et al. (Eds.) *Philosophical and mathematical correspondence* Oxford: Blackwell. (trans: H. Kaal).

Frege, G. (1984). B. McGuinness (Ed.) *Collected papers. On mathematics, logic, and philosophy* Oxford: Blackwell. (trans: M. Black, et al.).

Goodman, N. (1941). Sequences. *The Journal of Symbolic Logic, 6,* 150–153.

Husserl, E. (1900–1901). *Logische Untersuchungen*. Halle: M. Niemeyer. (2nd ed., 1913–1921).

Kanamori, A. (2003) The empty set, the singleton, and the ordered pair. *The Bulletin of Symbolic Logic, 9,* 273–298.

Parsons, T. (1986). Why Frege should not have said "The concept *horse* is not a concept". *History of Philosophy Quarterly, 3,* 449–465.

Quine, W. V. O. (1945). On ordered pairs. *The Journal of Symbolic Logic, 10,* 95–96.

Quine, W. V. O. (1952). On an application of Tarski's theory of truth. *Proceedings of the National Academy of Science, 38,* 430–433.

Rouilhan, Ph. de (1988). *Frege. Les paradoxes de la représentation*. Paris: Editions de Minuit.

Rouilhan, Ph. de (2002). On what there are. *Proceedings of the Aristotelian Society, 102,* 183–200.

Rouilhan, Ph. de (2012). In defense of logical universalism: Taking issue with Jean van Heijenoort. *Logica Universalis, 6*, 553–586.

Scott, D., & Dominic, M. (2008). Reconsidering ordered Pairs. *The Bulletin of Symbolic Logic, 14,* 379–397.

Tarski, A. (1933). *Pojecie prawdy w jezykach nauk dedukcyjnych* (The concept of truth in the languages of deductive sciences), Warszawa. (German ed. Tarski 1935, and English ed., "The Concept of Truth in Formalized Languages", in Tarski 1956).

Tarski, A. (1935). Der Wahrheitsbegriff in den formalisierten Sprachen. *Studia Philosophica, 1,* 261–405. (1935). (English ed., The Concept of Truth in Formalized Languages, in Tarski 1956).

Tarski, A. (1956). J. H. Woodger (Ed.), *Logic, semantics, metamathematics*. *Papers from 1923 to 1938*. Oxford: Oxford University Press. (2d ed., J. Corcoran, Hackett Publishing Company, 1983).

Wittgenstein, L. (1921). Logisch-philosophische Abhandlung. *Annalen der Naturphilosophie,* 14, 185–262. (German-English ed. Wittgenstein 1961).

Wittgenstein, L. (1961). *Tractatus Logico-Philosophicus.* London: Rouledge and Kegan Paul. (bilingual ed., trad. D. F. Pears and B. F. McGuinness).
Wright, C. (1998). Why Frege did not deserve his *granum salis*. A note on the paradox of 'The Concept Horse' and the ascription of bedeutungen to predicates. *Grazer philosophische Studien, 55,* 239–363.

Chapter 6
Sets, Truth, and Recursion

Reinhard Kahle

Abstract We discuss some philosophical aspects of an intensional set theory based on an axiomatic truth theory. This set theory gains its justification from natural truth axioms combined with standard recursion-theoretic operations.

6.1 Introduction

In this paper, we present a set theory based on an axiomatic truth theory. This paper is a companion to the more technical paper (Kahle 2011); here we elaborate on the philosophical foundation of the proposed framework. Based on an idea of Scott (1975), and following approaches of Aczel (1980), Beeson (1985), and Cantini (1996), we define a notion of set using a partial truth predicate in applicative theories. Our aim is to compare the philosophical justification of the resulting sets with the justification of those in ordinary axiomatic set theory, like ZFC. Although our set theory is mathematically weak, we will discuss how the framework can be tailored to set theories with differing mathematical strength.

Axioms and Axiomatic Set Theory
The traditional view of an axiom is that it is an "evident truth" that does not require proof, or even cannot be proven. With Hilbert the mathematical notion of axiom was shifted towards the assumption of an arbitrary mathematical statement which may serve as a starting point for derivations (Kahle 2015). While this rather liberal use of the notion of axiom is perfectly normal in abstract algebra, philosophers still tend

The work was partially supported by the ESF research project *Dialogical Foundations of Semantics* within the ESF Eurocores program *LogICCC*, LogICCC/0001/2007 and by the projects *Hilbert's Legacy in the Philosophy of Mathematics*, PTDC/FIL-FCI/109991/2009, and *The notion of mathematical proof*, PTDC/MHC-FIL/5363/2012, as well as PEst-OE/MAT/UI097/2013, all funded by the Portuguese Science Foundation, FCT.

R. Kahle
CMA and DM, FCT, Universidade Nova de Lisboa, 2829-516 Caparica, Portugal
e-mail: kahle@mat.uc.pt

T. Achourioti et al. (Eds.), *Unifying the Philosophy of Truth,* Logic, Epistemology, and the Unity of Science 36, DOI 10.1007/978-94-017-9673-6_6

to try to *justify* axioms, in particular, when we turn to foundational theories, like set theory.

Zermelo was forced to defend his axiomatization of set theory which, to some extent, remains controversal today. However, Zermelo–Fraenkel set theory *including the Axiom of Choice* became the universally accepted foundational base of Mathematics. Although mathematical logicians may still discuss the status of these axioms[1], working mathematicians are content with the current state of affairs. Yet, philosophically, ZFC is questionable as an axiomatic base for mathematics, *if* axioms ought to come with an intrinsic justification. Here, it is not even necessary to draw on the notorious Axiom of Choice: already the Replacement Scheme cannot deny some contingency.

In the following, we present an alternative set theory whose axiomatic base is, we believe, philosophically more satisfying. However, it has, admittedly, two drawbacks: it is an intensional set theory and it will turn out to be mathematically rather weak.

6.2 An Applicative Truth Theory and its Sets

The Applicative Theory TON

Applicative Theories are the first-order part of Feferman's systems of *explicit mathematics* (Feferman 1975, 1979). They are usually formulated over the *logic of partial terms* (Beeson 1985), including an existence predicate which allows one to speak directly about termination of computations. Due to a subtle problem with partiality in the context of truth predicates, cf. (Kahle 1999), we confine ourselves here to *total* versions of applicative theories, that is, applicative theories where all computations are considered as terminating.

Total applicative theories are based on the theory TON (total theory of operations and numbers), introduced and studied by Jäger and Strahm (1995).[2]

Roughly, TON is a *type-free* system which formalizes combinatory logic, a pairing operation, and an axiomatically given "datatype" N for natural numbers. The theory gains proof-theoretic strength when we add induction. A formal presentation of TON is given in the Appendix; as an illustration, here are the axioms for the combinatory algebra:

I. **Combinatory algebra.**
(1) $\mathsf{k}\, x\, y = x$,
(2) $\mathsf{s}\, x\, y\, z = x\, z\, (y\, z)$.

[1] Cf. the discussion in (Feferman et al. 2000) or the discussion of the set theoretic axioms in (Maddy 1997).

[2] The partial version is usually called BON—basic theory of operations and numbers, see (Feferman and Jäger 1993; Jäger et al. 1999b) or the textbooks (Beeson 1985; Troelstra and Dalen 1988). More about applicative theories can be found in (Kahle 2007).

It is well-known that, in a type-free context, combinatory algebra allows to introduce the standard notion of λ-abstraction and a recursion operator rec (which, in fact, makes use of the possible self-application on the term level):

Proposition 1

1. *For every variable x and every term t, there exists a term $\lambda x.t$ whose free variables are those of t, excluding x, such that* TON *proves* $(\lambda x.t)\,x = t$.
2. *There exists a term* rec *such that* TON *proves* $\forall x.\mathsf{rec}\,x = x\,(\mathsf{rec}\,x)$.

This *combinatorial completeness* is important because it forms the recursive base of our truth theory. With respect to the axiomatized natural numbers, it allows to introduce, in TON, terms which represent arbitrary partial recursive functions, in the sense that, for every partial recursive function F (on the natural numbers) there is a term t_F such that if $F(n_1,\ldots,n_m)$ is defined, TON proves that $t_F\,\overline{n_1}\,\ldots\overline{n_m} = \overline{F(n_1,\ldots,n_m)}$ ($\overline{n}$ being the canonical representation of the natural number n in TON).

The Truth Theory FON

The truth theory introduced in this section is, essentially, an applicative version of the *Kripke–Feferman theory of truth*, cf. (Halbach 2009, § 4.3). The specific feature of the applicative setting is the possibility to represent formulas "without Gödelization".

FON, *Frege structures*[3] *over* TON, is defined as an extension of the applicative theory TON by a (partial) truth predicate. For it, we extend the language of TON by the new relation symbol T (truth) and new individual constants $\doteq$, $\dot{\mathsf{N}}$, $\dot{\neg}$, $\dot{\wedge}$, and $\dot{\forall}$.

The axioms of FON are the axioms of TON extended to the new language plus the following axioms:[4]

I. **Closure under prime formulae of TON.**
   (1) $x = y \leftrightarrow \mathsf{T}(x \doteq y)$,
   (2) $\neg x = y \leftrightarrow \mathsf{T}(\dot{\neg}\,(x \doteq y))$,
   (3) $\mathsf{N}(x) \leftrightarrow \mathsf{T}(\dot{\mathsf{N}}\,x)$,
   (4) $\neg\mathsf{N}(x) \leftrightarrow \mathsf{T}(\dot{\neg}\,(\dot{\mathsf{N}}\,x))$.

II. **Closure under composed formulae.**
   (5) $\mathsf{T}(x) \leftrightarrow \mathsf{T}(\dot{\neg}\,(\dot{\neg}\,x))$,
   (6) $\mathsf{T}(x) \wedge \mathsf{T}(y) \leftrightarrow \mathsf{T}(x \mathbin{\dot{\wedge}} y)$,
   (7) $\mathsf{T}(\dot{\neg}\,x) \vee \mathsf{T}(\dot{\neg}\,y) \leftrightarrow \mathsf{T}(\dot{\neg}\,(x \mathbin{\dot{\wedge}} y))$,
   (8) $(\forall x.\mathsf{T}(f\,x)) \leftrightarrow \mathsf{T}(\dot{\forall}\,f)$,
   (9) $(\exists x.\mathsf{T}(\dot{\neg}\,(f\,x))) \leftrightarrow \mathsf{T}(\dot{\neg}\,(\dot{\forall}\,f))$.

III. **Consistency.**
   (10) $\neg(\mathsf{T}(x) \wedge \mathsf{T}(\dot{\neg}\,x))$.

---

[3] The designation *Frege structures* origins from Aczel's aim to use a related theory to recast Frege's (inconsistent) system from the *Grundgesetze der Arithmetik* (Frege 1893, 1903) by use of a partial truth predicate (Aczel 1980). Following Flagg and Myhill (1987a, b) we use here, under the same name, a more liberal account than Aczel, dispensing with a primitive notion of proposition.

[4] For readability we employ infix notations like $x \doteq y$ instead of the formal applicative term $\doteq x\,y$.

As said, these are essentially the Kripke–Feferman axioms for truth, with an apparent lack of *self-reference*; however, the Kripke–Feferman version of self-reference for T can be introduced by simply defining $\dot{\mathsf{T}}$ as the identity function $\lambda x.x$:

Self-reference

- $\mathsf{T}(x) \leftrightarrow \mathsf{T}(\dot{\mathsf{T}}\, x)$,
- $\mathsf{T}(\dot{\neg}\, x) \leftrightarrow \mathsf{T}(\dot{\neg}\, (\dot{\mathsf{T}}\, x))$.

Now, using the following representation of formulas by terms, we get from our theory the so-called *T-sentences* for T-positive formulae.

Definition 2 *By induction of the build up of a formula of* FON*, we define:*

$$\dot{\overbrace{t = s}} \equiv t \doteq s,$$

$$\dot{\overbrace{\mathsf{N}(t)}} \equiv \dot{\mathsf{N}}\, t,$$

$$\dot{\overbrace{\mathsf{T}(t)}} \equiv \dot{\mathsf{T}}\, t \quad \equiv \quad t,$$

$$\dot{\overbrace{\neg\varphi}} \equiv \dot{\neg}\, \dot{\varphi},$$

$$\dot{\overbrace{\varphi \wedge \psi}} \equiv \dot{\varphi} \mathbin{\dot{\wedge}} \dot{\psi},$$

$$\dot{\overbrace{\forall x.\varphi}} \equiv \dot{\forall}\, (\lambda x.\dot{\varphi}).$$

By classical logic, this notation can be extended to: $(t \mathbin{\dot{\vee}} s) := \dot{\neg}\, (\dot{\neg}\, t \mathbin{\dot{\wedge}} \dot{\neg}\, s)$, $(t \mathbin{\dot{\rightarrow}} s) := \dot{\neg}\, t \mathbin{\dot{\vee}} s$, and $(\dot{\exists} x.t) := \dot{\neg}\, (\dot{\forall} x.\dot{\neg}\, t)$.

Proposition 3 [Cantini 1996, Theorem 8.8] *If φ is a T-positive formula, then we have:*

$$\mathsf{FON} \vdash T(\dot{\varphi}) \leftrightarrow \varphi.$$

We dispense here with a further philosophical discussion of (the pros e contras of) the Kripke–Feferman account to truth and refer to literature, for instance (Halbach 2009) or (Halbach and Horsten 2015). We will, however, invoke the *naturalness* of these axioms as a minimal base. In fact, for T*-free* formulas, the T-sentences as well as the consistency axiom seem to be a base to which everybody, formalizing truth, should be committed to. While we get the T-sentences for T-positive formulas quite naturally, it is straightforward (by a Liar argument) that the T-sentences for T-negative statements would lead to a contradiction with consistency.

Remark 4 There are interesting mathematical studies about theories where consistency is replaced by completeness, i.e., the axiom $\forall x.\mathsf{T}(x) \vee \mathsf{T}(\dot{\neg}\, x)$, cf. (Friedman and Sheard 1982; Cantini 1996; Leigh and Rathjen 2010). These studies are very valuable from a technical perspective; for instance, with respect to the replacement of a least fixed point for T by a greatest fixed point. However, on philosophical grounds,

we can hardly maintain a "complete T-predicate" as a *truth* predicate, as sentences—first of all, the Liar sentence—will turn out to be true and false at the same time. If this is not yet considered as contrary to our notion of truth—as paraconsistency tries to argue—it simply corrupts our notion of *negation*.

Sets via Truth

Based on an idea of Dana Scott (1975) we may now introduce *abstraction terms* together with an *element relation* on the basis of the truth predicate.

Definition 5 *Given two terms t and s and a formula φ, we define:*

$$\{x \mid \varphi\} := \lambda x.\dot{\varphi},$$
$$t \in s :\Leftrightarrow \mathsf{T}(s\,t).$$

The term $\{x \mid \varphi\}$ is defined for arbitrary formulae φ, and one may say that our truth theory allows unrestricted (full) comprehension. However, the element-of relation is only partial, i.e., we get—syntactically—the equivalence of $t \in \{x \mid \varphi\}$ with $\varphi[t/x]$ for T-positive formulae only.

Corollary 6 *If φ is a T-positive formula, then we have:*

$$\mathsf{FON} \vdash t \in \{x \mid \varphi\} \leftrightarrow \varphi[t/x].$$

Since this equivalence extends, of course, to arbitrary formulas provably equivalent to a T-positive one, we do not introduce *sets*[5] via a syntactic criterion, but via the characteristic property, the totality of the element relation, which may be proven case-by-case:

Definition 7 (Sets in FON)

$$\mathsf{Set}(t) :\Leftrightarrow \forall x.\mathsf{T}(t\,x) \vee \mathsf{T}(\dot{\neg}(t\,x)).$$

If for a term t, $\mathsf{FON} \vdash \mathsf{Set}(t)$, we say that t is a set (in FON).

One way to look at these sets is to say they are *T-characteristic functions* in FON.

As such they are *intensional*, i.e., in the case that we have for two applicative terms t and s, $\mathsf{Set}(t) \wedge \mathsf{Set}(s) \wedge \forall x.\mathsf{T}(t\,x) \leftrightarrow \mathsf{T}(s\,x)$ we may not have $t = s$; In (Cantini 1996, Sect. II.11), Cantini discusses the non-extensionality in detail, showing how extensionality may result in contradictions. Although extensionality is a quite natural property for sets, the intensional version is natural from the perspective of the applicative base.

The main point is that the sets are *defined* in terms of the applicative ground structure. Therefore, the set formation principles are induced by this ground structure (rather than given by explicit axioms as in usual set theories).

[5] These "sets" are called *propositional functions* by Aczel (1980, Def. 3.4) and *classes* by Cantini (1996, p. 55).

As a matter of fact, the applicative ground structure provides us with a handy collection of such set formation principles, cf. (Cantini 1996, Chap. II). It includes, in particular, the possibility to form the *universal set*, whose (T-)characteristic function is, in fact, trivial. With it, it should not be surprising that we do not have a standard power set operation, (Kahle 2011). The situation is similar to Feferman's system of *explicit mathematics*, and we can, indeed, formally embed the theory EM_0 plus Join in our theory of Frege structures (in the presence of the appropriate induction principle).

Due to the lack of a power set operation, our theory will stay mathematically comparatively weak. It is, however, shown for explicit mathematics (and, more generally, by research in *reverse mathematics*) that a significant part of "everyday mathematics" can be formalized in this framework.

6.3 Discussion

Our theory draws its *philosophical* justification from the following two ingredients:

- Natural truth axioms,
- Standard recursion-theoretic operations.

In contrast to ZFC, our axioms are not arbitrary. With respect to Hilbert's famous dictum concerning the axiomatic method (Hilbert 1918) one can, indeed, speak of a *lowering of the foundations*. As already mentioned, the mathematical weakness of our theory makes it unlikely that this approach can compete as foundational base with ZFC. And in the end, the decision for or against a formal framework is taken on mathematical grounds not on philosophical ones.[6]

But our setup permits variations that yield set theories with specific mathematical strength. We will shortly address these possibilites which, in part, are still subject to further research.

Strengthenings

1. *Iterating truth*. The first, quite natural idea would be to *iterate* truth, such that negative truth statements can be captured "positively" on a higher level.
   This was investigated by Cantini (1996) using an external levelling and by Kahle (2003) with an internal levelling. The resulting theories have *metapredicative* strength and can be compared with the transfinitely iterated fixed-point theories $\widehat{\mathsf{ID}}_\alpha$, (Jäger et al. 1999a).
2. *Stronger truth theories*. To reach a theory of the strength of the impredicative theory ID_1 we may replace the Kripke–Feferman truth axioms by axioms for *Supervaluation*. The idea of supervaluation is due to van Fraassen (1968, 1970). It expresses that formulae which follow by pure logic are true independently of the logical complexity and syntactical structure of their subformulae. For example, we have $\mathsf{T}(\overbrace{\varphi \to \varphi}^{\cdot})$ for arbitrary formulae φ (see also Halbach 2009, § 4.4).

[6] See also our note (Kahle 2009).

Cantini studied a truth theory of supervaluation over Peano Arithmetic in (Cantini 1990), the corresponding theory over applicative theories is given in (Cantini 1996). Kahle in (Kahle 2001) refined Cantini's analysis by giving a syntactical embedding of ID_1 in the applicative theory. Iterated supervaluation should lead to even stronger theories. This was carried out, so far, only for a theory over Peano Arithmetic by Fujimoto (2011).
Also, Cantini's account to *stratified truth* (Cantini 2015) can be considered as an alternative to get stronger truth theories.

3. *Other datatypes*. With respect to the applicative base, one has the possibility to replace the natural numbers by other datatypes. Fujimoto (2012), gives an approach where a truth theory is defined over Zermelo–Fraenkel set theory. Technically, it should be possible to define also an applicative theory which formalizes set theory at its base and a truth theory over it. Philosophically, however, it stays unclear whether it then would make much sense to put "another" (intensional!) concept of set, now defined in terms of truth, on top of it.
4. *Strengthening the recursion-theoretic base*. Finally, we may strengthen our theory by adding additional recursion-theoretic operators. This is the way explicit mathematics is going, for instance, by adding a *non-constructive μ operator* (Feferman and Jäger 1993, 1996; Glaß and Strahm 1996) or even stronger recursion-theoretic operators (Jäger and Strahm 2002).

Weakenings

We may ask whether we can weaken our theory, for instance, to a theory characterizing functions computable in polynomial time.

5. *Weaker truth*. Eberhard and Strahm proposed an interesting truth theory which stays "polynomial" (Eberhard and Strahm 2015). It should give rise to a corresponding "polynomial set theory" following the same steps as done in FON.
6. *Weaker recursion*. With respect to the recursion-theoretic base, there exists a very interesting approach by Schlüter (1995) which allows to restrict the recursive power of the combinatorial algebra, for instance, to primitive-recursion. The exploration of such modified recursion for the resulting set theory is still a desideratum.

In conclusion, the two parameters we have to tailor our set theories, the truth axioms and the recursion-theoretic base, allow for an investigation of an interesting range of mathematical theories. In particular, the truth predicate serves as an *indepedent* parameter if we compare this framework with a plain recursion-theoretic set theory where sets are identified with their (recursive) characteristic functions.

Appendix

The Applicative Theory TON

TON is formulated in $\mathcal{L}_t$, the first order language of operations and numbers, comprising of individual variables $x, y, z, v, w, \ldots$, individual constants k, s (combinators), p, p_0, p_1 (pairing and projection), 0, s_N, p_N (zero, successor and predecessor), d_N (definition by cases), a binary function symbol $\cdot$ for term application, and the relation symbols $=$ and N. Terms $(r, s, t, \ldots)$ are built up from individual variables and individual constants by term application. Formulas $(\varphi, \psi, \ldots)$ are constructed from by $\neg$, $\wedge$ and $\forall$ in the usual manner, starting from the atomic formulas $t = s$ and $\mathsf{N}(t)$.

We write $s\,t$ for $(s \cdot t)$ with the convention of association to the left. The connectives $\vee$, $\rightarrow$, $\leftrightarrow$, and $\exists$ are defined as usual from the other connectives.

The logic of TON is classical first-order predicate logic with equality, formulated in a Hilbert-style calculus. The non-logical axioms of TON include:

I. **Combinatory algebra.**
(1) $\mathsf{k}\,x\,y = x$,
(2) $\mathsf{s}\,x\,y\,z = x\,z\,(y\,z)$.

II. **Pairing and projection.**
(3) $\mathsf{p}_0\,(\mathsf{p}\,x\,y) = x \wedge \mathsf{p}_1\,(\mathsf{p}\,x\,y) = y$.

III. **Natural numbers.**
(4) $\mathsf{N}(0) \wedge \forall x.\mathsf{N}(x) \rightarrow \mathsf{N}(\mathsf{s}_\mathsf{N}\,x)$,
(5) $\forall x.\mathsf{N}(x) \rightarrow \mathsf{s}_\mathsf{N}\,x \neq 0 \wedge \mathsf{p}_\mathsf{N}\,(\mathsf{s}_\mathsf{N}\,x) = x$,
(6) $\forall x.\mathsf{N}(x) \wedge x \neq 0 \rightarrow \mathsf{N}(\mathsf{p}_\mathsf{N}\,x) \wedge \mathsf{s}_\mathsf{N}\,(\mathsf{p}_\mathsf{N}\,x) = x$.

IV. **Definition by cases on N.**
(7) $\mathsf{N}(v) \wedge \mathsf{N}(w) \wedge v = w \rightarrow \mathsf{d}_\mathsf{N}\,x\,y\,v\,w = x$,
(8) $\mathsf{N}(v) \wedge \mathsf{N}(w) \wedge v \neq w \rightarrow \mathsf{d}_\mathsf{N}\,x\,y\,v\,w = y$.

We may add the following natural induction scheme to TON:

Formula induction on N ($\mathcal{L}_t$-I_N)

$$\varphi(0) \wedge (\forall x.\mathsf{N}(x) \wedge \varphi(x) \rightarrow \varphi(\mathsf{s}_\mathsf{N}\,x)) \rightarrow \forall x.\mathsf{N}(x) \rightarrow \varphi(x).$$

TON + ($\mathcal{L}_t$-I_N) is proof-theoretically equivalent to Peano arithmetic PA.

References

Aczel, P. (1980) Frege structures and the notion of proposition, truth and set. In J. Barwise, H. Keisler, & K. Kunen (eds.), *The Kleene symposium* (pp. 31–59). Amsterdam: North-Holland.

Beeson, M. (1985). *Foundations of constructive mathematics. Ergebnisse der Mathematik und ihrer Grenzgebiete; 3.Folge* (Bd. 6). Berlin: Springer.

Cantini, A. (1990). A theory of formal truth arithmetically equivalent to ID_1. *Journal of Symbolic Logic, 55*(1), 244–259.

Cantini, A. (1996). *Logical frameworks for truth and abstraction, vol. 135 of Studies in Logic and the Foundations of Mathematics*. Amsterdam: North-Holland.

Cantini, A. (2015). On stratified truth. In D. Achourioti, K. Fujimoto, H. Galinon, & J. Martinez (Eds.), *Unifying the philosophy of truth, volume 36. Logic, epistemology and the unity of science*. Dordrecht: Springer.

Eberhard, S., & Strahm, T. (2015). Unfolding feasible arithmetic and weak truth. In D. Achourioti, K. Fujimoto, H. Galinon, & J. Martinez (Eds.), *Unifying the philosophy of truth, volume 36. Logic, epistemology and the unity of science*. Dordrecht: Springer.

Feferman, S. (1975). A language and axioms for explicit mathematics. In J. Crossley (ed.) *Algebra and logic, vol. 450 of Lecture Notes in Mathematics* (pp. 87–139). Berlin: Springer.

Feferman,S. (1979). Constructive theories of functions and classes. In M. Boffa, D. van Dalen, & K. McAloon (eds.) *Logic colloquium 78* (pp. 159–224). Amsterdam: North-Holland.

Feferman, S. & Jäger, G. (1993). Systems of explicit mathematics with non-constructive μ-operator. Part I. *Annals of Pure and Applied Logic, 65*(3), 243–263.

Feferman, S. & Jäger, G. (1996). Systems of explicit mathematics with non-constructive μ-operator. Part II. *Annals of Pure and Applied Logic, 79,* 37–52.

Feferman, S., Friedman, H. M., Maddy, P., & Steel, J. R. (2000). Does mathematics need new axioms? *Bulletin of Symbolic Logic, 6*(4), 401–446.

Flagg, R. & Myhill, J. (1987a). An extension of frege structures. In D. Kueker, E. Lopez-Escobar, & C. Smith (eds.), *Mathematical logic and theoretical computer science* (pp. 197–217). New York: Dekker.

Flagg, R. & Myhill, J. (1987b). Implication and analysis in classical frege structures. *Annals of Pure and Applied Logic, 34,* 33–85.

Frege, G. (1893). *Grundgesetze der Arithmetik; begriffschriftlich abgeleitet*, vol. 1. Jena: Hermann Pohle.

Frege, G. (1903). *Grundgesetze der Arithmetik, begriffschriftlich abgeleitet*, vol. 2. Jena: Hermann Pohle. Reprinted together with vol. 1, Hildesheim: Olms. 1966.

Friedman, H., & Sheard, M. (1982). An axiomatic approach to self-referential truth. *Annals of Pure and Applied Logic, 33,* 1–21.

Fujimoto, K. (2011). Autonomous progression and transfinite iteration of self-applicable truth. *Journal of Symbolic Logic, 76*(3), 914–945.

Fujimoto, K. (2012). Classes and truths in set theory. *Annals of Pure and Applied Logic, 163,* 1484–1523.

Glaß, T. & Strahm, T. (1996). Systems of explicit mathematics with non-constructive μ-operator and join. *Annals of Pure and Applied Logic, 82*(2), 193–219.

Halbach, V. (2009). Axiomatic theories of truth. In E. N. Zalta (ed.), *The stanford encyclopedia of philosophy*. Winter 2009 edition. http://plato.stanford.edu/entries/truth-axiomatic/.

Halbach, V. & Horsten, L. (2015). Norms for theories of reflexive truth. In D. Achourioti, K. Fujimoto, H. Galinon, & J. Martinez (Eds.), *Unifying the philosophy of truth, volume 36. Logic, epistemology and the unity of science*. Dordrecht: Springer.

Hilbert, D. (1918). Axiomatisches Denken. *Mathematische Annalen, 78*(3/4), 405–415.

Jäger, G., Kahle, R., Setzer, A., & Strahm, T. (1999a). The proof-theoretic analysis of transfinitely iterated fixed point theories. *Journal of Symbolic Logic,* 64(1), 53–67.

Jäger, G., Kahle, R., & Strahm, T. (1999b). On applicative theories. In A. Cantini, E. Casari, & P. Minari (eds.) *Logic and foundation of mathematics*. (pp. 88–92). Dordrecht: Kluwer.

Jäger, G. & Strahm, T. (1995). Totality in applicative theories. *Annals of Pure and Applied Logic, 74*, 105–120.

Jäger, G. & Strahm, T. (2002). The proof-theoretic strength of the Suslin operator in applicative theories. In W. Sieg, R. Sommer, & C. Talcott (eds.) *Reflections on the foundations of mathematics: Essays in honor of Solomon Feferman. Vol. 15 of Lecture Notes in Logic* (pp. 270–292). ASL and AK Peters.

Kahle, R. (1999). Frege structures for partial applicative theories. *Journal of Logic and Computation, 9*(5), 683–700.

Kahle, R. (2001). Truth in applicative theories. *Studia Logica, 68*(1), 103–128.

Kahle, R. (2003). Universes over Frege structures. *Annals of Pure and Applied Logic, 119*(1–3), 191–223.

Kahle, R. (2007). *The applicative realm volume 40, of Textos de Matemática.*. Coimbra: Departamento de Matemática Coimbra

Kahle, R. (2009). The universal set—a (never fought) battle between philosophy and mathematics. In O. Pombo & Á. Nepomuceno (eds.), *Lógica e Filosofia da Ciência, vol. 2 of Colecção Documenta* (pp. 53–65). Lisboa: Centro de Filosofia das Ciências da Universidade de Lisboa.

Kahle, R. (2011). The universal set and diagonalization in Frege structures. *Review of Symbolic Logic, 4*(2), 205–218.

Kahle, R. (2015). Axioms as hypotheses. In T. Piecha & P. Schroeder-Heister (Eds.), Proceedings of the Conference on Hypothetical Reasoning, 23–24 August 2014 (pp. 47–54). Tübingen: University of Tübingen.

Leigh, G. & Rathjen, M. (2010). An ordinal analysis for theories of self-referential truth. *Archive for Mathematical Logic, 49*, 213–247.

Maddy, P. (1997). *Naturalism in mathematics*. Oxford: Clarendon.

Schlüter, A. (1995). A theory of rules for enumerated classes of functions. *Archive for Mathematical Logic, 34*, 47–63.

Scott, D. (1975). Combinators and classes. In C. Böhm (ed.), λ*-Calculus and computer science theory. Vol. 37 of Lecture Notes in Computer Science*. (pp. 1–26). Berlin: Springer.

Troelstra, A. & van Dalen, D. (1988). *Constructivism in mathematics, vol. II*. Amsterdam: North Holland.

van Fraassen, B. (1968). Presupposition, implication and self-reference. *Journal of Philosophy, 65*, 135–152.

van Fraassen, B. (1970). Inference and self-reference. *Synthese, 21*, 425–438.

Chapter 7
Unfolding Feasible Arithmetic and Weak Truth

Sebastian Eberhard and Thomas Strahm

Abstract In this paper we continue Feferman's unfolding program initiated in (Feferman, vol. 6 of Lecture Notes in Logic, 1996) which uses the concept of the unfolding $\mathcal{U}(\mathsf{S})$ of a schematic system S in order to describe those operations, predicates and principles concerning them, which are implicit in the acceptance of S. The program has been carried through for a schematic system of non-finitist arithmetic NFA in Feferman and Strahm (Ann Pure Appl Log, 104(1–3):75–96, 2000) and for a system FA (with and without Bar rule) in Feferman and Strahm (Rev Symb Log, 3(4):665–689, 2010). The present contribution elucidates the concept of unfolding for a basic schematic system FEA of feasible arithmetic. Apart from the operational unfolding $\mathcal{U}_0(\mathsf{FEA})$ of FEA, we study two full unfolding notions, namely the predicate unfolding $\mathcal{U}(\mathsf{FEA})$ and a more general truth unfolding $\mathcal{U}_{\mathsf{T}}(\mathsf{FEA})$ of FEA, the latter making use of a truth predicate added to the language of the operational unfolding. The main results obtained are that the provably convergent functions on binary words for all three unfolding systems are precisely those being computable in polynomial time. The upper bound computations make essential use of a specific theory of truth T_{PT} over combinatory logic, which has recently been introduced in Eberhard and Strahm (Bull Symb Log, 18(3):474–475, 2012) and Eberhard (A feasible theory of truth over combinatory logic, 2014) and whose involved proof-theoretic analysis is due to Eberhard (A feasible theory of truth over combinatory logic, 2014). The results of this paper were first announced in (Eberhard and Strahm, Bull Symb Log 18(3):474–475, 2012).

7.1 Introduction

The notion of *unfolding a schematic formal system* was introduced in Feferman (1996) in order to answer the following question:

S. Eberhard · T. Strahm
Institut für Informatik und angewandte Mathematik, Universität Bern,
Neubrückstrasse 10, CH-3012 Bern, Switzerland.
e-mail: eberhard@iam.unibe.ch

T. Strahm
e-mail: strahm@iam.unibe.ch

T. Achourioti et al. (Eds.), *Unifying the Philosophy of Truth,* Logic, Epistemology,
and the Unity of Science 36, DOI 10.1007/978-94-017-9673-6_7

Given a schematic system S*, which operations and predicates, and which principles concerning them, ought to be accepted if one has accepted* S*?*

A paradigmatic example of a schematic system S is the basic system NFA of non-finitist arithmetic. In Feferman and Strahm (2000), three unfolding systems for NFA of increasing strength have been analyzed and characterized in more familiar proof-theoretic terms; in particular, it was shown that the full unfolding of NFA, $\mathcal{U}(\mathsf{NFA})$, is proof-theoretically equivalent to predicative analysis. For more information on the path to the unfolding program, especially with regard to predicativity and the implicitness program, see Feferman (2005).

More recently, the unfolding notions for a basic schematic system of finitist arithmetic, FA, and for an extension of that by a form BR of the so-called bar rule have been worked out in Feferman and Strahm (2010). It is shown that $\mathcal{U}(\mathsf{FA})$ and $\mathcal{U}(\mathsf{FA}+\mathsf{BR})$ are proof-theoretically equivalent, respectively, to primitive recursive arithmetic, PRA, and to Peano arithmetic, PA.

The aim of the present contribution is to elucidate the concept of unfolding in the context of a natural schematic system FEA for *feasible arithmetic*. We will sketch various unfoldings of FEA and indicate their relationship to weak systems of explicit mathematics and partial truth.

The basic schematic system FEA of feasible arithmetic is based on a language for binary words generated from the empty word by the two binary successors S_0 and S_1; in addition, it includes some natural basic operations on the binary words like, for example, word concatenation and multiplication. The logical operations of FEA are conjunction ($\wedge$), disjunction ($\vee$), and the bounded existential quantifier ($\exists^{\leq}$). FEA is formulated as a system of sequents in this language: apart from the defining axioms for basic operations on words, its heart is a schematically formulated, i.e. open-ended induction rule along the binary words, using a free predicate letter P.

The operational unfolding $\mathcal{U}_0(\mathsf{FEA})$ of FEA extends FEA by a general background theory of combinatory algebra and tells us which operations on words are implicit in the acceptance of FEA. It further includes the generalized substitution rule from Feferman and Strahm (2010), which allows arbitrary formulas to be substituted for free predicates in derivable rules of inference such as, for example, the induction rule. We will see that $\mathcal{U}_0(\mathsf{FEA})$ derives the totality of precisely the polynomial time computable functions.

The full predicate unfolding $\mathcal{U}(\mathsf{FEA})$ of FEA tells us, in addition, which predicates and operations on them ought to be accepted if one accepts FEA. It presupposes each logical operation of FEA as an operation on predicates. Predicates themselves are just represented as special operations equipped with an elementhood relation on them. We may further accept the formation of the disjoint union of a (bounded with respect to $\leq$) sequence of predicates given by a corresponding operation. It will turn out that the provably convergent functions of $\mathcal{U}(\mathsf{FEA})$ are still the polynomial time computable ones.

We will also describe an alternative way to define the full unfolding of FEA which makes use of a truth predicate T which mimics the logical operations of FEA in a natural way and makes explicit the requirement that implicit in the acceptance of FEA

is the ability to reason about truth in FEA. Using a truth predicate in order to expand a given theory is straightforward and standard approach in the so-called implicitness program: one prominent example is Feferman (1991) where the reflective closure of a schematic system is introduced via the famous Feferman-Kripke axioms of partial truth. More recently, in Feferman's original definition of unfolding in (Feferman 1996), a truth predicate is used in order to describe the full unfolding of a schematic system.

The truth unfolding $\mathcal{U}_{\mathsf{T}}(\mathsf{FEA})$ is obtained by extending the combinatory algebra by a unary truth predicate. Indeed, $\mathcal{U}_{\mathsf{T}}(\mathsf{FEA})$ contains the predicate unfolding $\mathcal{U}(\mathsf{FEA})$ in a natural way, including the disjoint union operator for predicates. Moreover, the truth unfolding is proof-theoretically equivalent to the predicate unfolding in the sense that its provably convergent functions on the binary words are precisely the polytime functions.

The upper bound computations for both $\mathcal{U}(\mathsf{FEA})$ and $\mathcal{U}_{\mathsf{T}}(\mathsf{FEA})$ will be obtained via the weak truth theory T_{PT} introduced in Eberhard and Strahm (2012b) and Eberhard (2014), whose analysis and polynomial time upper bound is achieved in Eberhard (2014). The embedding of our two unfolding systems into T_{PT} is rather straightforward, but some special care and additional considerations are needed in order to treat their generalized substitution rules.

We end this introduction by giving a short outline of the paper. In Sect. 2 we describe in detail the basic schematic formulation of feasible arithmetic FEA. In Sect. 3 we extend FEA to its operational unfolding $\mathcal{U}_0(\mathsf{FEA})$, thus introducing its underlying abstract theory of operations in the sense of a combinatory algebra and the generalized substitution rule. We will show that the polynomial time computable functions are very naturally and directly proved to be total in $\mathcal{U}_0(\mathsf{FEA})$, hence establishing the lower bound for $\mathcal{U}_0(\mathsf{FEA})$. In Sect. 4 we turn to the full predicate unfolding $\mathcal{U}(\mathsf{FEA})$ of FEA and Sect. 5 describes the truth unfolding $\mathcal{U}_{\mathsf{T}}(\mathsf{FEA})$. The final section of the paper is devoted to the upper bound of $\mathcal{U}(\mathsf{FEA})$ and $\mathcal{U}_{\mathsf{T}}(\mathsf{FEA})$ via the above-mentioned truth theory T_{PT}.

7.2 The Basic Schematic System FEA

In this section we introduce the basic schematic system FEA of feasible arithmetic. Its intended universe of discourse is the set $\mathbb{W} = \{0,1\}^*$ of finite binary words and its basic operations and relations include the binary successors S_0 and S_1, the predecessor Pd, the initial subword relation $\subseteq$, word concatenation $\circledast$ as well as word multiplication $\boxtimes$.[1] The logical operations of FEA are conjunction ($\wedge$), disjunction ($\vee$), and bounded existential quantification ($\exists^{\leq}$). As in the case of finitist arithmetic FA, the statements proved in FEA are sequents of formulas in the given language, i.e. implication is allowed at the outermost level.

[1] Given two words w_1 and w_2, the word $w_1 \boxtimes w_2$ denotes the length of w_2 fold concatenation of w_1 with itself.

7.2.1 The Language of FEA

The language $\mathcal{L}$ of FEA contains a countably infinite supply of variables $\alpha, \beta, \gamma, \ldots$ (possibly with subscripts). These variables are interpreted as ranging over the set of binary words $\mathbb{W}$. $\mathcal{L}$ includes a constant ϵ for the empty word, three unary function symbols $\mathsf{S}_0, \mathsf{S}_1, \mathsf{Pd}$ and three binary function symbols $\circledast$, $\boxtimes$, $\subseteq$.[2] Terms of $\mathcal{L}$ are defined as usual and are denoted by $\sigma, \tau, \ldots$. Further, $\mathcal{L}$ contains the binary predicate symbol $=$ for equality, and an infinite supply $P_0, P_1, \ldots$ of free predicate letters.

The atomic formulas of $\mathcal{L}$ are of the form $(\sigma = \tau)$ and $P_i(\sigma_1, \ldots, \sigma_n)$ for $i \in \mathbb{N}$. The formulas are closed under $\wedge$ and $\vee$ as well as under bounded existential quantification. In particular, if A is an $\mathcal{L}$ formula, then $(\exists\alpha \preceq \tau)A$ is an $\mathcal{L}$ formula as well, where τ is not allowed to contain α. Further, as usual for theories of words, we use $\sigma \preceq \tau$ as an abbreviation for $1\boxtimes\sigma \subseteq 1\boxtimes\tau$, thus expressing that the length of σ is less than or equal to the length of τ. We use $\bar{\alpha}, \bar{\sigma}$, and $\bar{A}$ to denote finite sequences of variables, terms, and formulas, respectively. Moreover, the notation $\sigma[\bar{\alpha}]$ and $A[\bar{\alpha}]$ is used to indicate a sequence of free variables possibly occurring in a term σ or a formula A; finally, $\sigma[\bar{\tau}]$ and $A[\bar{\tau}]$ are used to denote the result of substitution of $\bar{\alpha}$ by $\bar{\tau}$ is those expressions.

7.2.2 Axioms and Rules of FEA

FEA is formulated as a system of sequents Σ of the form $\Gamma \to A$, where Γ is a finite sequence of $\mathcal{L}$ formulas and A is an $\mathcal{L}$ formula. Hence, we have the usual Gentzen-type logical axioms and rules of inference for our underlying restricted language. In particular, the bounded existential quantifier is governed by the following rules of inference, where the usual variable conditions apply:

$$\frac{\Gamma \to \sigma \preceq \tau \wedge A[\sigma]}{\Gamma \to (\exists\beta \preceq \tau)A[\beta]} \qquad (\mathsf{E1})$$

$$\frac{\Gamma, \alpha \preceq \tau, A[\alpha] \to B}{\Gamma, (\exists\beta \preceq \tau)A[\beta] \to B} \qquad (\mathsf{E2})$$

Further, in our restricted logical setting, we adopt the following rule of term substitution:

$$\frac{\Gamma[\alpha] \to A[\alpha]}{\Gamma[\tau] \to A[\tau]} \qquad (\mathsf{S0})$$

The non-logical axioms of FEA state the usual defining equations for the function symbols of the language $\mathcal{L}$, see, e.g., Ferreria (1987) for similar axioms. Finally, we

[2] We assume that $\subseteq$ defines the characteristic function of the initial subword relation. Further, we employ infix notation for these binary function symbols.

have the schematic induction rule formulated for a free predicate P as follows:

$$\frac{\Gamma \to P(\epsilon) \qquad \Gamma, P(\alpha) \to P(\mathsf{S}_i(\alpha)) \quad (i = 0, 1)}{\Gamma \to P(\alpha)} \tag{Ind}$$

In the various unfolding systems of FEA introduced below, we will be able to substitute an arbitrary formula for an arbitrary free predicate letter P.

7.3 The Operational Unfolding $\mathcal{U}_0(\mathbf{FEA})$

In this section we are going to introduce the *operational unfolding* $\mathcal{U}_0(\mathsf{FEA})$ of FEA. It tells us which operations from and to individuals, and which principles concerning them, ought to be accepted if one has accepted FEA.

In the operational unfolding, we make these commitments explicit by extending FEA by a partial combinatory algebra. Since it represents any recursion principle and thus any recursive function by suitable terms, it is expressive enough to reflect any ontological commitment we want to reason about. Using the notion of *provable totality*, we single out those functions and recursion principles we are actually committed to by accepting FEA.

Let us explain some properties of the operations we use in the above mentioned extension of FEA. We employ a general notion of (partial) operation, belonging to a universe V including the universe of discourse of FEA. Operations are not bound to any specific mathematical domain, but have to be considered as pre-mathematical in nature. Operations can apply to other operations. Some operations are universal and are naturally self-applicable as a result, like the identity operation or the pairing operation, while some are partial and presented to us on the binary words only. Operations satisfy the laws of a partial combinatory algebra with pairing, projections, and definition by cases.

7.3.1 The Language $\mathcal{L}_1$

The language $\mathcal{L}_1$ is an expansion of the language $\mathcal{L}$ including new constants k, s, π, p_0, p_1, d, tt, ff, e, ϵ, s_0, s_1, pd, $\mathsf{c}_\subseteq$, $*$, $\times$, and an additional countably infinite set of variables $x_0, x_1, \ldots$.[3] The new variables are supposed to range over the universe of operations and are usually denoted by $a, b, c, x, y, z, \ldots$. The $\mathcal{L}_1$ terms $(r, s, t, \ldots)$ are inductively generated from variables and constants of $\mathcal{L}$ and $\mathcal{L}_1$ by means of the function symbols of FEA and the application operator $\cdot$. We use the usual abbreviations for applicative terms and abbreviate $s \cdot t$ as (st), st or $s(t)$ as long as no confusion arises. We further adopt the convention of association to the

[3] These variables are syntactically different from the $\mathcal{L}$ variables $\alpha_0, \alpha_1, \ldots$.

left so that $s_0 s_1 \cdots s_n$ stands for $(\cdots(s_0 s_1)\cdots s_n)$; we sometimes write $s(t_0, \ldots, t_n)$ for $s t_0 \cdots t_n$. We have $(s = t)$, $s\!\downarrow$ and $P_i(\bar{s})$ for $i \in \mathbb{N}$ as atoms of $\mathcal{L}_1$. The formula $s\!\downarrow$ is interpreted as definedness of s. The formulas $(A, B, C, \ldots)$ are built from the atoms as before using $\vee, \wedge$ and the bounded existential quantifier, where as above the bounding term is a term of $\mathcal{L}$ not containing the bound variable.

For s a term of $\mathcal{L}_1 \setminus \mathcal{L}$ we write $s \preceq \tau$ for $(\exists \beta \preceq \tau)(s = \beta)$. We use the pairing operator π to introduce n-tupling $\langle t_1, \ldots, t_n \rangle$ of terms as usual.

7.3.2 Axioms and Rules of $\mathcal{U}_0(\mathbf{FEA})$

The operational unfolding $\mathcal{U}_0(\mathsf{FEA})$ is formulated as a system of sequents $\Gamma \to A$ of formulas in the language $\mathcal{L}_1$. $\emptyset \to A$ will just be displayed as A. Apart from the axioms for FEA, $\mathcal{U}_0(\mathsf{FEA})$ comprises the following axioms and rules of inference.

I. Applicative counterpart of the initial functions.
   (1) $\mathsf{s}_i\alpha = \mathsf{S}_i(\alpha)$, $\quad \mathsf{pd}\alpha = \mathsf{Pd}(\alpha)$,
   (2) $*\alpha\beta = \alpha \circledast \beta$, $\quad \times\alpha\beta = \alpha \boxtimes \beta$, $\quad \mathsf{c}_{\subseteq}\alpha\beta = \alpha \subseteq \beta$.

II. Partial combinatory algebra, pairing, definition by cases.
   (3) $\mathsf{k}ab = a$,
   (4) $\mathsf{s}ab\!\downarrow$, $\quad \mathsf{s}abc \simeq ac(bc)$,
   (5) $\mathsf{p}_0\langle a, b\rangle = a$, $\quad \mathsf{p}_1\langle a, b\rangle = b$,
   (6) $\mathsf{d}ab\,\mathsf{tt} = a$, $\quad \mathsf{d}ab\,\mathsf{ff} = b$.

III. Equality on the binary words.
   (7) $\mathsf{e}\alpha\beta = \mathsf{tt} \vee \mathsf{e}\alpha\beta = \mathsf{ff}$,
   (8) $\mathsf{e}\alpha\beta = \mathsf{tt} \leftrightarrow \alpha = \beta$.[4]

The operational unfolding of FEA includes the rules of inference of FEA (extended to the new language). In addition, in analogy to the rule (S0), we have the following new substitution rule for terms of $\mathcal{L}_1$:

$$\frac{\Gamma[u] \to A[u]}{\Gamma[t], t\!\downarrow \to A[t]} \qquad \text{(S1)}$$

The next useful substitution rule (S2) can be derived easily from the other axioms and rules. It tells us that bounded terms can be substituted for word variables:[5]

$$\frac{\Gamma[\alpha] \to A[\alpha] \quad \Gamma[t] \to t \preceq \tau}{\Gamma[t] \to A[t]} \qquad \text{(S2)}$$

Finally, $\mathcal{U}_0(\mathsf{FEA})$ includes the generalized substitution rule for derived rules of inference as it is developed in Feferman and Strahm (Feferman and Strahm 2010).

[4] To be precise, this equivalence is a shorthand for the two sequents $\mathsf{e}\alpha\beta = \mathsf{tt} \to \alpha = \beta$ and $\alpha = \beta \to \mathsf{e}\alpha\beta = \mathsf{tt}$.

[5] Note that for an $A[\alpha]$ with α occurring in a bound and a term $t \in \mathcal{L}_1 \setminus \mathcal{L}$, the rule (S2) cannot be derived because then $A[t]$ is not a formula.

Towards a more compact notation, let use write $\Sigma_1, \Sigma_2, \ldots, \Sigma_n \Rightarrow \Sigma$ to denote a rule of inference with premises $\Sigma_1, \ldots, \Sigma_n$ and conclusion Σ. We let $A[\bar{B}/\bar{P}]$ denote the formula $A[\bar{P}]$ with each subformula $P_i(\bar{t})$ replaced by $\bar{t}\downarrow \wedge B_i[\bar{t}]$, where the length of $\bar{t}$ equals the arity of P_i. The generalized substitution rule (S3) can now be described as follows: Assume that the rule of inference $\Sigma_1, \Sigma_2, \ldots, \Sigma_n \Rightarrow \Sigma$ is derivable from the axioms and rules at hand. Then we can adjoin an arbitrary substitution instance

$$\Sigma_1[\bar{B}/\bar{P}], \ldots, \Sigma_n[\bar{B}/\bar{P}] \Rightarrow \Sigma[\bar{B}/\bar{P}] \tag{S3}$$

as new rule of inference to our system. Here $\bar{P}$ and $\bar{B}$ are finite sequences of free predicates and $\mathcal{L}_1$ formulas, respectively. Note that the notion of *derivability of a rule of inference* is dynamic as one unfolds a given system. Clearly, using the generalized substitution rule, the induction rule in its usual form can be derived for an arbitrary $A \in \mathcal{L}_1$:

$$\frac{\Gamma \to A[\epsilon] \qquad \Gamma, A[\alpha] \to A[\mathsf{S}_i(\alpha)] \quad (i = 0, 1)}{\Gamma \to A[\alpha]}$$

Moreover, the usual substitution rule for sequents, $\Sigma[\bar{P}] \Rightarrow \Sigma[\bar{B}/\bar{P}]$ can be obtained as an admissible rule of inference. This ends the description of the operational unfolding $\mathcal{U}_0(\mathsf{FEA})$ of FEA.

Next we want to show that the polynomial time computable functions can be proved to be total in $\mathcal{U}_0(\mathsf{FEA})$. We call a function $F : \mathbb{W}^n \to \mathbb{W}$ provably total in a given axiomatic system whose language includes $\mathcal{L}_1$, if there exists a closed $\mathcal{L}_1$ term t_F such that (i) t_F defines F pointwise, i.e. on each standard word, and, moreover, (ii) there is a $\mathcal{L}$ term $\tau[\alpha_1, \ldots, \alpha_n]$ such that the assertion

$$t_F(\alpha_1, \ldots, \alpha_n) \leq \tau[\alpha_1, \ldots, \alpha_n]$$

is provable in the underlying system. Thus, in a nutshell, F is provably total iff it is provably and uniformly bounded.

Lemma 1 *The polynomial time computable functions are provably total in the operational unfolding $\mathcal{U}_0(\mathsf{FEA})$.*

Proof We use Cobham's characterization of the polynomial time computable functions (cf. (Cobham 1965),(Clote 1999): starting off from the initial functions of $\mathcal{L}$ and arbitrary projections, the polynomial time computable functions can be generated by closing under composition and bounded recursion. First of all, the initial functions of $\mathcal{L}$ and projections represented using lambda abstraction are obviously total. Closure of the provably total functions under composition is established by making use of the substitution rules (S1) and (S2) as well as the fact that the $\mathcal{L}$ functions are provably monotone. In order to show closure under bounded recursion, assume that F is defined by bounded recursion with initial function G and step function H,

where τ is the corresponding bounding polynomial.[6] By the induction hypothesis, G and H are provably total via suitable $\mathcal{L}_1$ terms t_G and t_H. Using the recursion or fixed point theorem of the partial combinatory algebra, we find an $\mathcal{L}_1$ term t_F which provably in $\mathcal{U}_0(\mathsf{FEA})$ satisfies the following recursion equations for $i = 0, 1$:

$$t_F(\bar{\alpha}, \epsilon) \simeq t_G(\bar{\alpha}) \mid \tau[\bar{\alpha}, \epsilon],$$
$$t_F(\bar{\alpha}, \mathsf{s}_i(\beta)) \simeq t_H(t_F(\bar{\alpha}, \beta), \bar{\alpha}, \beta) \mid \tau[\bar{\alpha}, \mathsf{s}_i(\beta)]$$

Here $\mid$ is the usual truncation operation such that $\alpha \mid \beta$ is α if $\alpha \preceq \beta$ and β otherwise. Now fix $\bar{\alpha}$ and let $A[\beta]$ be the formula $t_F(\bar{\alpha}, \beta) \preceq \tau[\bar{\alpha}, \beta]$[7] and simply show $A[\beta]$ by induction on β. Thus F is provably total in $\mathcal{U}_0(\mathsf{FEA})$ which concludes the proof of the lower bound lemma. □

7.4 The Full Predicate Unfolding $\mathcal{U}(\mathsf{FEA})$

In this section we will define the full predicate unfolding $\mathcal{U}(\mathsf{FEA})$ of FEA. It tells us, in addition, which predicates and operations on predicates ought to be accepted if one has accepted FEA. By accepting $\mathcal{U}_0(\mathsf{FEA})$ one implicitly accepts an equality predicate and operations on predicates corresponding to the logical operations of $\mathcal{U}_0(\mathsf{FEA})$. Finally, we may accept the principle of forming the predicate for the disjoint union of a (bounded) sequence of predicates given by an operation.

As before the equality predicate and the above-mentioned operations will be given as elements of an underlying combinatory algebra which is extended by a binary relation $\in$ for elementship, so predicates are represented via classifications in the sense of Feferman's explicit mathematics (Feferman 1975, 1979). We additionally use a relation Π to single out the operations representing predicates one is committed to by accepting FEA.

The language $\mathcal{L}_2$ of $\mathcal{U}(\mathsf{FEA})$ is an extension of $\mathcal{L}_1$ by new individual constants id (identity), inv (inverse image), con (conjunction), dis (disjunction), leq (bounded existential quantifier), and j (bounded disjoint unions); further new constants are $\pi_0, \pi_1, \ldots$ which are combinatorial representations of free predicates. Finally, $\mathcal{L}_2$ has a new unary relation symbol Π in order to single out the predicates we are committed to as well as a binary relation symbol $\in$ for elementhood of individuals in predicates. The terms of $\mathcal{L}_2$ are generated as before but now taking into account the new constants. The formulas of $\mathcal{L}_2$ extend the formulas of $\mathcal{L}_1$ by allowing new atomic formulas of the form $\Pi(t)$ and $s \in t$.

The axioms of $\mathcal{U}(\mathsf{FEA})$ extend those of $\mathcal{U}_0(\mathsf{FEA})$ by the following axioms about predicates.

---

[6] We can assume that only functions built from concatenation and multiplication are permissible bounds for the recursion.

[7] Recall that by expanding the definition of the $\preceq$ relation, the formula $A[\beta]$ stands for the assertion $(\exists \gamma \preceq \tau[\bar{\alpha}, \beta])(t_F(\bar{\alpha}, \beta) = \gamma)$.

I. Identity predicate
(1) $\Pi(\mathsf{id})$,
(2) $x \in \mathsf{id} \to \mathsf{p}_0 x = \mathsf{p}_1 x \wedge x = \langle \mathsf{p}_0 x, \mathsf{p}_1 x \rangle$,
(3) $\mathsf{p}_0 x = \mathsf{p}_1 x,\ x = \langle \mathsf{p}_0 x, \mathsf{p}_1 x \rangle \to x \in \mathsf{id}$.

II. Inverse image predicates
(4) $\Pi(a) \to \Pi(\mathsf{inv}(f, a))$,
(5) $\Pi(a), x \in \mathsf{inv}(f, a) \to fx \in a$,
(6) $\Pi(a), fx \in a \to x \in \mathsf{inv}(f, a)$.

III. Conjunction and disjunction
(7) $\Pi(a), \Pi(b) \to \Pi(\mathsf{con}(a, b))$,
(8) $\Pi(a), \Pi(b), x \in \mathsf{con}(a, b) \to x \in a \wedge x \in b$,
(9) $\Pi(a), \Pi(b), x \in a, x \in b \to x \in \mathsf{con}(a, b)$,
(10) $\Pi(a), \Pi(b) \to \Pi(\mathsf{dis}(a, b))$,
(11) $\Pi(a), \Pi(b), x \in \mathsf{dis}(a, b) \to x \in a \vee x \in b$,
(12) $\Pi(a), \Pi(b), x \in a \vee x \in b \to x \in \mathsf{dis}(a, b)$.

IV. Bounded existential quantification
(13) $\Pi(a) \to \Pi(\mathsf{leq} a)$,
(14) $\Pi(a), \langle y, \alpha \rangle \in \mathsf{leq}(a) \to (\exists \beta \leq \alpha)(\langle y, \beta \rangle \in a)$,
(15) $\Pi(a), (\exists \beta \leq \alpha)(\langle y, \beta \rangle \in a) \to \langle y, \alpha \rangle \in \mathsf{leq}(a)$.

V. Free predicates
(16) $\Pi(\pi_i)$,
(17) $\langle \bar{x} \rangle \in \pi_i \to P_i(\bar{x}),\ P_i(\bar{x}) \to \langle \bar{x} \rangle \in \pi_i$.
Further, the full unfolding $\mathcal{U}(\mathsf{FEA})$ includes axioms stating that a bounded sequence of predicates determines the predicate of the disjoint union of this sequence. We use the following three rules to axiomatize the join predicates in our restricted logical setting.

VI. Join rules [8]

(18) $$\frac{\Gamma, \beta \leq \alpha \to \Pi(f\beta)}{\Gamma \to \Pi(\mathsf{j}(f, \alpha))}$$

(19) $$\frac{\Gamma, \beta \leq \alpha \to \Pi(f\beta)}{\Gamma, x \in \mathsf{j}(f, \alpha) \to x = \langle \mathsf{p}_0 x, \mathsf{p}_1 x \rangle \wedge \mathsf{p}_0 x \leq \alpha \wedge \mathsf{p}_1 x \in f(\mathsf{p}_0 x)}$$

(20) $$\frac{\Gamma, \beta \leq \alpha \to \Pi(f\beta)}{\Gamma, x = \langle \mathsf{p}_0 x, \mathsf{p}_1 x \rangle, \mathsf{p}_0 x \leq \alpha, \mathsf{p}_1 x \in f(\mathsf{p}_0 x) \to x \in \mathsf{j}(f, \alpha)}$$

The rules of inference of $\mathcal{U}_0(\mathsf{FEA})$ are also available in $\mathcal{U}(\mathsf{FEA})$. In particular, $\mathcal{U}(\mathsf{FEA})$ contains the generalized substitution rule (S3): the formulas $\bar{B}$ to be substituted for $\bar{P}$ are now in the language of $\mathcal{L}_2$; the rule in the premise of (S3), however, is required to be in the language $\mathcal{L}_1$.[9] This concludes the description of the predicate unfolding $\mathcal{U}(\mathsf{FEA})$ of FEA.

[8] In the formulation of these rules, it is assumed that β does not occur in Γ.
[9] This last restriction is imposed since predicates may depend on $\bar{P}$.

7.5 The Truth Unfolding $\mathcal{U}_{\mathsf{T}}(\mathsf{FEA})$

In this section we describe an alternative way to define the full unfolding of FEA. The truth unfolding $\mathcal{U}_{\mathsf{T}}(\mathsf{FEA})$ of FEA makes use of a truth predicate T which reflects the logical operations of FEA in a natural and direct way. We will see that the full predicate unfolding $\mathcal{U}(\mathsf{FEA})$ is directly contained in $\mathcal{U}_{\mathsf{T}}(\mathsf{FEA})$.

As in the last section, we want to make the commitment to the logical operations of FEA explicit. This is done by introducing a truth predicate for which truth biconditionals defining the truth conditions of the logical operations hold. The axiomatization of the truth predicate relies on a coding mechanism for formulas. In the applicative framework, this is achieved in a very natural way by using new constants designating the logical operations of FEA. The language $\mathcal{L}_{\mathsf{T}}$ of $\mathcal{U}_{\mathsf{T}}(\mathsf{FEA})$ extends $\mathcal{L}_1$ by new individual constants $\dot{=}$, $\dot{\wedge}$, $\dot{\vee}$, $\dot{\exists}$, as well as constants $\pi_0, \pi_1, \ldots$. In addition, $\mathcal{L}_{\mathsf{T}}$ includes a new unary relation symbol T. The terms and formulas of $\mathcal{L}_{\mathsf{T}}$ are defined in the expected manner. Moreover, we will use infix notation for $\dot{=}$, $\dot{\wedge}$ and $\dot{\vee}$.

The axioms of $\mathcal{U}_{\mathsf{T}}(\mathsf{FEA})$ extend those of $\mathcal{U}_0(\mathsf{FEA})$ by the following axioms about the truth predicate T:

$$
\begin{array}{lrcl}
(\dot{=}) & \mathsf{T}(x \mathrel{\dot{=}} y) & \leftrightarrow & x = y \\
(\dot{\wedge}) & \mathsf{T}(x \mathrel{\dot{\wedge}} y) & \leftrightarrow & \mathsf{T}(x) \wedge \mathsf{T}(y) \\
(\dot{\vee}) & \mathsf{T}(x \mathrel{\dot{\vee}} y) & \leftrightarrow & \mathsf{T}(x) \vee \mathsf{T}(y) \\
(\dot{\exists}) & \mathsf{T}(\dot{\exists}\alpha x) & \leftrightarrow & (\exists \beta \leq \alpha)\mathsf{T}(x\beta) \\
(\pi_i) & \mathsf{T}(\pi_i(\bar{x})) & \leftrightarrow & P_i(\bar{x})
\end{array}
$$

It is easy and natural to assign $\mathcal{L}_{\mathsf{T}}$ terms to $\mathcal{L}_{\mathsf{T}}$ formulas in the following way.

Definition 2 For each formula A of $\mathcal{L}_{\mathsf{T}}$ we inductively define an $\mathcal{L}_{\mathsf{T}}$ term $[A]$ whose free variables are exactly the free variables of A:

$$
\begin{array}{rcl}
[t = s] & := & t \mathrel{\dot{=}} s \\
[P_i(\bar{t})] & := & \pi_i(\bar{t}) \\
[\mathsf{T}(t)] & := & t \\
[A \wedge B] & := & [A] \mathrel{\dot{\wedge}} [B] \\
[A \vee B] & := & [A] \mathrel{\dot{\vee}} [B] \\
[(\exists \alpha \leq \tau) A[\alpha]] & := & \dot{\exists}\tau(\lambda\alpha.[A[\alpha]])
\end{array}
$$

The following lemma can be proved by a trivial induction on the complexity of formulas.

Lemma 3 (**Tarski biconditionals**) Let A be a $\mathcal{L}_{\mathsf{T}}$ formula. Then we have

$$\mathcal{U}_{\mathsf{T}}(\mathsf{FEA}) \vdash A \leftrightarrow \mathsf{T}([A])$$

This lemma shows that in our weak setting, full Tarski biconditionals can be achieved without having to type the truth predicate. Of course, this is due to the fact that negation is only present at the level of sequents.

We close this section by noting that the generalized substitution rule (S3) can be stated in a somewhat more general form for $\mathcal{U}_{\mathsf{T}}(\mathsf{FEA})$. Recall that in $\mathcal{U}(\mathsf{FEA})$, the rule in the premise of (S3) is required to be in $\mathcal{L}_1$. Due to the fact that each $\mathcal{L}_{\mathsf{T}}$ formula can be represented by a term, we can allow rules in $\mathcal{L}_{\mathsf{T}}$ in the premise of the generalized substitution rule, as long as we substitute formulas and associated terms for the predicates P_i and constants π_i simultaneously. In the following we denote by $\Sigma[\overline{B}/\overline{P};\overline{t}/\overline{\pi}]$ the simultaneous substitution of the predicates $\overline{P}$ by the formulas $\overline{B}$ and of the constants $\overline{\pi}$ by the $\mathcal{L}_{\mathsf{T}}$ terms $\overline{t}$. The generalized substitution rule for $\mathcal{U}_{\mathsf{T}}(\mathsf{FEA})$ can now be stated as follows. Assume that the rule $\Sigma_1, \ldots, \Sigma_n \Rightarrow \Sigma$ is derivable with the axioms and rules at hand. Assume further that the terms $\overline{t_B}$ correspond to the $\mathcal{L}_{\mathsf{T}}$ formulas $\overline{B}$ according to the lemma above. Then we can adjoin the rule

$$\Sigma_1[\overline{B}/\overline{P};\overline{t_B}/\overline{\pi}], \ldots, \Sigma_n[\overline{B}/\overline{P};\overline{t_B}/\overline{\pi}] \Rightarrow \Sigma[\overline{B}/\overline{P};\overline{t_B}/\overline{\pi}]$$

as a new rule of inference to our unfolding system $\mathcal{U}_{\mathsf{T}}(\mathsf{FEA})$. This concludes the description of $\mathcal{U}_{\mathsf{T}}(\mathsf{FEA})$.

It is easy to see that the full predicate unfolding $\mathcal{U}(\mathsf{FEA})$ is contained in the truth unfolding $\mathcal{U}_{\mathsf{T}}(\mathsf{FEA})$. The argument proceeds along the same line as the embedding of weak explicit mathematics into theories of truth in Eberhard and Strahm (2012b), which will also be described in some detail in the next section.

7.6 Proof-Theoretical Analysis

In this section we will find a suitable upper bound for $\mathcal{U}(\mathsf{FEA})$ and $\mathcal{U}_{\mathsf{T}}(\mathsf{FEA})$ thus showing that their provably total functions are indeed computable in polynomial time. We will obtain the upper bound via the weak truth theory T_{PT} introduced in Eberhard and Strahm (2012b) and Eberhard (2014), whose detailed and very involved proof-theoretic analysis is carried out in (Eberhard 2014). To be precise, we consider a slight (conservative) extension of T_{PT} which facilitates the treatment of the generalized substitution rule.

Let us sketch the theory T_{PT}; for a more detailed description, the reader is referred to Eberhard and Strahm (2012b), Eberhard (2014). For a more extensive survey on similar kinds of theories in a stronger setting, see Cantini (1996) and Kahle (2007). T_{PT} is based on a total version of the basic applicative theory B for words which was developed in Strahm (2003). In particular, we have a word predicate W which is interpreted as the type of binary strings, constants for some simple functions on the words and a computationally complete combinatory algebra. T_{PT} contains, in addition, a unary truth predicate T which formalizes a compositional truth predicate, where we have constants for the basic logical operations as in the case of the truth unfolding above. The axioms for this predicate T are as usual for theories of truth over an applicative setting with the exception of the axiom for the word predicate.

Only *bounded* elementship in the words can be reflected by T, thus the low proof theoretic strength of T_{PT}.[10] In the axioms below, $y \preceq_{\mathsf{W}} x$ is short for $y \preceq x \wedge y \in \mathsf{W}$. The truth axioms now read as follows:

$$
\begin{array}{lrcl}
(\dot{=}) & \mathsf{T}(x \mathbin{\dot{=}} y) & \leftrightarrow & x = y \\
(\dot{\mathsf{W}}) & x \in \mathsf{W} \to (\mathsf{T}(\dot{\mathsf{W}}xy) & \leftrightarrow & y \preceq_{\mathsf{W}} x) \\
(\dot{\wedge}) & \mathsf{T}(x \mathbin{\dot{\wedge}} y) & \leftrightarrow & \mathsf{T}(x) \wedge \mathsf{T}(y) \\
(\dot{\vee}) & \mathsf{T}(x \mathbin{\dot{\vee}} y) & \leftrightarrow & \mathsf{T}(x) \vee \mathsf{T}(y) \\
(\dot{\forall}) & \mathsf{T}(\dot{\forall} f) & \leftrightarrow & (\forall x)\mathsf{T}(fx) \\
(\dot{\exists}) & \mathsf{T}(\dot{\exists} f) & \leftrightarrow & (\exists x)\mathsf{T}(fx)
\end{array}
$$

In addition, T_{PT} contains unrestricted truth induction on the binary words:

$$\mathsf{T}(r\epsilon) \wedge (\forall x \in \mathsf{W})(\mathsf{T}(rx) \to \mathsf{T}(r(\mathsf{s}_0 x)) \wedge \mathsf{T}(r(\mathsf{s}_1 x))) \to (\forall x \in \mathsf{W})\mathsf{T}(rx)$$

It is shown in Eberhard (2014) that the provably total operations of T_{PT} are precisely the polynomial time computable functions.[11] Moreover, T_{PT} proves the Tarski biconditionals for formulas that contain only bounded occurrences of the word predicate, e.g. formulas of the form $s \preceq_{\mathsf{W}} t$. The corresponding terms for such formulas A are denoted by $\ulcorner A \urcorner$, see Eberhard and Strahm (2012b; Eberhard (2014) for details.

In order to deal with the generalized substitution rule below, we will consider a slight extension T^*_{PT} of T_{PT}. Its language extends the one of T_{PT} by the predicates $P_0, P_1, \ldots$ and the constants $\pi_0, \pi_1, \ldots$. It contains the additional axiom $\mathsf{T}(\pi_i \bar{x}) \leftrightarrow P_i(\bar{x})$ for every $i \in \mathbb{N}$. Since no other axioms for the P predicates and the π constants are present, T^*_{PT} is clearly a conservative extension of T_{PT}.

Next we describe a direct embedding of $\mathcal{U}(\mathsf{FEA})$ into T^*_{PT} which resembles a standard embedding of a theory of explicit mathematics into a theory of truth: We translate the elementship relation with help of the truth predicate and the type constructors by formulating their elementhood condition as in Eberhard and Strahm (2012b). Nevertheless, we have to consider some peculiarities of our system: we take special care of the FEA function constants which are not present in the language of T^*_{PT} and map the two kinds of variables to disjoint sets of T^*_{PT} variables.

Definition 4 (**Translation * of $\mathcal{L}_2$ terms**) The translation of $\mathcal{L}_2$ terms is given inductively on their complexity.

---

[10] We note that T_{PT} can be seen as a polynomial time analogue of a theory of truth of primitive recursive strength studied in Cantini (1995, 2005).

[11] As usual for applicative systems, we call a function $F : \mathbb{W}^n \to \mathbb{W}$ provably total in T_{PT}, if there exists a closed term t_F such that (i) t_F defines F pointwise, i.e. on each standard word, and, moreover, (ii) the following assertion is provable in T_{PT}:

$$x_1 \in \mathsf{W}, \ldots, x_n \in \mathsf{W} \to t_F(x_1, \ldots, x_n) \in \mathsf{W}$$

- Let c be an applicative constant. Then $c^* \equiv c$.
- Let α_i be an $\mathcal{L}$ variable. Then $\alpha_i^* \equiv x_{2i}$.
- Let x_i be a variable of $\mathcal{L}_1 \setminus \mathcal{L}$. Then $x_i^* \equiv x_{2i+1}$.
- $\mathsf{leq}^* \equiv \lambda a.\lambda z.z \doteq \langle \mathsf{p}_0 z, \mathsf{p}_1 z \rangle \dot{\wedge} \dot{\exists} \lambda y.\dot{\mathsf{W}}(\mathsf{p}_1 z) y \dot{\wedge} a\langle \mathsf{p}_0 z, y\rangle$
- $\mathsf{id}^* \equiv \lambda z.z \doteq \langle \mathsf{p}_0 z, \mathsf{p}_1 z\rangle \dot{\wedge} \mathsf{p}_0 z \doteq \mathsf{p}_1 z$
- $\mathsf{con}^* \equiv \lambda a.\lambda b.\lambda z.\ az \dot{\wedge} bz$
- $\mathsf{dis}^* \equiv \lambda a.\lambda b.\lambda z.\ az \dot{\vee} bz$
- $\mathsf{inv}^* \equiv \lambda f.\lambda a.\lambda z.\ a(fz)$
- $\mathsf{j}^* \equiv \lambda f.\lambda a.\lambda z.z \doteq \langle \mathsf{p}_0 z, \mathsf{p}_1 z\rangle \dot{\wedge} \dot{\mathsf{W}} a(\mathsf{p}_0 z) \dot{\wedge} f(\mathsf{p}_0 z)(\mathsf{p}_1 z)$
- $\pi_i^* \equiv \pi_i$
- Let t be $s_0 s_1$. Then $t^* \equiv s_0^* s_1^*$.
- Let G be an n-ary $\mathcal{L}$ function symbol, g_{App} its applicative analogue, and $\bar{t}$ a sequence of terms of suitable arity. Then $G(\bar{t})^* \equiv g_{App}\overline{t^*}$.

For the translation of $\mathcal{L}_2$ formulas, we interpret elementship using the truth predicate as usual and trivialize the relation Π.

Definition 5 (**Translation * of $\mathcal{L}_2$ formulas**) The translation of $\mathcal{L}_2$ formulas is given inductively on their complexity.

- $\Pi(s)^* \equiv 0 = 0$
- $(s = t)^* \equiv s^* = t^*$
- $(s \in t)^* \equiv \mathsf{T}(t^* s^*)$
- $(\exists \alpha \leq \tau) A[\alpha]^* \equiv (\exists \alpha^* \leq_{\mathsf{W}} \tau^*) A^*[\alpha^*]$
- The translation commutes with the connectives $\wedge$ and $\vee$.

The translation * is extended in the obvious way to sequences and sequents of $\mathcal{L}_2$ formulas. Further, for the statement of the embedding theorem below, the following notation is handy.

Definition 6 Let $\diamond$ be an $\mathcal{L}_2$ term, formula or sequence of formulas. Then $\overline{y}(\diamond) \in \mathsf{W}$ denotes the sequence $x_{n_0} \in \mathsf{W}, \ldots, x_{n_m} \in \mathsf{W}$ where the x_{n_i} enumerate the variables with even subscripts occurring freely in $\diamond^*$.

The next two lemmas will be used in the proof of the embedding theorem below. Lemma 7 can be proved by a trivial induction on the complexity of the FEA term. Lemma 8 can be proved by induction on the complexity of A. For the case where A is of the form $(\exists \alpha \leq \tau) B[\alpha]$, we use lemma 7.

Lemma 7 *Let τ be an $\mathcal{L}$ term. Then we have*

$$T^*_{PT} \vdash \overline{y}(\tau) \in W \rightarrow \ \tau^* \in W.$$

Lemma 8 *Let A be an $\mathcal{L}_2$ formula. Then we have*

$$T^*_{PT} \vdash \overline{y}(A) \in W \rightarrow \ T(\ulcorner A^* \urcorner) \leftrightarrow A^*.$$

We are now ready to state the main embedding lemma of $\mathcal{U}(\mathsf{FEA})$ into T^*_{PT} and sketch its proof.

Lemma 9 (**Embedding lemma**) Assume $\mathcal{U}(\mathsf{FEA}) \vdash \Gamma \to A$. Then we have

$$\mathsf{T}^*_{PT} \vdash \overline{y}(\Gamma, A) \in \mathsf{W}, \Gamma^* \to A^*.$$

Proof (Sketch) In order to prove the lemma, one shows a stronger assertion, namely that the * translation of each derivable rule of $\mathcal{U}(\mathsf{FEA})$ is also derivable in T^*_{PT}. Let us exemplary discuss some crucial examples. First, let us look at * translations of axioms of $\mathcal{U}(\mathsf{FEA})$ and distinguish the following cases:

(i) The translations of the axioms about the (word) function symbols of $\mathcal{L}$ hold, because the $\mathcal{L}$ variables are assumed to range over words;
(ii) The translations of the axioms about the applicative combinators clearly hold;
(iii) The translations of the axioms about the correspondence between the function symbols and the applicative constants follow directly from the definition of the translation;
(iv) The translations of the axioms about the predicate constructors hold because of their translation by suitable elementhood conditions and because of the trivial interpretation of the relation Π;
(v) The translations of the axioms $P_i(\bar{x}) \leftrightarrow \langle \bar{x} \rangle \in \pi_i$ clearly hold.

Towards the treatment of the generalized substitution rule, assume that the rule with premises $\Gamma_i[\overline{P}] \to A_i[\overline{P}]$ for $1 \leq i \leq m$ and conclusion $\Gamma[\overline{P}] \to A[\overline{P}]$ is derivable in $\mathcal{U}(\mathsf{FEA})$. Let us look at the * translation of the proof for derivability which is a proof of derivability in T^*_{PT} by induction hypothesis. It can be easily seen that for each sequence of formulas $\overline{B} \in \mathcal{L}_2$ we still have a proof if we replace each occurrence of P_i by B_i^* and each occurrence of π_i by $[B_i^*]$ and add $\overline{y}(\overline{B}) \in \mathsf{W}$ to each antecedent. Here, we use lemma 8 to justify induction and the substituted P biconditionals. Thus the * translation of the rule with conclusion $\Gamma[\overline{B}] \to A[\overline{B}]$ and premises $\Gamma_i[\overline{B}] \to A_i[\overline{B}]$ for $1 \leq i \leq m$ is derivable in T^*_{PT} as desired. This ends the treatment of the generalized substitution rule and hence the proof sketch of the embedding lemma. □

The embedding lemma immediately implies that each function which is provably total in $\mathcal{U}(\mathsf{FEA})$ is also provably total in T^*_{PT} in the usual sense. Since T^*_{PT} is conservative over T_{PT} and the latter proves totality exactly for the polynomial time computable functions (cf. Eberhard (2014)) this delivers the desired upper bound for the unfoldings $\mathcal{U}_0(\mathsf{FEA})$ and $\mathcal{U}(\mathsf{FEA})$. Together with Lemma 1, we obtain sharp proof theoretic bounds.

Theorem 10 *The provably total functions of $\mathcal{U}_0(\mathsf{FEA})$ and $\mathcal{U}(\mathsf{FEA})$ are exactly the polynomial time computable functions.*

An embedding of $\mathcal{U}_\mathsf{T}(\mathsf{FEA})$ into T^*_{PT} can be found in a very similar way as for $\mathcal{U}(\mathsf{FEA})$. Just interpret the constants $\doteq$, $\dot{\wedge}$ and $\dot{\vee}$ as themselves and $\dot{\exists}$ as $\lambda y.\lambda z.\dot{\exists}\lambda x.\dot{\mathsf{W}}yx \,\dot{\wedge}\, zx$. Thus we obtain the following theorem.

Theorem 11 *The provably total functions of $\mathcal{U}_\mathsf{T}(\mathsf{FEA})$ are exactly the polynomial time computable functions.*

This concludes the computation of the upper bounds and hence the proof-theoretic analysis of our various unfolding systems.

References

Cantini, A. (1996). *Logical frameworks for truth and abstraction*. Amsterdam: North-Holland.

Cantini, A. (1997). Proof-theoretic aspects of self-referential truth. In M. L. D. Chiara, et al. (Ed.), *Tenth international congress of logic, methodology and philosophy of science, Florence, August 1995* (Vol. 1, pp. 7–27). Dodrecht: Kluwer.

Cantini, A. (2005). Choice and uniformity in weak applicative theories. In M. Baaz, S. Friedman, & J. Krajíček (Eds.), *Logic colloquium '01, vol. 20 of lecture notes in logic* (pp. 108–138). Association for Symbolic Logic: A K Peters, Wellesley.

Clote, P. (1999). Computation models and function algebras. In E. Griffor (Ed.) *Handbook of computability theory* (pp. 589–681) Amsterdam: Elsevier.

Cobham, A. (1965). The intrinsic computational difficulty of functions. In *Logic, methodology and philosophy of science II* (pp. 24–30). Amsterdam: North Holland.

Eberhard, S. A. (2014). Feasible theory of truth over combinatory logic. *Annals of Pure and Applied Logic*, 165(5), 1009–1033.

Eberhard, S., & Strahm, T. (2012a). Towards the unfolding of feasible arithmetic (Abstract). *Bulletin of Symbolic Logic,* 18(3), 474–475.

Eberhard, S., & Strahm, T. (2012b). Weak theories of truth and explicit mathematics. In U. Berger, H. Diener, P. Schuster, & M. Seisenberger (Eds.), *Logic, construction, computation* (pp. 157–184). Ontos.

Feferman, S. (1975). A language and axioms for explicit mathematics. In J. Crossley (Ed.), *Algebra and Logic, vol. 450 of lecture notes in mathematics* (pp. 87–139). Berlin: Springer.

Feferman, S. (1979). Constructive theories of functions and classes. In M. Boffa, D. van Dalen, & K. McAloon (Eds.), *Logic colloquium '78* (pp. 159–224). Amsterdam: North Holland.

Feferman, S. (1991). Reflecting on incompleteness. *Journal of symbolic logic,* 56(1), 1–49.

Feferman, S. (1996). Gödel's program for new axioms: Why, where, how and what? In P. Hájek (Ed.), *Gödel '96 vol. 6 of Lecture Notes in Logic* (pp. 3–22). Berlin: Springer.

Feferman, S. (2005). Predicativity. In S. Shapiro (Ed.), *The Oxford handbook of philosophy of mathematics and logic* (pp. 590–624). Oxford University Press.

Feferman, S., & Strahm, T. (2000). The unfolding of non-finitist arithmetic. *Annals of Pure and Applied Logic,* 104(1–3), 75–96.

Feferman, S., & Strahm, T. (2010). Unfolding finitist arithmetic. *Review of Symbolic Logic,* 3(4), 665–689.

Ferreira, F. (1990). Polynomial time computable arithmetic. In W. Sieg (Ed.), *Logic and computation, proceedings of a workshop held at Carnegie Mellon University, 1987 vol. 106 of contemporary mathematics* (pp. 137–156). Rhode Island: American Mathematical Society, Providence.

Kahle, R. (2007). *The Applicative Realm*. Habilitation Thesis, Tübingen, 2007. Appeared in Textos de Mathemática 40, Departamento de Mathemática da Universidade de Coimbra, Portugal.

Strahm, T. (2003). Theories with self-application and computational complexity. *Information and Computation,* 185*,* 263–297.

Chapter 8
Some Remarks on the Finite Theory of Revision

Riccardo Bruni

Abstract The *Revision theory of truth* is known, in its *full* (transfinite) form, as one way of dealing with circular concepts (see Gupta and Belnap 1993). The restriction of this approach which is obtained by limiting it to arbitrary, but finite steps of revision is less known, and less studied instead. In this paper we try to assess it, both from the point of view of its motivations, and of those properties which are relevant for establishing a connection with the logical investigation. Finally, we try to see how much of this approach can we make use of in the case of truth.

8.1 Introduction

Alice and Bob are given one euro each. They are told that in case they both decide to invest it on stocks of the Safe Deal Corp., they will be given the euro back. If only Bob decides to make the investment while Alice refuses to do it, the latter will loose the euro she was given, and Bob will receive thrice the investment. If it is Alice who decides to invest and Bob who refuses to do it, the opposite will happen. If they both refuse to make the investment, they will be given twice the initial sum. After pondering over the proposal for a short while, both Alice and Bob (who have been asked to make a decision independently), decide to invest the given euro. Have they made a rational decision?

Let us deal with the situation game–theoretically. The columns of the diagram below correspond to Alice's options, with "I" staying for "Invest" and "N" for "Not invest". Rows correspond to Bob's actions instead, and they are marked by the corresponding lower–case letters. Payoffs take the form of pairs, the first member of which is Alice's payoff, the second is Bob's. Then we have:

R. Bruni
Università degli Studi di Firenze, Firenze, Italy
e-mail: riccardobruni@hotmail.com

T. Achourioti et al. (Eds.), *Unifying the Philosophy of Truth,* Logic, Epistemology, and the Unity of Science 36, DOI 10.1007/978-94-017-9673-6_8

	i	n
I	1,1	3,0
N	0,3	2,2

Let us try to figure out how the two characters might have reasoned. Alice assumes that Bob decides to invest. She then notices that by investing she gets the best payoff. If Bob refuses to invest instead, then investing yields again the best outcome for her. So, she decides to invest. Bob reasons in a similar manner and, being the payoffs distributed symmetrically, he is brought to conclude the same.

This reasoning seems unexceptionable[1]. Let us make the situation more realistic, and say that Alice has an intuition regarding what is more convenient to do for her. For instance, she could be hesitating toward the idea of entering the stocks market. Hence, the previous reasoning would have the effect of making Alice change her mind (we are here assuming that Alice and Bob trust logic more then their own prejudices). In case she was already oriented toward stock investment from the very beginning, Alice would pursue this option even more firmly afterwards.

We are tempted to extract some kind of a moral from this story. Alice and Bob are dealing with a problem of *circularity*. The situation requires indeed each of them to single out their best option, on the basis of what is best to do for the other. The argument by means of which they do this is based on *making hypothesis* regarding what might be convenient for the other to do, and for the reasoner also. In the one case, the initial hypothesis is confirmed, in the other it is *revised*. Moreover, the story seems to be tied up with the proposed analysis in a very *natural* way.

This is a good basis for introducing, and evaluating the machinery of finite revision for dealing with circular concepts.

8.2 Circular Concepts by Finite Revision

In the fifth chapter of their book (Gupta and Belnap 1993), Gupta and Belnap introduce a cut–down version of the Revision Theory of Truth as an intermediate step toward the full theory as it is best known. As this part of Gupta and Belnap's work, as well as the articles by Gupta (2000) and Chapuis (2000) which are based on it, seem to have attracted little attention so far, our plan is to review the theory of finite revision in some details here, and show how it reconciles with the introductory example.

[1] Notice that it is irrelevant that Alice knows Bob's payoffs and vice versa. This depends upon their distribution in *this* game, and the fact that one and the same option for both players (not investing), is strictly "dominated" by the other alternative, as one would say it in game–theoretic terms.

Let $\mathcal{L}$ be a given first–order predicate formal language. Let $\mathcal{L}^+$ indicate the extension of it by means of a unary predicate $G(x)$. Assume that $A_G(x, G)$ is a formula of $\mathcal{L}^+$ which is an accepted *definition* of G. The natural way to interpret this is to say that the intended model $\mathbf{M}^+$ of $\mathcal{L}^+$ is an expansion of the intended model $\mathbf{M}$ of $\mathcal{L}$ by the set $\mathbf{G}$ of those elements of the domain of $\mathbf{M}$ which produce valid instances of $A_G(x, G)$. However, since the latter is a formula of the extended language, it features occurences of G itself and requires that one knows the set $\mathbf{G}$ already, in order to determine which instances of the formula are valid in a classical setting.

The revision theory is conceived in order to overcome such a difficulty. In order to make use of the example, we may assume that $\mathcal{L}$ is a language featuring terms a_1, a_2, b_1, b_2 for *actions* of the players (with the intended meaning that a_1 represents Alice's option of investing, and a_2 of refusing to invest, while b_1 and b_2 are names for Bob's corresponding actions). Furthermore, the alphabet of $\mathcal{L}$ is equipped with a function symbol u which, given one of the two players, say A or B for simplicity, and any given play, that is any pair of actions, represents the utility for the chosen player under the given scenario. So, for instance, by $u_A(a_1b_1)$ it is meant Alice's payoff in the event that both she and Bob decide to invest. Payoffs are then compared by means of an order relation $\geq$.

This base language is upgraded to $\mathcal{L}^+$ by means of a new unary predicate $R(x)$ for "x is a rational action". By taking Alice and Bob's reasoning in the example as a paradigm, the two players favour the action which is best in terms of utility. This property is easy to describe by the linguistic means that we have at our disposal. As a matter of fact, let φ_1^A be the formula

$$(R(b_1) \wedge u_A(a_1b_1) \geq u_A(a_2b_1)) \vee (R(b_2) \wedge u_A(a_1b_2) \geq u_A(a_2b_2))$$

This expresses the fact that a_1 is the best possible action for Alice, either if it is rational for Bob to invest, or not.

Similarly, one can easily imagine how to write down a formula φ_2^A which does the same for Alice's action a_2, and formulas φ_1^B, φ_2^B for Bob's actions.

By logic, one obtains a general definition of rationality as the following combination of these formulas

$$R(x) \Leftrightarrow (x = a_1 \wedge \varphi_1^A) \vee (x = a_2 \wedge \varphi_2^A) \vee (x = b_1 \wedge \varphi_1^B) \vee (x = b_2 \wedge \varphi_2^B)$$

Clearly, this says that x is rational in case it is the most convenient action for one of the two players. Let the righthand side of this defining equivalence be abbreviated by $A_R(x, R)$ in the following.

Let us go back to the general case, and let an interpretation $\mathbf{M}$ of $\mathcal{L}$ be fixed. An *hypothesis* is a subset H of the domain $|\mathbf{M}|$ of $\mathbf{M}$. Let us assume that the set of terms of $\mathcal{L}$ (hence, of $\mathcal{L}^+$) contains in addition all of the names for elements of $|\mathbf{M}|$ ($\overline{a}$ being the name of $a \in |\mathbf{M}|$). This is known to be possible without loss of generality.

Let $\models$ indicate the usual, classical validity relation for formulas of $\mathcal{L}$ with respect to the interpretation $\mathbf{M}$. For every hypothesis H, and A, B formulas of $\mathcal{L}^+$, this relation is extended to the expanded language by the following clauses where P

stays for any atomic predicate of $\mathcal{L}$ (hence P is different from R):

$$\begin{array}{lll}
(\mathbf{M}, H) \models R(\overline{a}) & \text{iff} & a \in H \\
(\mathbf{M}, H) \models P(t_1, \ldots, t_n) & \text{iff} & \mathbf{M} \models P(t_1, \ldots, t_n) \\
(\mathbf{M}, H) \models \neg A & \text{iff} & (\mathbf{M}, H) \not\models A \\
(\mathbf{M}, H) \models A \wedge B & \text{iff} & (\mathbf{M}, H) \models A \text{ and } (\mathbf{M}, H) \models B \\
(\mathbf{M}, H) \models \exists x A & \text{iff} & \text{for some } a \in |\mathbf{M}|, (\mathbf{M}, H) \models A[x/\overline{a}]
\end{array}$$

Let this relation be defined for the other connectives of $\mathcal{L}^+$ and for the universal quantifier as usual.

The idea of the revision theory is that hypotheses fix the extension of the circular predicate temporarily. Then, this tentative extension is refined by means of a *revision operator* δ_A (depending on the defining condition $A_G(x, G)$ of the circular predicate G), which associates H to the hypothesis $\delta_A(H)$ according to the condition, for every $a \in |\mathbf{M}|$:

$$a \in \delta_A(H) \Leftrightarrow (\mathbf{M}, H) \models A_G(\overline{a}, G)$$

If one thinks of this condition with respect to the example (i.e., with respect to the formula $A_R(x, R)$ above), it is very easy to see that this is actually what Alice and Bob are using in order to refine their intuitions regarding what is best for them to opt for. For, according to the description we have made, they "calculate" the action yielding the best payoff, under some initial suppositions about what to do. Suppositions correspond to hypotheses. The revision operator translates at the formal level Alice and Bob's calculation of the action which fulfils the underlying definition of rational choice. The latter, in turn, embodies the idea that it is rational what guarantees the best payoff.

Payoffs distribution determines whether the outcoming hypothesis $\delta_R(H)$ (δ_R being defined on the basis of the definition $A_R(x, R)$ of rational choice), can either revise H, or not. So, for instance, in the proposed example we have that $\delta_R(H) = H$ for $H = \{a_1, b_1\}$, while for every $H' \neq H$ one has $\delta_R(H') \neq H'$ instead[2].

In general, one can expect that the revision procedure requires further applications of the base machinery in order to yield solutions. Hence, the following iteration of the operator along $\mathbb{N}$ for every hypothesis H is defined:

$$\delta_A^0(H) = H$$
$$\delta_A^{n+1}(H) = \delta_A(\delta_A^n(H))$$

Let us call $(\delta_A^n(H) \mid n \in \mathbb{N}, H \text{ hypothesis})$ a *revision evaluation sequence* (RES, henceforth).

---

[2] For the sake of this account, we only consider hypotheses which are different from the empty set (i.e., hypotheses that appear in a diagram representing the given situation as a game). It is clear that if one relaxes this condition then for $H = \emptyset$ one has $\delta_R(H) = H$ as well.

The next problem is to set a standard on how to single out "solutions" of any given RES. In the chosen example, that was an easy task due to the fact that investing was best for both players independently of the opponent's choice. It would be too strict to stick to hypotheses which can replicate this condition in all possible situations. Rather, Gupta and Belnap (but see also Gupta (2000)), have opted for the property of being *reflexive* in order to stress the fact that an hypothesis is "reliable":

Definition 1 1. An hypothesis H is said to be *n–reflexive* if $\delta_A^n(H) = H$.

2. An hypothesis H is *reflexive*, if it is n–reflexive, for some $n > 0$.

The reader should notice, and keep in mind the restriction by means of which the notion of reflexivity is defined out of the more general notion of n–reflexivity (where $n = 0$ is allowed).

The property of (n–)reflexivity relaxes the feature of the solution $H = \{a_1, b_1\}$ from the chosen example, which is a *fixed point* of the revision operator (i.e., such that $\delta_R(H) = H$, hence 1–reflexive). In addition, a reflexive hypothesis *needs not* to be such that $\delta(H') = H$ for every $H' \neq H$[3].

Informally speaking, the idea is that if an hypothesis recurs in this sense in a RES, then it is deemed to provide reliable information. This latter fact, is explained in terms of a notion of *validity* for formulas of $\mathcal{L}^+$, which is defined accordingly. In fact, there are two slightly different versions of such a notion.

The first one comes from Gupta and Belnap's book (Gupta and Belnap 1993, Def. 5A.2, p. 147), and it makes a direct use of the notion of n–reflexivity:

Definition 2 Let $\mathcal{L}$ be a first–order language, and $\mathcal{L}^+ = \mathcal{L} \cup \{G(\cdot)\}$. Let $A_G(x, G)$ be the defining condition for the predicate G. Then, we say that:

1. a sentence B of $\mathcal{L}^+$ is *m–valid in* a given interpretation $\mathbf{M}$ of $\mathcal{L}$ ($m \in \mathbb{N}$), if, and only if there exists $k \in \mathbb{N}$, such that, for every m–reflexive hypothesis H, $(\mathbf{M}, \delta_A^k(H)) \models B$;
2. a sentence B of $\mathcal{L}^+$ is *m–valid* ($m \in \mathbb{N}$) if, and only if B is *m–valid in* $\mathbf{M}$, for every interpretation $\mathbf{M}$ of the base language.

Let, in the following, $\Vdash_n B$ stay for "B is n–valid".

One can try to give an explanation of this definition along the following lines. Reflexivity is, as we said, a feature that makes an hypothesis reliable. Since the revision process itself is trusty, reliability is preserved along any RES which starts from a reflexive hypothesis. A formula is m–valid if it is validated *at one and the same stage k* of every RES starting from an m–reflexive hypothesis H. So, validity here depends upon fulfilling a condition which is made strong by the requirement of the stage k in question being uniform for every m–reflexive hypothesis.

[3] This feature of the given game depends upon the payoffs distribution, as we said. Games like the one we have considered in the introduction are indeed *regular*, as Gupta (2000) calls them (as they refer to circular definitions which can be said to be regular in the sense of (Gupta and Belnap 1993, Def. 5A.8, p. 149)). As it is made clear by Gupta himself, regularity plays an important role in the revision–theoretic account of finite games, though it is too much a special feature to become a standard (see also Bruni and Sillari 2011 on this).

It should be noticed that this (maybe) disturbing aspect of the above definition can be eliminated, though only partially. This has been shown, in a remark by Gupta and Belnap (1993, p. 147) themselves. We reproduce it here, along with its proof, for the sake of self–containedness:

Lemma 1 *If $n > 0$, any given formula B of $\mathcal{L}^+$ is n–valid if, and only if, for every interpretation* $\mathbf{M}$ *of* $\mathcal{L}$, $(\mathbf{M}, H) \models B$ *for every n–reflexive hypothesis H.*

Proof Let $\mathcal{L}$, $\mathcal{L}^+ = \mathcal{L} \cup \{G(\cdot)\}$, and δ_A be as before. Let $\mathbf{M}$ be an arbitrary, but fixed interpretation of $\mathcal{L}$.

The direction from right to left of the lemma is trivial. For, if $(\mathbf{M}, H) \models B$ holds for every n–reflexive hypothesis H, then $(\mathbf{M}, \delta_A^0(H)) \models B$ holds too by definition of δ_A^k for every n–reflexive hypothesis H. Hence, B is n–valid in $\mathbf{M}$, which was so chosen to be generic.

For the other direction, assume that H is n–reflexive with $n > 0$. Then, clearly $\delta_A^k(\delta_A^{(nk)-k}(H)) = H$ for every $k \in \mathbb{N}$ (since $\delta_A^n(\delta_A^m(H)) = \delta_A^{n+m}(H)$ for every $n, m \in \mathbb{N}$, easily follows from the definition of $\delta_A^n(\cdot)$). Moreover, $\delta_A^{(nk)-k}(H)$ is n–reflexive. As a matter of fact, having noticed that $\delta_A^n(\delta_A^m(H)) = \delta_A^{n+m}(H)$, one has also that

$$\delta_A^n(\delta_A^m(H)) = \delta_A^{n+m}(H) = \delta_A^{m+n}(H) = \delta_A^m(\delta_A^n(H))$$

Hence,

$$\delta_A^n(\delta_A^{(nk)-k}(H)) = \delta_A^{(nk)-k}(\delta_A^n(H)) = \delta_A^{(nk)-k}(H)$$

by H being n–reflexive.

Now, assume that $(\mathbf{M}, \delta_A^k(H)) \models B$ holds for some $k \in \mathbb{N}$, and for every n–reflexive hypothesis H (that is, assume that B is n–valid in $\mathbf{M}$). Then, in particular, $(\mathbf{M}, \delta_A^k(\delta_A^{(nk)-k}(H))) \models B$ for every H which is n–reflexive. Then, $(\mathbf{M}, H) \models B$ holds for every such an H. ■

Having shown this, one can stick to the alternative, simplified definition of n–validity *for* $n > 0$ which goes as follows:

Definition 3 Let $\mathcal{L}$ be a first–order language, and $\mathcal{L}^+ = \mathcal{L} \cup \{G(\cdot)\}$. Let $A_G(x, G)$ be the defining condition for the predicate G. Then, we say that:

1. a sentence B of $\mathcal{L}^+$ is *m–valid in* a given interpretation $\mathbf{M}$ of $\mathcal{L}$ ($m > 0$), if, and only if $(\mathbf{M}, H) \models B$ for every m–reflexive hypothesis H;
2. a sentence B of $\mathcal{L}^+$ is *m–valid* ($m > 0$) if, and only if B is *m–valid in* $\mathbf{M}$, for every interpretation $\mathbf{M}$ of the base language.

Why getting rid of the $m = 0$ case? The reason one can think of so far is that the concept of 0–reflexivity is trivial since it applies to *every hypopthesis* by definition of the revision operator.

Now, Gupta (2000, p. 126) introduces another notion of validity, which, as we shall notice, is related to the previous one. It goes as follows:

Definition 4 Let $\mathcal{L}$ be a first–order language, and $\mathcal{L}^+ = \mathcal{L} \cup \{G(\cdot)\}$. Let $A_G(x, G)$ be the defining condition for the predicate G. Then, we say that:

1. a sentence B of $\mathcal{L}^+$ is *valid in* $\mathbf{M}$ if, and only if $(\mathbf{M}, H) \models B$ for every reflexive hypothesis H;
2. a sentence B of $\mathcal{L}^+$ is *valid* if, and only if B is *valid in* $\mathbf{M}$, for every interpretation $\mathbf{M}$ of the base language.

Let $\Vdash B$ stay in the following for "B is valid" in this latter sense.

Notice that by referring to hypotheses which are *reflexive*, this definition is meaningful for all hypotheses which are n–reflexive for $n > 0$.

Having shown that the original uniformity requirement can be dispensed with in the case of n–validity, if $n > 0$, there is no simplification in this latter definition of validity in this sense. Actually, one has, that any given formula B of $\mathcal{L}^+$ is valid if, and only if B is n–valid for every $n > 0$.

The real advantage is that, by sticking to the notion of validity, one has to deal with *just one* semantical notion, rather than with infinitely many validity notions, or, to say it differently, with a notion which is stratified in infinitely many layers.

However, this has a cost. For, in the case of validity one has to make sure that reflexive hypotheses do exist. As a consequence, Gupta (2000, p. 126) notices that the semantics of finite revision as defined in this latter way, is meaningful for circular definitions respecting the following *finiteness requirement* FR:

FR For all models $\mathbf{M}$ of $\mathcal{L}$, there exists a natural number k such that, for every hypothesis H, $\delta_A^k(H)$ is reflexive.

Notice that FR does not limit itself to state that reflexive hypotheses exist. It says that they do exist in a uniform manner: there exists a $k \in \mathbb{N}$, such that the k–th stage of *any* given RES is reflexive. So, it seems that a uniform condition that we were able to throw out of the door in the case of n–validity, came back in from the window in the case of validity *where it seemed that it was not*[4]. The observation is less harmful than it seems, since the only known application of the semantics of finite revision is the notion of rational choice in finite games of (Gupta 2000), and the circular definition of it that one can give along the previous lines of argument *does* respect FR in this form. However, dealing with truth by finite revision allows to make a (partial) case for the missing $n = 0$ clause. This we shall see in § 8.4.

Of course, this is not to say that the notion of n–validity does not give rise to worries of any sort. Conceptually speaking, for instance, one would like to find grounds supporting the intuition that there is a difference between a formula A which requires the revision process to be applied, say, once in order to get a reason for believing that it is valid, and a formula A' which requires that the whole procedure be carried out 70 times to reach a similar conclusion. At the present stage of development

[4] As a matter of fact, this impression is illusory: the reason why FR needs to be formulated in this way is stated in § 8.3, and is related to the uniform way in which the original concept of n–validity was defined.

of the theory, and to the best of the author's knowledge, there is nothing that one can make use of in order to say what this difference amounts to in the end.

This suggests, at least, that we should ponder the features of the finite theory of revision with a bit more care.

8.3 Naturality, Complexity and Logicality

As the story of Alice and Bob, and our analysis of it make clear, the treatment of circular concepts by finite revision is very *natural*, one could say first. The revisionary way of identifying solutions for situations involving circular concepts, is tied up with the informal reasoning that "real" people can be thought of pursuing, while making decisions in "real" situations (some caution is required as to how much game theory is regarded as a legitimated way of representing real people dealing with real situations).

In order to stress the point, it can be useful to make a comparison with the full, transfinite revision theory. This latter extension of the finite theory, is obtained by giving a limit clause for the iteration of the revision operator δ. Notoriously, different proposals have been made in this respect (the interested reader is again referred to (Gupta and Belnap 1993)). Here we confine ourselves to the simplest of them, which is due to Hans Herzberger (1982).

Having defined $\delta_A^n(H)$ for every $n \in \mathbb{N}$ and for every hypothesis H, we define $\delta_A^\alpha(H)$ for every ordinal number α and H by:

$$\begin{aligned} \delta_A^0(H) &= H \\ \delta_A^{\alpha+1}(H) &= \delta_A(\delta_A^\alpha(H)) \\ \delta_A^\lambda(H) &= \{x | \exists \alpha < \lambda \forall \beta < \lambda (\alpha \leq \beta \rightarrow x \in \delta_A^\beta(H))\}, \lambda \text{ limit} \end{aligned}$$

Clearly, the first two clauses are as in the finite theory, except that they have been extended to ordinals. The idea behind the last clause in this definition should be clear: one retains at limits what behaves stably below the limit ordinal, where "stably" here means that it remains inside the revision sequence from one point onwards.

Now, this clearly represents a legitimate extension of the notion of reliability for an hypothesis, that we were referring to before. As far as the successor stage is concerned, reliability stems from accepting the revision–theoretic machinery. In turn, the intuitiveness of it comes from the analysis of actual forms of reasoning. This connection with reality is lost for the transfinite extension of the theory. This is due to the need of referring to ordinal numbers already. Apart from that, it is a notable fact that no proposal concerning a limit clause received a general consensus. Beside Herzberger's limit rule, we have also Gupta's and Belnap's.

By the way, it is hard to imagine that things could be different. For, any limit clause is motivated by the nature of the process being transfinite, and by the involved stage being a limit ordinal, rather than by the need of adding another operation to the revision process due to the informal reasoning that we want to capture. There seems

to be no real story here providing the required motivation. Clearly, the transfinite iteration seen as a completion of the finite theory, must be done coherently with the "spirit" of revision. As we said, it is easy to acknowledge coherence to the above extension proposal. The remarks we have just made, show that coherence seems not to be enough.

So, the finite theory of revision is natural, and its naturality comes from the fact that it needs not to compromise itself to any limit rule as it is shown by the comparison with the transfinite theory. More than that, the theory is "acceptable" also from the point of view of its complexity[5]. This observation can be made precise in the following manner.

By taking inspiration from the transfinite case, we focus on the standard structure $\mathbb{N}$ of arithmetic. Hypotheses take here the form of subsets of $\mathbb{N}$. The base language $\mathcal{L}$ is the language $\mathcal{L}_{PA}$ of Peano arithmetic, and $\mathcal{L}^+ = \mathcal{L}_{PA} \cup \{G(\cdot)\}$. Let the circular predicate be defined by any formula $A(x, G)$ of $\mathcal{L}^+$, and set, for every $X \subseteq \mathbb{N}$

$$\delta_A(X) = \{n \in \mathbb{N} | (\mathbb{N}, X) \models A[x/\overline{n}, G]\}$$

Then, put:

Definition 5 Let, for every $X \subseteq \mathbb{N}$

$$ref_A(n, X) := \delta^n(X) = X$$
$$ref_A(X) := (\exists n \in \mathbb{N}^+) ref_A(n, X)$$

(where $\mathbb{N}^+ = \mathbb{N} \setminus \{0\}$).
Then, we say of any $Z \subseteq \mathbb{N}$:

1. For every $n \in \mathbb{N}^+$, Z is n–definable if, and only if

$$x \in Z \Leftrightarrow \forall X[ref_A(n, X) \rightarrow x \in X]$$

2. Z is definable if, and only if

$$x \in Z \Leftrightarrow \forall X[ref_A(X) \rightarrow x \in X]$$

Notice that we have here referred to the simplified definition of n–reflexivity, the one for the case $n > 0$ only, and to the related notion of reflexivity.

The question is what kind of sets (i.e., of what logical complexity) get defined in these senses as $A(x, G)$ is allowed to vary. It is an easy task to verify the following:

Lemma 2 *1. Every n–definable set Z is* Π^1_1.
2. Every definable set Z is Π^1_1.

Proof Fixed an interpretation **M** of the given language, one notices that $x \in \delta_A(X)$ has the same logical complexity as the relation $\models$ we have introduced in § 2, for every

[5] The material below follows a comment and a suggestion in this sense due to Philip Welch.

$X \subseteq \mathbb{N}$ and $x \in \mathbb{N}$. By known arguments the latter can be shown to be of complexity Δ^1_1 as the defining condition $A(x, G)$ is allowed to vary (see, e.g., Takeuti 1987). This can be used to prove by induction that $x \in \delta^n_A(X)$ is of complexity Δ^1_1 as well. It follows that $ref_A(n, X)$ is Σ^1_1 definable. Hence the theorem. ■

Of course, this is only a sketch of a more comprehensive study of set–definability under finite revisions. Since this would be largely useless for the aim of the present paper, we will eventually take the issue up elsewhere. However, this is enough to claim a superiority of the finite theory over the transfinite one in this sense. The available results for the transfinite theory make it legitimate to conclude that the finite theory is more satisfactory than the full one from the viewpoint of these complexity calculations[6].

However, one may argue that this way of extracting the complexity of the theory can only give partial information. For, in the end we have defined a notion of validity for formulas of a language (two, in fact). So, we have a semantics. Then, should not be better to ask the question: do we have a logical calculus which is adequate to this semantics? This is another way of dealing with complexity in the end (in particular, to deal with the complexity of the involved notion of validity). Also, it is another way of carrying out the comparison with the transfinite theory, which lacks such a connection with an axiom system.

In the case of the finite theory of revision, the answer is positive if one bases oneself on the notion of n–reflexivity (with $n = 0$ included). Recall that a formula B from a language with a circular predicate is said to be n–valid if, given any interpretation $\mathbf{M}$ of the base language, A is made true by any set $\delta^k(H)$ for some $k \in \mathbb{N}$ and for every n–reflexive hypothesis H (Def. 2 from § 2). Then, Gupta and Belnap (1993, Chap. 5) devised a family $(\mathbf{C}_n)_{n\in\mathbf{N}}$ of Fitch–style natural deduction calculi for which they prove the following[7]:

Theorem 1 *Let $\mathcal{L}$ be any first–order predicate language. Let $G(x)$ be a unary predicate, $\mathcal{L}^+ = \mathcal{L} \cup \{G\}$, and $A_G(x, G)$ be a formula of $\mathcal{L}^+$ which is a defining condition for G. Then for every $n \in \mathbb{N}$ and for every formula B of $\mathcal{L}^+$*

$$\Vdash_n B \Leftrightarrow \vdash_{\mathbf{C}_n} B$$

In the case of the other definition of validity, that is the one of Def. 4, the answer is positive as well. As a matter of fact, by exploiting the fact that this semantics is meaningful only for circular definitions respecting the requirement FR, one easily proves[8]:

---

[6] See (Welch 2003). There it is shown that sets which are revision–theoretically definable with respect to the transfinite process are of a complexity that is at least Δ^1_2 (Π^1_2 in certain cases), independently of what limit rule is used.

[7] Gupta and Belnap's book aims at dealing with a more general situation, where a base language $\mathcal{L}$ is inflated to a language $\mathcal{L}^+$ which contains a *countable set* $\{G_i\}_{i\in\mathbb{N}}$ of circular concepts defined by formulas A_i of $\mathcal{L}^+$ (hence, a situation where any of these G_i *depends* upon all G_j's which occur in A_i for $i, j \in \mathbb{N}$). The result we are referring to, that is (Gupta and Belnap 1993, Thrm. 5B.1, pp. 162 ff.), is proved to hold with respect to this expanded setting.

[8] The result is stated, without proof, by Gupta in (2000).

Theorem 2 *For every formula B of $\mathcal{L}^+$*

$$\Vdash B \Leftrightarrow \vdash_{C_0} B$$

Proof By the previous result, this amounts to proving that B is valid in the sense of Def. 4 if, and only if B is 0–valid. Now, if B is 0–valid, then there exists a natural number m such that, for every hypothesis H

$$(\mathbf{M}, \delta^m(H)) \models B$$

In particular, if H is reflexive $(\mathbf{M}, \delta^m(H)) \models B$. Then, by the argument we have used for the sake of Lemma 1 above, we have that $(\mathbf{M}, H) \models B$ must hold for every reflexive H. Hence, B is valid.

Conversely, assume that B is valid instead. Then, $(\mathbf{M}, H) \models B$ is the case for every reflexive hypothesis H. Since moreover FR is respected, there exists $k \in \mathbb{N}$ such that, for every hypothesis H', $\delta_A^k(H')$ is reflexive. Hence, $(\mathbf{M}, \delta_A^k(H')) \models B$ holds for every H'. This means that B is 0–valid. ■

The reader should notice that it is necessary for this argument that FR be formulated in an uniform manner, i.e. that there exists one and the same $k \in \mathbb{N}$ such that $\delta_A^k(H')$ is reflexive for every hypothesis H'. If that were not the case, than we would have no adequate syntactic notion of derivability to the notion of validity $\Vdash$. This shows also why the case $n = 0$ in the definition of n–reflexivity can be dispensed with, at least for circular definitions respecting FR. It is because that case is captured by the notion of validity already.

Finally, the answer to the question we raised (do we have axiom systems adequate to the semantics of finite revision?) is now complete.

However, still someone may not be entirely satisfied by the way we answered this question in the positive. For, one may also wish to know whether we have *nice* calculi which do the job. In terms of standard proof–theory, this means whether we have Hilbert–style calculi, of which we have a sequent (Gentzen–style) versions, that further feature a cut–elimination theorem showing them to be analytic.

The answer is again positive, but in order to say something more about that, we need to go a little bit into the details of the axiomatization.

8.3.1 On the Logic of Finite Revision

The aim here is to illustrate some conceptual aspects related to logical investigations over the semantics of finite revision. For all technical details, we refer the reader to (Bruni 2012) and (Gupta and Belnap 1993).

The main idea on which the formalism is based, an idea which comes from Gupta and Belnap, is to use an indexed language, with indices representing stages in a revision–theoretic evaluation of a formula. So, for a given formula A of the chosen language, and for an index i, A^i means: "*A holds at the i–th stage of a RES*".

Having fixed a standard first–order predicate language $\mathcal{L}$ and having upgraded it to $\mathcal{L}^+ = \mathcal{L} \cup \{G(\cdot)\}$ as before, index terms are numerals $\overline{p}$ for every $p \in \mathbb{Z}$. In the following, we are going to use $i, j, h, \ldots$ as metavariables for index terms and, having made clear that they are numerals, we use $i+1, i+n, \ldots$ with the expected meaning.

Indexed formulas are defined in such a way that, for every formula A of $\mathcal{L}^+$, A^i is a formula of the indexed language for every index term i (so, notice that formulas of $\mathcal{L}^+$ are indexed by placing the index *outside* it).

In order to reach an Hilbert–style axiomatization which would correspond to Gupta and Belnap calculi, our idea was to mimic derivability in $\mathbf{C}_n$ by means of a specifically devised implication connective $\rightarrow$. So, to every rule $(\mathcal{R})$ of $\mathbf{C}_n$ there corresponds an axiom (R) of the Hilbert–style system $\mathbf{HC}_n$, according to the schema

$$\frac{A}{B}\ (\mathcal{R}) \Rightarrow A \rightarrow B\ (R)$$

Since A, B are indexed formulas already, hence they are formulas C^i, D^j, one has to allow formulas of the form $C^i \rightarrow D^j$ on the side of Hilbert–style systems. This means that *indices need to distribute over the new implication connective*. Moreover, we need to allow the antecedent and the consequent of $\rightarrow$–formulas to carry *different* index terms (see the special axioms below).

By referring to the previous informal reading of indexed formulas, the meaning of a formula of the form $A^i \rightarrow B^j$ is then: "*If* A^i *is the case*" (hence, if A holds at the ith–stage of a RES) "*then* B^j *is the case as well*"[9].

Besides that, introducing a new connective has the expected effect on the logical base of the axiom systems $(\mathbf{HC}_n)_{n\in\mathbb{N}}$. As a matter of fact, one has to fix the "meaning" of the new arrow by setting some logical principles. It turns out that, insofar as the goal of representing derivability in systems $\mathbf{C}_n$ is concerned, this is possible by indexing one's preferred version of first–order classical logic[10]. The only prescription is that for those logical axioms and rules featuring occurrences of the new implication connective (which is necessary since we are trying to mimic axiomatically the logical inference rules of the original calculi $\mathbf{C}_n$), *one and the same index term is distributed over the antecedent and the consequent*. To make a straightforward example, this is how appears one of the basic axioms for conjunction:

$$(A \wedge B)^i \rightarrow A^i$$

for every formulas A, B of $\mathcal{L}^+$, and index term i.

[9] It would be natural to ask to what an extent one can represent derivability in systems $\mathbf{HC}_n$ themselves by the new implication connective. Despite its interest, this topic would take us afield. Hence, we refer the interested reader to (Bruni 2012), where the issue is dealt with at length.

[10] The process of indexing a standard logic calculus hides some subtleties, which is nonetheless unnecessary to go into here. We refer the interested reader to (Bruni 2012) for details.

Anyway the new connective *is* an implication. Hence, the logical calculus will include the following indexed version of the rule of modus ponens $\mathrm{MP}_{\rightarrow}$:

$$\frac{A^i \rightarrow B^j \quad A^i}{B^j} \ (\mathrm{MP}_{\rightarrow})$$

This should be compared with the rule of modus ponens for the usual, material implication, which would rather read[11]:

$$\frac{(A \supset B)^i \quad A^i}{B^i} \ (\mathrm{MP})$$

This entails that *derivations employing only logical axioms and rules of inference cannot cause the index term of the involved formulas to increase, or decrease*.

The special axioms of systems $(\mathbf{HC}_n)_{n\in\mathbb{N}}$, take direct inspiration from the basic features of the finite revision semantics, since they derive from rules of systems $\mathbf{C}_n$ according to the above correspondence schema. So, for instance, if $n = 0$, then the goal is to devise derivability in $\mathbf{HC}_0$ so to let it correspond to 0–validity. As a consequence, this system features an axiom of *index shift*, which, for every formula B of $\mathcal{L}$ (hence, any formula which *does not feature occurrences of the circular concept* G) and for every index terms i, j, reads

$$B^i \rightarrow B^j \ (\mathrm{IS})$$

In the proposed, informal interpretation, this says that if the formula B holds at any given stage i in a RES, then it holds also at any other stage j. This corresponds to the fact that formulas of the base language retain, in the semantics of finite revision, the truth value they have on the basis of the chosen model of $\mathcal{L}$, *independently of the procedure of revising hypotheses*.

Moreover, the system $\mathbf{HC}_0$ features *G–definition* axioms, namely, for every index term i, and t term of $\mathcal{L}$, instances of the form

$$A_G(t, G)^i \leftrightarrow G(t)^{i+1} \ (\mathrm{DEF})$$

Clearly, this axiom tries to capture the very definition of the revision operator δ. If one looks at revision sequences as processes by means of which the extension of a circular predicate is fixed through stages of approximation, this can be viewed as saying: (i) having verified that an instance $A_G(t, G)$ of the condition defining the circular predicate G holds at stage i, *allows us to conclude* that $G(t)$ holds at stage $i + 1$; (ii) that $G(t)$ holds at any stage $i + 1$ (remember that index terms are *integers*), *requires that* the corresponding instance $A_G(t, G)$ is validated at stage i. So, the two directions of this axiom can be seen as representing two ways of going through any RES: upwards (the left–to–right direction), and downwards (the right–to–left one). This will be used for the sake of Theorem 4 in § 8.4[12].

[11] We use the symbol $\supset$ for the standard implication connective.

[12] Also, the reader is invited to see Gupta and Standefer (2011) for an actual elaboration of this view.

Systems $\mathbf{HC}_n$ for $n > 0$, are obtained from $\mathbf{HC}_0$ by extending it with a generalization of index shift that reads, for every formula B *of the extended language* $\mathcal{L}^+$

$$B^i \leftrightarrow B^{i+n} \ (\mathrm{IS}_n)$$

for every index term i.

Again, one can find a natural explanation for this in the semantics. Since $n > 0$, the notion of validity which corresponds to derivability in systems $\mathbf{HC}_n$ relies upon n–reflexive hypotheses H, which are such that $\delta_A^n(H) = H$. This means that n–distant stages in a RES starting with an n–reflexive hypothesis are identical. Hence, expansions of any given model of the base language which are obtained by n–distant hypotheses actually validate the same formulas of $\mathcal{L}^+$, as the axiom says.

One thing to notice about this axiomatization is that there is no need for axioms for index terms. This is worth emphasizing as an axiomatization of the transfinite theory along similar lines would require a fragment of ordinal arithmetic to be chosen as a base system. Index terms are indeed required to behave, arithmetically speaking, as ordinal numbers in that case (see Bruni 2009). Here, instead, the arithmetical behaviour of index terms is too basic for requiring any such a thing. In the end, this is another little point in favour of the finite theory.

Once the Hilbert–style systems have been devised, to provide the Gentzen–style counterparts $\mathbf{GC}_n$ is more or less a matter of routine. One thing is worth noticing anyway. The sequent–style rules corresponding to the definition axioms are[13]:

$$\frac{A_G(t,G)^i, \Gamma \Rightarrow \Delta}{G(t)^{i+1}, \Gamma \Rightarrow \Delta} \ (L1) \qquad \frac{\Gamma \Rightarrow \Delta, A_G(t,G)^i}{\Gamma \Rightarrow \Delta, G(t)^{i+1}} \ (R1)$$

It is immediate to notice that, in passing from the premise to the conclusion of both these rules, one cannot assume that the logical complexity of the principal formula is lower than the complexity of the active formula. This is a relevant issue as far as the aim of providing a syntactic argument of cut–elimination is concerned. The rules above behave indeed as naive abstraction rules, which are known to clash with the usage of classical logic (while they have been proved to be consistent with substructural logics instead—see, e.g., (Cantini 2003)). So, the plain fact that the cut–elimination result, as well as the corollary about the consistency of the systems are achieved, is a notable feature itself.

In addition, this is made possible by the very use of indices: the main theorem is proved by a triple induction argument, whose first parameter is specifically defined by referring to index terms. As a matter of fact, the reader can appreciate that the involved index term increases while passing from the premise to the conclusion of the rules in question.

[13] Here we use a standard sequent notation, with Γ and Δ indicating multisets of indexed formulas of $\mathcal{L}^+$.

8.4 Truth by Finite Revision

So, the approach by finite revision is a natural way of dealing with circular concepts. It also proves to have some pleasing features as far the underlying complexity, and the possibility of extracting the logic of it are concerned. What about truth then? Since truth is a circular concept, it seems natural to ask: *how much of this machinery can we expect to be able to use in the case of truth?* Very little, we must admit.

The negative part of the answer goes back to an observation which was made by Volker Halbach (1994).

Let $\mathcal{L} = \mathcal{L}_{PA}$, where the latter is the language of Peano arithmetic. Let $\mathcal{L}^+ = \mathcal{L} \cup \{T(\cdot)\}$. Fix $\mathbb{N}$ as the chosen interpretation of the base language. Accordingly, hypotheses take the form of sets of natural numbers. Let the revision operator δ_T be defined by, for every $X \subseteq \mathbb{N}$[14]

$$\delta_T(X) = \{\ulcorner B \urcorner | (\mathbb{N}, X) \models B\}$$

By taking into account the function of revision operators, δ_T is defining the extension of the truth predicate according to a disquotationalist principle: $T(\ulcorner B \urcorner)$ holds at the "revised" stage if, and only if B held at the previous one.

Then, part (ii) of Lemma 4.1 from (Halbach 1994, p. 317) shows that the following holds:

Lemma 3 *For every $X \subseteq \mathbb{N}$, for every $n \in \mathbb{N}$ such that $n > 0$, there exists a formula L_n of $\mathcal{L}^+$ such that*

$$(\mathbb{N}, X) \models L_n \Leftrightarrow (\mathbb{N}, \delta^n(X)) \not\models L_n$$

According to the terminology we have introduced in § 8.2, this entails that, over the standard model of arithmetic, *there are no m–reflexive hypotheses, for $m > 0$*. So, there is no chance of applying the semantics of finite revision. At least, if one bases himself on the simplified notion of validity from Def. 4.

The same source of information contains nonetheless another result, which is worth noticing.

Remember that the system FS by H. Friedman and M. Sheard (1987) contains the following list of axioms and rules of inference:

1. axioms PA of Peano arithmetic, where the induction schema

$$B(0) \wedge \forall x(B(x) \supset B(x+1)) \supset \forall x B(x)$$

   is allowed to be instantiated by formulas $B(x)$ of the extended language $\mathcal{L}^+$;
2. axioms for self–referential truth[15]:

[14] We assume that the formulas of $\mathcal{L}^+$ have been assigned a Gödel number according to the application of a standard arithmetization tecnique, and write $\ulcorner A \urcorner$ for the code of the formula A.

[15] We use here the symbol $\equiv$ for logical equivalence based on material implication $\supset$.

(i) $\forall x[At(x) \supset (T(x) \equiv Ver(x))]$
(ii) $\forall x[Sent_{\mathcal{L}^+}(x) \supset (T(\dot{\neg}x) \equiv \neg T(x))]$
(iii) $\forall x \forall y[Sent_{\mathcal{L}^+}(x) \wedge Sent_{\mathcal{L}^+}(y) \supset (T(x \dot{\supset} y) \equiv (T(x) \supset T(y)))]$
(iv) $\forall x \forall v[Sent_{\mathcal{L}^+}(x(\overline{0}/v)) \wedge Var(v) \supset (T(\dot{\exists} vx) \equiv \exists y T(x(\dot{y}/v)))]$
where the additional symbols here present come from the arithmetization of the syntax, and have the expected meaning: $At(x)$ means "x is (the code of) an atomic formula of $\mathcal{L}$"; $Ver(x)$ means "x is (the code of) a true atomic sentence of $\mathcal{L}$"; $Sent_{\mathcal{L}^+}(x)$ means "x is a sentence of $\mathcal{L}^+$"; $Var(x)$ means "x is a variable".

3. the rules of inference

$$\frac{A}{T(\ulcorner A \urcorner)}\ \text{NEC} \qquad \frac{T(\ulcorner A \urcorner)}{A}\ \text{CONEC}$$

Then, it seems legitimate to consider FS as *the* system of truth by finite revision. Indeed, let FS_n be the systems obtained by FS when rules NEC and CONEC are allowed to apply $(n-1)$–times at most. Then, Halbach (1994, Thrm. 4.2, p. 318) shows that:

Theorem 3 *For every $n \in \mathbb{N}$, and for every $X \subseteq \mathbb{N}$, $(\mathbb{N}, X) \models \mathsf{FS}_n$ if, and only if $X = \delta_T^n(Y)$ for some $Y \subseteq \mathbb{N}$.*

This means that, sticking with the standard model of arithmetic, every model of FS_n has the form of an expansion of it where the set interpreting the true formulas is produced by applying n–times the machinery of finite revision[16]. Furthermore, it entails that every model of this latter sort, that is every structure of the form $(\mathbb{N}, \delta_T^n(X))$ for any $X \subseteq \mathbb{N}$, is a model of FS_n. So the semantics of finite revision applied to (a disquotationalist definition of) truth, is adequately captured at the syntactic level by the layers FS_n of FS.

In fact, the second direction of the above result tells us a bit more. For, it entails that if $\mathsf{FS}_n \vdash A$ is the case (A formula of $\mathcal{L}^+$) and Y is any given subset of $\mathbb{N}$, then $(\mathbb{N}, \delta_T^n(Y)) \models A$. Since obviously $\mathsf{FS} \vdash A$ if, and only if there exists $n \in \mathbb{N}$ such that $\mathsf{FS}_n \vdash A$, it follows that every theorem of FS is 0–valid[17]. This suggests that one could use the trick of representing revision stages by index terms, in order to obtain a syntactic version of Halbach's theorem.

Let then T be the system made out of the following groups of axioms:

---

[16] Non–standard models of theories FS_n could be also provided. For instance, any model of FS is also a model of FS_n. Among other things, this was pointed out to the author by an anonymous referee. The author would like to thank the referee, whose comments helped to improve the previous version of the paper.

[17] To be more precise: let $i(A)$ be defined for every formula A of $\mathcal{L}^+$ by

$$i(A) = \min k \in \mathbb{N}.\mathsf{FS}_k \vdash A$$

Then, the previous theorem ensures that if $\mathsf{FS} \vdash A$, then $(\mathbb{N}, \delta_T^{i(A)}(Y)) \models A$ holds for every $Y \subseteq \mathbb{N}$. Hence, A is 0–valid.

1. an indexed calculus of of first–order classical logic[18];
2. for every index term i, B^i is an axiom of T, where B is:
   a) either an axiom of PA, or
   b) the induction schema instantiated by any formula of the language $\mathcal{L}^+$, or
   c) one of the axioms (i)–(iv) of FS for self–referential truth;
3. any instance of the schema $B^i \leftrightarrow T(\ulcorner B \urcorner)^{i+1}$, for every index term i and B formula of $\mathcal{L}^+$[19];

The idea is to reproduce the revision–theoretic semantics at the syntactic level, as in § 8.3.1. The additional feature here is that the revision semantics is built–up on top of a standard arithmetical one. So, at every stage in a RES, not just the laws of classical logic, but also all and the same arithmetical truths are valid. This explains point 2 in the list[20].

The truth axiom schema from point 3 is obviously devised to allow the embedding of NEC and CONEC rules. Similarly to what we were noticing in § 8.3.1, this corresponds to see NEC as representing one step forward in any RES, and CONEC being one step backwards. This must be kept in mind for the definition of the partial function $j(\cdot)$ in Theorem 4 below.

Having said that, we prove:

Theorem 4 *Set, for every formula B of $\mathcal{L}^+$, $j(B) \in \mathbb{Z}$ to be such that:*

$$\begin{cases} j(B) = 0, \textit{ if } B \textit{ is derivable in } \mathsf{FS} \textit{ with no applications} \\ \textit{of NEC and CONEC;} \\ j(B) = \min k.(k = m - m'), \textit{ if } B \textit{ is derivable in } \mathsf{FS} \textit{ with } m \textit{ applications} \\ \textit{of NEC, and } m' \textit{ applications of CONEC.} \end{cases}$$

Then:

$$\mathsf{FS} \vdash B \Rightarrow \mathsf{T} \vdash B^{\overline{j(B)}}$$

Proof The proof is by induction on $m + m'$ where m is the number of occurrences of the rule NEC in the given proof, and m' is the number of occurrences of the rule CONEC.

---

[18] Since the system FS we are trying to embed has a logic base system in axiomatic form, the required indexed version needs not to feature fully nested occurrences of the new $\rightarrow$ connective. For the sake of Theorem 4 below, the reader can think of $\rightarrow$ as occurring solely in the truth–definition axioms from point 3 of the present list. The reader should consult (Bruni 2012) for related remarks.

[19] This requires that the indexed modus ponens $\mathrm{MP}_{\rightarrow}$ from § 8.3.1 be among the logical inference rules of T.

[20] Notice that, among the axioms we have spoken of in § 8.3.1 we have *not* included in T the schema of index shift (IS). Hence, the need of assuming A^i as axiom for every formula A which is valid at every stage i. Anyway, due to the possibility that the induction schema be instantiated by formulas of $\mathcal{L}^+$, the inclusion of (IS) in the above list would not have been sufficient to obtain an equivalent, and maybe more elegant system.

$m + m' = 0$ The theorem then follows by definition of T.

$m + m' > 0$ Assume that B is obtained by an application of NEC. This means that $B = T(\ulcorner C \urcorner)$ and C is derivable with m''–many applications of NEC, and $m'' < m$. By the induction hypothesis, $\mathsf{T} \vdash C^{\overline{j(C)}}$. This means that $B^{\overline{j(C)+1}}$ is derivable in T by an application of the the definition axiom and modus ponens $\mathrm{MP}_{\rightarrow}$. Since $j(C) + 1 = j(B)$, this ends the proof.

If B is obtained by an application of the CONEC rule instead, $T(\ulcorner B \urcorner)$ is derivable in FS with m''–many applications of CONEC with $m'' < m'$. By the induction hypothesis, $\mathsf{T} \vdash T(\ulcorner B \urcorner)^{\overline{j(T(\ulcorner B \urcorner))}}$, and by the definition axiom of T and $\mathrm{MP}_{\rightarrow}$, $\mathsf{T} \vdash B^{\overline{j(T(\ulcorner B \urcorner))-1}}$. But, $j(B) = j(T(\ulcorner B \urcorner)) - 1$, hence the theorem.

If B is obtained by neither NEC, nor CONEC, though these rules have been used in the course of the proof, one simply observes that the usage of all axioms and rules in T which are different from the definition axiom, cannot increase or decrease the index of any derivable formula (see § 8.3.1). Hence, the theorem. ■

A similar result going in the other direction, i.e. from T to FS, would be possible. However, with Theorem 4 in mind, this is no surprise. Nor it would add relevant information. Hence, we have decided to leave it out of this paper, for the sake of space consideration.

8.5 Conclusion

The proposed evaluation of the finite revision semantics yielded a two–sided result. On the one hand this approach has some nice features. It is satisfactory from the point of view of the involved complexity, especially if compared to the transfinite extension of it. The semantics of finite revision is also suitable for a logical development, since it has a corresponding derivability notion which is adequate, and the syntactic approach can be dealt with in ways which are attractive for the proof–theorist. Last but not least, the machinery of revision seems to closely resemble the kind of reasoning which we would expect someone to follow, in situations where problems of circularity are involved. Hence, it is natural, as we claimed.

However, there are limits to the possibility of applying this approach. For, the theory in its nicest form requires that one is dealing with a circular concept whose defining condition respect the finitness requirement FR (see § 8.2). Luckily, we do have an interesting example of a circular concepts of this sort which is the concept of rationality in finite games. This is the reason why, however, this semantics has limited applications in the case of truth. The embeddability result we have presented is nonetheless interesting because it exploits the representation of stages in a revision path at the syntactic level via index terms. The way this can be done, in particular without any arithmetical assumption, is an additional advantage over the transfinite theory. On the basis of this result, there is an additional reason for stressing the tied connection between truth as formalised by FS, and validity under finite revision.

Can we go further? That is: can we do something similar for stability under *transfinite* revision? This is known to be one of the open problems related to this approach. Though certainly not something in the scope of the present paper.

References

Bruni, R. (2009). A note on theories for quasi–inductive definitions. *The Review of Symbolic Logic, 2,* 684–699.

Bruni, R. (2012). Analytic calculi for circular concepts by finite revision. *Studia Logica.* doi:10.1007/s11225-012-9402-2.

Bruni, R., & Sillari, G. (2011). A rational way of playing (In preparation).

Cantini, A. (2003). The undecidability of Grišhin's set theory. *Studia Logica, 74,* 345–368.

Chapuis, A. (2000). Rationality and circularity. In A. Chapuis & A. Gupta (Eds.), *Circularity, definition, and truth* (pp. 49–77). Indian Council of Philosohical Research.

Friedman, H., & Sheard, M. (1987). An axiomatic approach to self–referential truth. *Annals of Pure and Applied Logic, 33,* 1–21.

Gupta, A. (2000). On circular concepts. In A. Chapuis & A. Gupta (Eds.), *Circularity, definition, and truth* (pp. 123–153). Indian Council of Philosohical Research.

Gupta, A., & Belnap, N. (1993). *The revision theory of truth*. MIT Press.

Gupta, A., & Standefer, S. (2011). Conditionals in theories of truth (In preparation).

Halbach, V. (1994). A system of complete and consistent truth. *Notre Dame Journal of Formal Logic, 35,* 311–327.

Herzberger, H. (1982). Notes on naive semantics. *Journal of Philosophical Logic, 11,* 61–102 .

Takeuti, G. (1987). *Proof theory*. North Holland.

Welch, P. (2003). On revision operators. *Journal of Symbolic Logic, 68,* 689–711.

Part III
Truth as a Substantial Notion

Chapter 9
Truth as Composite Correspondence

Gila Sher

Abstract Is a substantive standard of truth for theories of the world by and for humans possible? What kind of standard would that be? How intricate would it be? How unified would it be? How would it work in "problematic" fields of truth like mathematics? The paper offers an answer to these questions in the form of a "composite" correspondence theory of truth. By allowing variations in the way truths in different branches of knowledge correspond to reality the theory succeeds in rendering correspondence universal, and by investigating, rather than taking as given, the structure of the correspondence relation in various fields of knowledge, it makes a substantive account of correspondence possible. In particular, the paper delineates a "composite" type of correspondence applicable to mathematics, traces its roots in views of other philosophers, and shows how it solves well-known problems in the philosophy of mathematics, due to Benacerraf and others.

9.1 The Problem

The problem that motivates me arises from a constellation of factors pulling in different, sometimes opposing directions. Simplifying, they are:

1. The complexity of the world;
2. Humans' ambitious project of theoretical knowledge of the world;
3. The severe limitations of humans' cognitive capacities;
4. The considerable intricacy of humans' cognitive capacities.

Earlier versions of this paper was presented at the *Truth at Work* Conference in Paris, 2011 and at the philosophy colloquium at UC Santa Barbara the same year. I would like to thank the audiences at both events for very constructive comments. This paper continues my earlier work on truth, knowledge, and logic, e.g., Sher (2004, 2010, 2011).

G. Sher
University of California, San Diego, USA
e-mail: gsher@ucsd.edu

T. Achourioti et al. (Eds.), *Unifying the Philosophy of Truth,* Logic, Epistemology, and the Unity of Science 36, DOI 10.1007/978-94-017-9673-6_9

Given these circumstances, the question arises whether a serious notion of truth is applicable to human theories of the world. In particular, I am interested in the questions:

a. Is a substantive standard of truth for human theories of the world possible?
b. What kind of standard would that be?
c. How intricate would it be?
d. How unified would it be?
e. How would it work in "problematic" fields of truth (e.g. mathematics, logic)?

Viewed constructively, the task is to develop a substantive theory of truth (and a standard of truth as part of it) that addresses itself to humans' desire to know and understand the world in its full complexity on the one hand and to their intricate yet limited cognitive resources on the other. Such a theory will be both critical and constructive. It will take a critical stance toward the epistemic project of discovery, understanding, and justification, yet it will contribute to, rather than interfere with, this project. And it will be both normative and descriptive. Normative in setting constraints on the pursuit of knowledge, descriptive in providing an informative account of humans' cognitive relation to the world (or certain aspects thereof). As such, the theory of truth will itself be a theory of the world, bound by the same veridical, justificatory, and pragmatic standards as other theories, and facing the same challenges.

Of course, in trying to tackle the problem of truth we are not starting from scratch. Nor do we purport to provide a complete, let alone final, solution to this problem. But the state of truth today makes our task urgent. Many philosophers have given up the goal of a substantive theory of truth, and even those who have not (like Wright 1992; Lynch 2009) have compromised the unity of truth as well as the connection between truth and reality.[1]

I should note that my focus on theoretical truth (truth of theories and statements within them) is not intended to drive a wedge between theoretical and non-theoretical truth. On the contrary. Inasmuch as I strive for a unified theory of truth, my aim is to encompass both. This, I believe, is possible because the project of theoretical knowledge is a continuation of the non-theoretical project of knowledge. The reason for focusing on the theoretical aspect of truth first is partly methodological. Philosophers have customarily started with everyday truths, yet this left them with many unsolved problems. It pays to see whether by starting with more complex truths we will not make progress in solving some of these problems.[2]

[1] The compromise concerning unity results from their willingness to assign altogether different standards of truth—e.g., correspondence vs. coherence standards—to different disciplines. The compromise concerning the connection between truth and reality arises from their willingness to assign a non-correspondence standard of truth to allegedly "problematic" disciplines. (It's important to note, however, that neither Wright nor Lynch is in principle averse to assigning a correspondence standard to any discipline).

[2] Indeed, given that we have more options in treating the simple cases than the complex cases, combinatorically it makes sense to start with the latter.

9.2 Methodological Flexibility

It's very easy to mismanage our project by rigidly following traditional templates of thought concerning truth. In particular, the "either-or" template, and its "if-then" correlate, which are very common in philosophy, make the space of options available to us unnecessarily limited. A few instances of this template that I find especially counterproductive are:

1. Either you provide a substantive *definition* of truth (a strict "if and only if (iff)" standard of truth), or you have not succeeded in developing a substantive theory (standard) of truth at all.
   There is no need to provide a definition *(an iff standard) of truth; it's fine to provide a more loosely structured account (standard) or even a family of such accounts (standards), so long as they are accurate and informative.* (Sher 2004)
2. If you are serious about a substantive theory of truth you cannot allow either circularity or infinite regress in your theory. More generally: If you are a substantivist about truth you cannot be a holist with respect to truth.
   Holism is perfectly compatible with a substantive theory of truth, and so are non-vicious circularity and infinite regress, which it sanctions. (Sher 2010)
3. If you acknowledge the plurality of truth, you cannot be a correspondence theorist with respect to *all* fields of truth.
   You can. You can say that different fields of truth are based on different (though, ideally, interconnected) correspondence principles. (Sher 2004)
4. Either you hold a copy or an isomorphism view of truth, or you are not a correspondence theorist at all.
   There can be other, possibly more intricate, correspondence conceptions of truth. In this paper I highlight "composite correspondence".
5. If you are a correspondence theorist of truth, you must allow only *one* pattern of correspondence.
   You can allow a variety of correspondence patterns, though you may wish to connect (unify) them in significant ways.
6. If you are a correspondence theorist with respect to logic and mathematics, you must be either a Platonist or an empiricist with respect to them.
   There is no need to postulate a separate, abstract, reality in order to affirm that objects and properties in the world (even physical objects and their physical properties) have formal features (like self-identify and cardinality) and that these (worldly) features, or the laws governing them, underlie logical and mathematical truth. But there is also no need to regard these features (laws) as empirical. It's an open question how mathematical truths are connected to formal features of reality. It's sufficient that they are systematically connected to them in some way that explains their correctness, not necessarily the traditional way that mandates the existence of mathematical individuals, for example.
7. Either you regard knowledge of logical and mathematical truths as apriori or you do not regard it as obtained primarily through use of reason.

It's possible that logical and mathematical knowledge is obtained largely *through reason without being* completely *isolated from (independent of) experience, i.e., without being apriori.*

8. Either you regard logic as grounded in reality, or you regard it as grounded exclusively in the mind.
   You can regard logic as grounded both in reality and in the mind (as you should all other veridical disciplines). (Sher 2011)

With this advance notice of my deviations from common practice, I am almost ready to proceed to my solution of the problem. First, however, I need to clarify an important methodological issue: *holism*. I have touched on holism in (2) above, but holism plays such a central role in my proposed solution to the problem that I need to clarify what exactly I mean by holism, and what my motivation for holism is.

Holism is my chosen methodology for philosophical theorizing. It contrasts most sharply with foundationalism. I don't reject all aspects of foundationalism: I think the search for the (or a) foundation of knowledge, truth, logic, morality, etc., is a worthwhile and legitimate task of philosophy. It's a search I am engaged in this paper (with respect to truth). But I think foundationalism goes about the foundational project in the wrong way. Foundationalism sets a strict ordering requirement on the grounding relation, one that renders this methodology self-defeating. Foundationalism requires that in grounding X we limit ourselves to resources more basic than X, and eventually to resources limited to a certain fixed trove—the so-called *base*. The problem is that it's impossible in principle to ground—or even to seriously justify—the base itself. We arrive at our holistic methodology by relaxing the strict ordering requirement of the foundationalist methodology.

Now, it's common to think of the renouncement of foundationalism as the renouncement of the foundational project itself. I think this is a mistake. The foundational project, as a dual, justificatory and explanatory project, can be carried out in multiple ways. By removing the unreasonable strictures of one methodology, foundationalism, holism provides us with a potentially unlimited number of ways to pursue it. Elsewhere (Sher 2010) I called this conception of holism as a foundational methodology "foundational holism". Among its advantages in pursuing the foundational project are:

- We need not worry about circularity or infinite regress per se. (Not all circularity and infinite regress are vicious.) ((2) above).
- We can use any resources that are helpful to us, in any combination (spanning any number of theories or fields of knowledge), to solve any problem.
- We can take multiple routes to reality—both routes of discovery and routes of justification (explanation)—including composite and/or circuitous routes.
- The justification/explanation process is a step by step process. Partial justification/explanation is worthwhile and superior to no justification/explanation.

This holistic method is not just an appropriate method for philosophical inquiry, but, in light of the four circumstances noted in the opening paragraph of this paper, also an appropriate method for humans to build their "system of knowledge" of the world.

As such, it is nicely captured by the Neurath-boat metaphor, though with one proviso. The boat, as we envision it, is not floating aimlessly in the water and is not isolated from the world. On the contrary, its sailors are engaged in exploration of the world (the sea and its environs), and their survival and success depend on taking the world into account and, indeed, focusing on the world. The Neurathian boat is sometimes associated with coherentism and rejection of the correspondence theory of truth. On our view, the boat pursues a discovery, understanding, and justification project, and its standard of truth is a correspondence standard: a standard of measuring up to (marine) facts. This leads to our proposed solution to the problem of truth.

9.3 Composite Correspondence

If the world is highly complex relative to our cognitive capacities and we nevertheless seek to know it in its full complexity, this requires stretching our cognitive endowments, devising multiple means for reaching its less accessible regions, improvising, experimenting, tinkering, exercising our imagination, etc. In short, it's quite likely that we use a wide variety of routes to reach the world cognitively, and some of these are complex, indirect, jagged. What this means for the study of truth is that it's seriously possible that there are multiple routes of correspondence between true cognitions and reality, and that some of these routes are quite intricate. This possibility concerns language as well. A proper use of, say, a *singular term* need not target an *individual* in the world. It may target something else (say, a property of individuals, a property of properties, etc.) that for one reason or another we reach through the use of a singular term and, possibly, an intermediate, posited, individual that is systematically connected to it.

Given this possibility, it's an open question what form correspondence takes in different areas of knowledge, and it's unreasonable either to assume or to require that it take the form of a copy, or a mirror-image, or even an isomorphism. At the same time it's also the case that in developing theories of the world humans aim at unity and systematicity, and that in seeking to develop a normative-descriptive theory of truth we, too, aim at these things. Put together, this means that the project of constructing a substantive theory of truth is not simple. It requires a balance between unity and diversity, between observing and proposing, between describing and constructing, between being critical and understanding. I will call a theory of truth that requires a substantial correspondence (of one kind or another) between true cognition and reality, allows multiple—including intricate—routes of correspondence from language to reality, yet seeks maximal unity and systematicity, a "composite correspondence" theory.

In attempting to develop a composite-correspondence theory of truth I use the holistic, Neurathian methodology delineated above. While keeping in mind a few milestones of the philosophy of truth (e.g., Tarski's T-Schema, his recursive definition of truth in terms of satisfaction, his semantic definition of logical truth), I will begin, in accordance with my policy of going from the complicated to the simple, with a

hitherto problematic field of truth, specifically, mathematics. Before attempting to deal with this field, however, I would like to briefly examine another conception of correspondence, one that is similar to mine in certain important respects yet does not extend to mathematics.

9.4 Horgan's Indirect Correspondence

Lynch (2001) summarizes Horgan's conception of correspondence as follows:

> In Horgan's view... truth in any discourse is determined jointly by the world and the semantic standards of the discourse. In short, truth is semantic correctness. Semantic correctness is a realist notion of truth, since it involves a type of correspondence with the world.... Nonetheless, the type of correspondence can vary according to what we are talking about. This is because the semantic norms governing truth vary with the context. Thus, there is a spectrum of ways in which statements can correspond to the world. On one end of the spectrum are statements governed by maximally strict semantic standards. Such statements are true just when they directly correspond, via causal/referential relations, to mind-independent objects and properties. On the other end are statements whose truth is determined almost entirely by the semantic standards alone. In between sits the majority of statements we make in life, such as those about corporations and works of art, which **indirectly** correspond to entities and attributes that are in many cases mind-dependent. [*Ibid.*, p. 13. My bolding]

Horgan himself describes his theory as a combination of metaphysical realism and a liberal approach to the correspondence theory of truth. Metaphysical realism, for him, is essentially the view that the world consists of a definite totality of discourse-independent objects and properties. Correspondence is the view that truth depends, at least in part, on the way the world is, and liberal correspondence is the view that this dependence need not involve a direct, 1-1 correlation between singular terms and objects in the world, predicates and properties in the world, quantifiers and sets of objects in the world (those over which they range), etc. The basic idea is that truth depends both on the world, independently of the mind, and on the mind in the sense of the totality of our semantic standards. Those standards vary from one context to another, and as a result there is a whole spectrum of ways in which, and degrees to which, a statement's truth value can depend upon the world. At one end of the spectrum are statements and contexts for which the semantic standards require a *direct* connection between true statements and reality, at the other end statements and contexts for which the semantic standards require no more than empty connection with reality, and the intermediate zone contains contexts and statements for which the semantic standards require indirect connections with reality of various types, levels, and degrees. Statements lying in the first region are subject to the strictest semantic standards: their truth requires a direct correspondence between their referential apparatus (singular terms, predicates, and quantifiers) and individuals, properties, ontologies (ranges of quantifiers) in the world. Statements lying in the second region are subject to the weakest semantic standards: their truth is altogether independent of the world; i.e., they exhibit null correspondence. Statements lying in the third, intermediate, region are subject to intermediate semantic standards:

their semantic correctness significantly depends on how things are with the world, but this dependence is *indirect*—does not involve a direct correlation between their referential apparatus and individuals, properties, or ontologies. Their truth (when they are true) is based on *indirect correspondence*.

Horgan's conception of correspondence, however, is empiricist. His empiricism manifests itself on three levels: (a) The ultimate ontology of the world is empirical: there are no abstract objects or properties. (b) The relevant relations between language and the world, as far as truth is concerned, are empirical: specifically, causal. (c) Disciplines whose statements do not obtain their truth-value from causal connections to empirical reality are viewed as governed by a *non-correspondence* standard of truth. For example, pure mathematical truths as well as moral truths are viewed as based on a non-correspondence standard of truth.

To see an example of a truth based on indirect correspondence, according to Horgan, consider the sentence

1. Beethoven's fifth symphony has four movements. (Horgan 2001, p. 73)

The truth of (1), Horgan points out, "does not require that there be some ENTITY answering to the term 'Beethoven's fifth symphony' and also answering to the predicate 'has four movements'" (*ibid.*). Rather, it's sufficient that there are "other, more indirect, connections between [(1)] and THE WORLD" (*ibid.*), connections with, in particular, Beethoven's and others' musical behavior: Beethoven's "composing his fifth symphony" (*ibid.*), his earlier compositional activities ("in virtue of which his later behavior counts as composing his *fifth* symphony" (*ibid.*)), "a broad range of human practices (including the use of handwritten or printed scores to guide orchestral performances) in virtue of which such behavior by Beethoven counts as 'composing a symphony' in the first place" (*ibid.*), and so on.

One consequence of accepting indirect correspondence is expanding the domain of correspondence. Even if you are a strict nominalist or empiricist who denies the existence of abstract objects like symphonies, you can acknowledge that a sentence like (1), which talks of abstract objects and properties (symphonies, movements of symphonies, the property of being fourth in a series, the property of having four parts), have a definite truth value, that their truth value is to a significant degree a matter of fact—a matter of how things are in the world. Another consequence is that indirect connections with reality are as significant for truth as direct connections.

As for (1) itself, it is linked to reality both directly, through a causal link to a bona fide physical object—Beethoven, and indirectly, through causal chains ending with the abstract terms "fifth symphony" and "four movements". But (1) is not linked to reality through any independent connections of "fifth" and "four" with reality, direct or indirect. Pure ordinals and cardinals are not connected to reality, either directly or indirectly. Instead they are connected to pure conventions, i.e., to the mind. While the abstract terms "symphony" and "movement" are anchored in reality on Horgan's conception, "five" and "four" are not. It's here that the empiricist bias of his correspondence theory comes into expression. His theory allows language to be connected to reality in two ways: through direct causal links to explicitly-mentioned empirically-acceptable objects and properties, and through indirect causal links to

mentioned and unmentioned empirically-acceptable objects, properties, events, etc. In both cases correspondence is limited to *empirical* reality and its routes are limited to *causal* routes.

One virtue of Horgan's conception is its ability to expand the range of correspondence acceptable to empiricists. Horgan shows that even if you are a strict empiricist or a nominalist who denies the existence of abstract objects like symphonies, you can acknowledge that a sentence like (1), whose subject matter is this abstract object, is true or false in the correspondence sense. A significant limitation of his approach (from our perspective) is that it leaves out of the correspondence account important parts of our system of knowledge, for example, logic and mathematics. On our own approach, *all* truths—including logical and mathematical truths—are based on correspondence.[3]

9.5 Mathematical Correspondence

The difficulties encountered by existent theories of mathematical truth—for example, difficulties concerning the identity of, and the access to, mathematical objects (Benacerraf 1965, 1973)—have led many philosophers to doubt that mathematical theories are true in the correspondence sense at all. In what follows I will present a tentative account of mathematical truth that avoids many of the extant objections to mathematical correspondence and is capable of doing the explanatory work that we would like a theory of mathematical truth to do.

In presenting this account, I will take a non-traditional approach. Discussions of truth and correspondence usually center on relatively small units of cognition, single sentences or units of a similar size. It seems to me that in investigating whether mathematical truth is based on correspondence we need to look at the discipline of mathematics as a whole. The main questions are whether mathematics aims (or large parts of mathematics, or some important parts of mathematics, aim) at *discovering facts*, what kind of facts these are, and whether such facts are in principle discoverable by humans. Before we can answer these questions, however, we must ask: Is there something in reality for mathematics to be true or false of? Does reality have mathematical features? The last question is rarely asked in contemporary discussions of truth, but I think it ought to be. Whether reality has features whose study requires a mathematical discipline, and whether the study of such features is an integral part of our body of theories of reality, is crucial for figuring out whether there is a need for correspondence truth in mathematics, and whether mathematical theorems are connected to reality. A positive answer to this question will not establish the existence

[3] It should be noted, though, that Horgan approaches truth from a somewhat different direction than we do. In particular, Horgan is interested in solving the problem of vagueness while we are interested in solving the epistemic problem of truth. These problems are not completely disconnected, but the difference is in certain ways insignificant.

of mathematical correspondence, but it will show that it's reasonable to investigate the possibility of correspondence in mathematics just as it is in physics.

My answer to the question "Does reality have mathematical, or mathematical-like, features?" is positive. You don't need to be a Platonist to realize that individuals (including physical individuals) have properties like self-identity, that properties of individuals (including physical properties) have properties like cardinality, that relations of individuals (including physical relations) have properties like reflexivity, symmetry, or transitivity, that cardinality properties (like the 2nd-level properties ZERO, ONE, TWO, THREE . . .) stand in an ordering relation, and so on. A natural name for such properties (relations) is "*formal properties (relations)*". It's hard to deny that objects in the world have some (indeed, many) *formal* properties. This is our starting point. Our next step is to try to understand what is unique about the formal.

One way to characterize the formal is as *invariant under isomorphisms.* An extended explanation is given in, e.g., Sher (1991, 2008). But briefly, we can see what this means by considering an example of a formal property, say, a cardinality property, $\mathbb{C}$. $\mathbb{C}$ is invariant under isomorphisms because given any isomorphic structures <U,P> and <U',P'>, where U,U' are universes[4] and P,P' are 1st-level properties, P has the property $\mathbb{C}$, (in U) iff P' has the property $\mathbb{C}$, (in U'). In contrast, the 2nd-level property "P is a property of some humans" is not invariant under all isomorphisms, hence it's not a formal property.

Now formal properties, like physical properties, are presumably governed by laws, or systematic regularities. (It's quite clear that cardinalities, for example, are governed by laws.) It stands to reason that humans are interested in discovering these laws, so the question arises: Which discipline studies these laws?—It's reasonable to presume that *mathematics* does, because mathematics does study things like cardinalities, reflexivity, symmetry, transitivity, etc, and it would be very strange if mathematicians knew there were "real" cardinalities and thought they were governed by laws yet studied the laws of other, non-real, imaginary, cardinalities. But if mathematics studies the laws governing formal objects in the world, the appropriate standard of truth for mathematical theories would be a *correspondence standard* of some kind, i.e., a standard that requires a systematic connection between true mathematical statements and laws governing the formal behavior (features) of objects in the world.

Here, however, we seem to encounter a problem: Many formal laws that govern objects in the world are higher-level laws, but the mathematical theories that study them are lower-order theories. For example, the laws of cardinality are laws of 2nd-level properties, but standard arithmetic and set-theory study them as laws of individuals (0-level objects). How can mathematical statements about individuals be true if what they say is true due to facts about higher-order properties? When put this way, the question appears difficult to answer. But if we put it in a different

[4] A universe is a non-empty set of individuals. The structures <U, P> and <U',P'> are isomorphic iff there is a 1-1 and onto function *f* from U to U' such that P' is the image of P in U' under *f*.

way, say, "How can laws of individual numbers *correspond* to laws of 2nd-level cardinality properties?", it's fairly easy to answer. In the same way that large objects can be represented by small objects (e.g., in a small-scale model), so higher-level objects can be represented by lower-level objects. 1st-order theories are capable, in principle, of correctly describing 2nd-level objects and their laws, *if we allow indirect representation* of these laws and their objects. 1st-order statements about individual numbers can correspond to facts, laws, phenomena involving 2nd-level properties, *if we allow indirect or composite correspondence*.

Why would humans use composite correspondence in studying these laws rather than simple correspondence? Why not opt for higher-order mathematics? This depends on the circumstances. One possible reason for preferring 1st-order theories might be that we, humans, are so wired that we are better at figuring out the formal laws governing structures of objects when we treat these structures as lower-level structures, say, structures of individuals. In that case, it would be advantageous to us to study higher-level cardinalities by 1st-order theories. There could, of course, be other reasons. The important thing is that this option is open to us.

How would humans go about constructing 1st-order theories of higher-level cardinality properties? One way they could go about it is by introducing a *postulated* level of *individual cardinals* which are systematically connected to *2nd-level* cardinality properties (*in the world*). Mathematical truth and reference would then be composite: Numerical singular terms would refer_1 to posited individual numbers and refer_2 to 2nd-level cardinality properties which are systematically represented by their referents_1. Similarly, 1st-order statements expressing laws of cardinal individuals would correspond_1 to posited 1st-level cardinality laws and correspond_2 to higher-level cardinality laws which their correspondents_1 systematically represent. The reference and correspondence relevant to a theory of mathematical truth are reference_{1+2} and $\text{correspondence}_{1+2}$. Put in terms of standards: an appropriate standard of truth for mathematics might be a standard of composite correspondence, like our $\text{correspondence}_{1+2}$. In short, an appropriate standard of truth for mathematics (or for any other discipline) need not be simple or direct, but it must measure mathematical theories and their statements against that facet of reality which is their target, and it must require mathematical theories to systematically represent this facet.

Systematic connection, however, does not mean translatability to an *equivalent* theory. Although we can translate 1st-order number statements to higher-order number statements, 1st-order arithmetic is *not equivalent* to 2nd-order arithmetic: 1st-order arithmetic has a logically complete proof system (the proof system of standard 1st-order logic) while 2nd-order arithmetic does not; 2nd-order arithmetic is categorical while 1st-order arithmetic isn't. The two are not equivalent, yet they are systematically connected. But even this kind of connection is not mandatory. A systematic connection between 1st-level posits and higher-level phenomena does not require the existence of a worked-out higher-order theory. It's acceptable if for some higher-order formal phenomena there is only a 1st-order theory that describes them, so long as this theory is systematically connected to them. Systematic connection can

take multiple forms, and it allows indirect and incremental demonstration, especially for holists.

This is our idea of composite correspondence: n-step correspondence ($n \geq 1$), possibly involving auxiliary posits. Composite correspondence does not give up the desideratum of simplicity, either on the normative or on the descriptive level (or, indeed, as a principle that might be at work in the world). But it regards it as secondary to the more important desiderata of setting up a realistic standard of correspondence and correctly accounting for the way human beings do, or may, cognitively reach the world, both on the level of theory and on the level of common-sense thought. Although our theory is still in its initial stages, we can already point out some of its methodological virtues. Three of these are: strong problem-solving capabilities, interesting connections with related philosophical theories, and potentially fruitful extensions.

9.6 Problem-Solving Capabilities

The composite account of mathematical correspondence, together with the underlying holistic approach to truth and knowledge, enable us to solve, or at least make some significant progress toward solving, a few widely discussed problems for mathematical correspondence. A full discussion of these problems would have to be quite lengthy, but briefly we can indicate the problems and solutions (or their direction) as follows:

a. *The Identity Problem* (Benacerraf 1965). This problem concerns the identity of mathematical individuals. For example: Is Zermelo's 2—namely, {{Ø}}—or von Neumann's 2—namely, {Ø, {Ø}}—the real 2? Our proposal easily dissolves this problem: If numbers (as individuals) are posited representations of real cardinality properties, there is no question of which individual is the *real* 2. Systems of posits are measured by their fruitfulness, systematicity, and representational success, not by their reality. Since Zermelo's and von Neumann's systems are equally fruitful, systematic, and representationally successful, it doesn't matter whether we choose one or the other.
b. *The Applicability Problem.* (Wigner 1960) The problem is to explain how abstract mathematical laws apply to empirical objects (events, phenomena, situations). The problem is especially difficult for traditional correspondence theorists of mathematics, who are platonists. It's hard to see how mathematical laws which govern a Platonic reality, completely separated from our mundane physical reality, can apply to physical objects. But this problem does not arise for our proposal. First, our proposal, being holistic, assumes one, interconnected reality, which has both physical and formal features. Second, formal features themselves, on our proposal, are in principle features of objects of various kinds, including physical objects. For example, cardinality properties are (direct) properties of physical properties of (physical) objects. The applicability of mathematical laws in the physical domain is, therefore, unproblematic.

c. *The Large Ontology Problem.*[5] The problem is to justify mathematics' claim to a super-large ontology of individuals (e.g., cardinals and ordinals). This problem is dissolved by our proposal: mathematical individuals are intermediary posits rather than real objects; as such their number is subject to standards of fruitfulness, not of reality or existence. Indeed, our proposal does more than just dissolve this "problem": it has the capacity to explain why mathematics needs a super-large collection of posited individuals. Mathematics, as a theory of the formal, is a theory of laws governing features of objects. But laws in general have a certain degree of necessity and as such require a counterfactual ontology. Formal laws have an especially strong degree of invariance; as such they require an especially large counterfactual ontology.[6]

d. *The Epistemic Access Problem.* (Benacerraf 1973) The epistemic access problem is the problem of how humans can cognitively access those aspects of reality that are relevant for mathematical truth and knowledge. This problem is especially difficult for those who regard mathematical truth as requiring a *real* ontology of mathematical individuals. It's harder to explain how we can *see* the number *3* than how we can see that there are *three* books on the table. Our account, however, does not require mathematical individuals: there is no need for humans to see the number 3 or any other numerical, set-theoretical, or geometrical individual. The problem of epistemic access is also extremely difficult for Platonists. It's difficult to explain how creatures who exist in one reality can access another reality. This problem, too, does not arise for us. Reality, for us, is one; it has both physical and formal features; there is no need to cross realities in order to access the formal. Finally, the problem of epistemic access is very difficult for empiricists. It's hard for empiricists to account for abstract properties, let alone for their laws. Most importantly, it's hard for empiricists to acknowledge the central role intellect (reason) plays in mathematical knowledge. None of these problems arise for us. Our approach to mathematical knowledge is neither empiricist nor apriorist. As holists we sanction multiple resources for reaching reality that are not available to the empiricist, including multiple configurations of intellect and sense perception that can be used to access the formal.

e. *The Mathematics-as-Algebra Problem.*[7] The problem is that some mathematical theories are not naturally viewed as theories of any particular formal feature (or family of features) of reality, but seem to engage with abstract structures that might have a variety of applications or no applications at all. Put otherwise, these "algebraic" theories create free-floating models: models that might represent multiple real structures (or none). This problem, too, is relatively easy for us to handle. First, we see no radical difference between a theory and a model. Theories

[5] This problem is raised by nominalists of various stripes (e.g., Goodman and Quine 1947), mathematical finitists, supporters of V = L, and others who feel uncomfortable with the huge ontology of contemporary (classical) set-theory. Here I focus on the issue of *size.*

[6] For further discussion of the philosophical ramifications of invariance, see Sher (1999a, 2008).

[7] The issue of algebraic vs. non-algebraic mathematical theories is discussed in, e.g., Shapiro (1997).

can involve the postulation of entities, which is what models do; and models, like theories, can account for a given set of phenomena in the world either accurately or inaccurately, either systematically or not systematically. Indeed, our dynamic holism sanctions fluctuations in the status of theories as being about the world or being mere algebras. In the course of history a theory like Euclidean geometry can turn into an "algebra", while an "algebra" like Riemannian geometry can turn into a theory of the world. When it comes to truth, the difference is between being *true simpliciter* and being *true of*; on our account there is no radical difference between the two.

9.7 Relation to Fictionalism and Other Philosophical Views

Our conception of mathematical truth as based on composite correspondence has points of contact with a number of other philosophical conceptions of mathematics, from Aristotle's to contemporary fictionalists'. I will not be able to trace all these contacts here (or, indeed, any one of them with great detail), but I will briefly discuss a few.

Aristotle Our treatment of mathematical truth bears some similarities to Aristotle's, especially as construed by Lear (1982). Reality, according to Aristotle, has multiple aspects, and this is reflected in the variety of features possessed by physical objects. In particular, physical objects have mathematical features (like being spherical), and it's these features that are the subject-matter of mathematics. Mathematics, therefore, studies *real* features of *real* objects, and its theories are genuinely true (or false). How does mathematics study these features? By separating them (in thought) both from the physical objects that possess them and from the other features these objects possess, and by studying them on their own. Thus, Aristotle says:

> Obviously physical bodies contain surfaces, volumes, lines, and points, and these are the subject matter of mathematics. . . . [T]he mathematician . . . separates them, for in thought they are separable from motion, and it makes no difference nor does any falsity result if they are separated. [Aristotle, *Physics*, Book II, Chapter 2: 193b23–35. Cited in Lear: 162]

This effective way of studying mathematical features might involve the positing of mathematical entities:

> The best way of studying [such features] would be this: to separate and posit what is not separate [i.e., what is not separate in reality but is separable in thought], as the arithmetician does and the geometer. [*Ibid. Metaphysics*, Book M, Chapter 3: 1078a21–3. Cited in Lear: 165. My square brackets.]

Under this conception, mathematical objects exist, but they exist in a special way: namely, as abstractions from physical objects. This relation between mathematics and physics explains how mathematics applies to physical objects. Lear concludes:

> For Aristotle, mathematics is true, not in virtue of the existence of separated mathematical objects to which its terms refer, but because it accurately describes the structural properties

> and relations which actual physical objects do have. ... [There is no need] to explain mathematical truth ... via the existence of mathematical objects. One can understand how mathematics can be true ... by understanding how it is applicable. (Lear 1982, p. 191)[8]

Our account is similar to Aristotle's in several significant ways: Both accounts start from the observation that objects in the world have formal/mathematical features and that one central task of mathematics is to study these features; both sanction the postulation of mathematical objects that represent such features; both regard mathematical truth as genuine truth; both regard mathematical truth as based on some kind of circuitous correspondence (Aristotle implicitly; we explicitly); both allow cognitive access to formal/mathematical features of reality; and both have resources for explaining the application of mathematics to physics in a relatively straightforward manner.

But there are also significant differences between the two accounts. Our own account is not committed either to Aristotle's special ontological theory of real, non-posited, objects, or to his views on the variety of formal features of objects in the world. Our account explicitly talks about laws governing the behavior of formal features of reality, and it sanctions laws that, I gather, are farther reaching than those Aristotle had in mind (for example, on our conception it's not unreasonable to expect that something like full-scale set theory is required for a comprehensive account of some formal laws). Our account has resources (e.g., Fregean resources), that were not available to Aristotle, for dealing with arithmetic truth, as well as resources (like invariance) for explaining the necessity of mathematical laws. Another difference is methodological: if, and to the extent that, Aristotle's methodology is foundationalist, our own, holistic, methodology provides us with a wider variety of tools for grounding mathematical theories in reality and for explaining mathematical truth and knowledge.[9]

Frege Our conception of cardinalities as 2nd-level properties, our focus on systematic connections between cardinality properties and numbers as individuals (0-level objects), our view of mathematics as objective, our emphasis on a close relation between mathematics and logic (see next section)—all have at least some roots in Frege. But there are also important differences between our approach and Frege's.

Frege gives more weight to language as a key to ontology than we do. For him, humans' use of singular terms to indicate numbers is a sign that numbers are individuals. (See, e.g., Frege 1884). For us, the relation between language and ontology is far less binding and direct. Natural language is a multi-dimensional tool, developed in a messy and often haphazard way, and there is no reason to assume a simple 1-1 correlation between its terms and things in the world. Theoretical language is less messy and haphazard, but it, too, cannot be taken as a guide for ontology. First,

[8] This account (and explanation) applies smoothly to geometry, but Aristotle uses a somewhat different (and arguably weaker) account for arithmetic. For us, post-Fregean philosophers, however, it's natural to extend Aristotle's account of geometry to arithmetic, by viewing numbers as representing cardinality properties of physical objects.

[9] For a discussion of this point, though not as it relates to Aristotle, see Sher (2010).

theoretical language is influenced by many things: natural language, earlier states of our knowledge, etc., and as such might not accurately reflect our current views of ontology. Second, theoretical language is bound by humans' limitations, and as such reflects the indirect and at times improvised routes that humans forge in an attempt to reach reality. In addition, theoretical language has multiple goals, including simplicity, efficiency, unity, and systematicity, and these might stand in some tension with the goal of ontological transparency. Finally, humans have the cognitive capacity to posit objects of various kind, including "fictional objects" systematically related to non-fictional objects, and there is no reason to think they don't exercise this capacity in developing theoretical (and other) languages. Indeed, not only do they have the capacity to posit new objects, they may very well have good reasons to do so, as we explained above. If, as we suggested above, (i) reality has formal features, (ii) these features are governed by laws, (iii) humans have difficulty in figuring out these laws directly, (iv) humans can figure out better, or more easily, what these laws are if they approach the task indirectly, using posited objects—then it's quite reasonable for them to develop languages that refer to these posited objects. This is something humans might do either instinctively and unconsciously or deliberately and in a planned manner, and therefore it might be reflected both in their natural language and in their theoretical languages. For that reason, we allow *posited* individual numbers where Frege requires *real* numbers

Another difference from Frege concerns the status of mathematical truths. While Frege characterizes mathematical truths as analytic and apriori, we, as holists, renounce the analytic-synthetic and apriori-aposteriori distinctions, thereby rejecting Frege's characterization.[10] Still another difference concerns logic. Although we share Frege's view that mathematics and logic are closely related, we differ on the precise nature and structure of this relation.[11] Finally, we reject Frege's claim (in, e.g., Frege 1918) that truth is primitive and unexplainable. Our account of truth as composite correspondence is based on a substantivist, non-primitivist approach to truth.[12]

Quine The idea that positing objects is central to human knowledge is a Quinean idea. Not just mathematical objects but also abstract physical objects, everyday common-sense objects, and even subconceptual sense data are posits, according to Quine—theoretical, conceptual, and evidential, respectively. What is the point of positing such diverse types of object?—For Quine, the motivation is pragmatic. In the case of molecular particles, for example, "the particles are posited for the sake of a simple physics" (Quine 1955, p. 250). Mathematical objects are posited to expedite scientific knowledge in general, and they are justified by their *indispensability* to (empirical) science. What can we conclude from the fact that objects of all types are posited?—"[T]hat posits are not *ipso facto* unreal" (*Ibid.*, p. 251). In fact, posits are essential for our very notion of reality: "it is by reference to [everyday bodies, which

[10] This is a significant issue. For discussion see Sher (1999b, 2010).
[11] This will be briefly discussed in the next section. See references there.
[12] This approach is argued for and explained in Sher (1998–9, 2004).

are posits] that the very notions of reality and evidence are acquired" (*Ibid.*, p. 252). In this way, the boundary between the real and the posited disappears.

Our own approach is similar to Quine's in its emphasis on the basic (and positive) role of posits in knowledge, but is different on other counts. First, we are not committed to the view that all objects are posits. Second, we think of posits as playing a broader (and more fundamental) role in knowledge than a purely pragmatic role. This is because posits, in our view, enable us not just to simplify our theories but also to do more fundamental things like *figure out* certain aspects of reality that we might not be able to figure out without them. Finally, while Quine regards mathematics as subsidiary to empirical science, we see it as standing on its own. Mathematics' primary role is to provide knowledge of the laws governing formal features of reality, and its role in physics, though very important, is secondary to this role.

Contemporary Fictionalists Contemporary fictionalists, like Field (1989), claim that theories need not refer to *real* objects in order to be "good". Theories whose ontology is *fictional* can make a significant contribution to knowledge and have the advantage of being immune to difficulties facing their *realistic* counterparts. Field is especially interested in fictional theories of a particular type: *mathematical* theories which are *conservative* with respect to (physical) *science*. A mathematical theory M is conservative with respect to science iff for any nominalistic scientific theory N and a nominalistic sentence S, M+N logically implies S only if N (by itself) logically implies S (*ibid.*, p. 58). The idea is that mathematics plays a purely *pragmatic* or *instrumental* role with respect to science, i.e., it does not contribute anything substantial to science. Science can dispense with mathematics altogether, at least in principle. This, however, does not mean that we can eliminate mathematics through direct translation to real-object language (*ibid.*, p. 7). Nor does it mean that there is a problem with accepting those parts of mathematics that have no application to science, like the higher reaches of set theory. These, Field says, are natural ways of extending the "story" told by the useful parts of mathematics (*ibid.*, p. 10). Field is flexible with respect to justifying his account. Justification, he emphasizes, is not an all-or-nothing thing (*ibid.*, p. 17). Among the methodological principles he appeals to is "inference to the best explanation" (*ibid.*, pp. 14–20).

Our view is similar to Field's in the central role it assigns to posits in knowledge, in its acknowledgment of the creativity of human cognition, in its acceptance of the non-applicable parts of mathematics as legitimate, in its appreciation of such methodological principles as inference to the best explanation, and in its rejection of the view that justification is an all-or-nothing affair. But we differ from Field in other ways, including his nominalistic outlook on mathematics, his lack of interest in mathematics as a branch of knowledge in its own right, and his approach to logic.

Field's basic approach to knowledge is physicalistic, and although he doesn't ban all abstract objects (e.g., he endorses space-time points), he is a strict nominalist when it comes to mathematics. We, in contrast, leave it an open question whether there are mathematical individuals, and, in the absence of compelling arguments to the contrary, we are committed to the reality of formal properties and formal laws. Unlike Field, we do not measure the epistemic value of mathematics just by its contribution to

(or instrumental value for) physical knowledge, but view it as an independent source of knowledge. As a result we, unlike Field, view mathematical laws as genuinely true, true in the sense of correspondence with reality, albeit composite correspondence. Finally, our view differs from Field's with respect to logic. Field's instrumentalist fictionalism puts a heavy burden on logic: on the one hand, mathematical fictions help the scientist to derive scientific truths from scientific truths using logic; on the other hand, the dispensability of mathematics is due to the fact that all nominalistic results arrived at using mathematics can be derived from other nominalistic results using only logic. Logic itself Field regards as "real", in contrast to mathematics. But logic, for him, is not real in the sense of being anchored in reality: logic is disengaged from reality. In this respect his view of logic is more traditional than ours. While we share Field's view that logic is real, we regard it as real in the sense of being anchored in reality, as will be made clear below.

Field's fictionalism is subject to several objections, most of which do not apply to us. One potential problem for fictionalism, however, discussed by Yablo (2001),[13] might be thought to apply to us as well. This problem can be formulated as follows: on the one hand the fictionalist is ready to assert that there is an even prime number, hence, that there are numbers; on the other hand, the fictionalist does not believe in the reality of numbers, hence is committed to asserting that there are no numbers. So the fictionalist is both committed to the view that there are numbers and committed to the view that there are no numbers. (Another way to put it is to say that the fictionalist is committed to the self-defeating sentence "The number of numbers is 0" (*ibid.*, p. 80).) Yablo dissolves the problem by showing it results from overlooking a simple distinction. This distinction he successively formulates as "representational aids" vs. "things-represented" (*ibid.*, p. 81), "engaged" vs. "disengaged" modes of speech (*ibid.*, p. 83)[14], "basic" vs. "parasitic" language-games (*ibid.*), "objectual" vs. "assertional" reality (*ibid.*, p. 85), and "figurative" vs. "non-figurative" speech (*ibid.*). The crucial point is that number words "can travel back and forth between the two categories"; indeed "they can do it ... within a single sentential move" (*ibid.*, p. 81).

Something like Yablo's solution is available to us as well. When we say that there are individual numbers we refer directly to the posits and indirectly to reality (cardinality properties); when we say that there are no individual numbers we refer directly to reality, and not at all to posits. (When we say that the number of numerical individuals is 0 we use "numerical individual" as a term ranging over real individuals and "0" as a term denoting a posit, 0, which represents a 2nd-level property, ZERO.) So there is no contradiction. It is possible, of course, that someone else speaks in a way that leads to a contradiction, but our holism provides us with resources for distinguishing these two ways of speaking.[15] We can move to a higher standpoint

[13] An earlier version, directed as *modal* fictionalism, is due to Rosen (1990).

[14] Our speech is engaged when we speak from within a given game, disengaged when we speak from outside it.

[15] For further discussion of our holistic, Neurathian conception of knowledge see Sher (1999b, 2010).

on Neurath's boat, or to a higher language in Tarski's hierarchy of languages, a standpoint or a language from which we can see both the world and the ways different people use language (object language) to speak about it. Some of these ways lead to a contradiction; others don't. Our claim is that on reflection, mathematics' contribution to knowledge is captured by the use of words spelled out above.

9.8 Truth in Logic and Beyond

Our conception of truth as based on composite correspondence is expandable to other fields besides mathematics, and as such it potentially contributes to the unity of the theory of truth and to the universality of the correspondence principle. Furthermore, our specific conception of mathematical truth can be used to establish the unity of logic and mathematics on a new ground.

It's not common to think of logic in the following way, but logic is a field that, under most of its conceptions, breaks the unity of truth and undermines the universality of correspondence. If truth in the empirical sciences is commonly viewed as based on correspondence, and truth in mathematics is sometimes viewed as based on correspondence, truth in logic is almost never viewed as based on correspondence. On the common view, therefore, not all truth is based on the same thing: some truth is based on correspondence, and other truth is based on something else. Composite correspondence enables us to restore the unity to truth and make a significant step toward removing counter-examples to correspondence. Furthermore, it enables us to do this in a way that explains the close connection between logic and mathematics (for example, the fact that every logical truth is an image of some mathematical truth)[16]. How can this be done?

Briefly, we could proceed by pointing out that: (a) Logic works—and has to work—in the world. (b) To work in the world—for example, to succeed in transmitting correspondence truth (truth that depends on the world) from premises to conclusion—logical laws cannot conflict with the basic laws governing the behavior of objects and structures of objects in the world. (c) If logic is grounded in certain universal laws governing objects and structures of objects in the world, this will explain how it works in the world (how it succeeds in transmitting truth from sentences grounded in the world to sentences grounded in the world). Furthermore, if the laws in question have an exceptionally strong modal force, this will explain the exceptionally strong modal force of logic. (d) Formal laws are universal. (e) Formal laws have an exceptionally strong modal force. (f) All logical laws have formal correlates. All this suggests that logical truth is based on correspondence with formal

[16] The logical truth "*Pa V* $\sim$*Pa*" is an image of the mathematical truth that a is in the union of P and its complement (in the given universe), the logical inference "*PaVQa,* $\sim$*Pa; therefore Qa*" is the image of the mathematical inference that if a is in the union of P and Q, yet is not in P, then it is in Q, and so on.

laws.[17] It would also explain the close connection between logic and mathematics: mathematics *studies* formal laws, while logic *applies* them in/through language).[18]

What about other areas?—Composite correspondence carries some promise for other areas as well. Take ethics, for example. Our approach suggests that in investigating truth in ethics we don't start by trying to fit ethics into the standard template of correspondence, inspired by Tarski's theory, with its emphasis on individuals and their properties. Instead, we try to understand morality in its own terms. We try to figure out whether there is, or should be, something objective in our moral judgments, something that transcends our subjective and communal emotions (preferences, etc.), and if there is, what kind of thing it is. And we try to understand how humans reach this objective ground of morality, what standard (pattern) of correspondence is involved, and how it relates to the standards (patterns) of correspondence in other fields. In this way, if there is truth in ethics it will be based on correspondence—composite correspondence, like all other fields.

References

Benacerraf, P. (1965). What numbers could not be. *Philosophical Review, 74,* 47–73.
Benacerraf, P. (1973). Mathematical truth. *Journal of Philosophy, 70,* 661–680.
Field, H. (1989). *Realism, mathematics & modality*. Oxford: Basil Blackwell.
Frege, G. (1884). *The foundations of arithmetic: A logico-mathematical enquiry into the concept of number* (trans: J. L. Austin). Evanston: Northwestern. (1968).
Frege, G. (1918). Thoughts. In P. T. Geach (Ed.), *Logical investigations* (trans: P. T. Geach & R. H. Stoothoff) (pp. 1–30). Oxford: Basil Blackwell.
Goodman, N., & Quine W. V. (1947). Steps toward a constructive nominalism. *Journal of Symbolic Logic, 12,* 105–122.
Horgan, T. (2001). Contextual semantics & metaphysical realism: Truth as indirect correspondence. In M. Lynch (Ed.), *The nature of truth: Classic & contemporary perspectives* (pp. 67–95). Cambridge: MIT.
Lear, J. (1982). Aristotle philosophy of mathematics. *Philosophical Review, 91,* 161–192.
Lynch, M. (2001). Realism and the correspondence theory: Introduction. In M. Lynch (Ed.), *The nature of truth: Classic & contemporary perspectives* (pp. 7–15). Cambridge: MIT.
Lynch, M. (2009). *Truth as one & many*. Oxford: Oxford University Press.
Quine, W. V. (1955). Posits & reality. *The ways of paradox and other essays* (pp. 246–254). Cambridge: Harvard University Press. (1976). (Revised and Enlarged Ed.).
Rosen, G. (1990). Modal fictionalism. *Mind, 99,* 327–354.
Shapiro, S. (1997). *Philosophy of mathematics: Structure & ontology*. Oxford: Oxford University Press.
Sher, G. (1991). *The bounds of logic: A generalized viewpoint*. Cambridge: MIT.
Sher, G. (1996). Did tarski commit 'tarski's fallacy'? *Journal of Symbolic Logic, 61,* 653–686.
Sher, G. (1998–9). On the possibility of a substantive theory of truth. *Synthese, 117,* 133–172.
Sher, G. (1999a). Is logic a theory of the obvious? *European Review of Philosophy, 4,* 207–238.

[17] This correspondence would be properly composite (i.e., will involve more than one step) if logical laws are construed as directly concerning linguistic entities and indirectly the world.
[18] For a detailed discussion of this approach to logic and mathematics from the perspective of logic, see, e.g., Sher (1996, 1999a, 2008).

Sher, G. (1999b). Is there a place for philosophy in quine's theory? *Journal of Philosophy, 96,* 491–524.
Sher, G. (2004). In search of a substantive theory of truth. *Journal of Philosophy, 101,* 5–36.
Sher, G. (2008). Tarski's thesis. In D. Patterson (Ed.), *New essays on tarski and philosophy* (pp. 300–339). Oxford: Oxford University Press.
Sher, G. (2010). Epistemic friction: Reflections on knowledge, truth, and logic. *Erkenntnis, 72,* 151–176.
Sher, G. (2011). Is logic in the mind or in the world? *Synthese, 181,* 353–365.
Wigner, E. P. (1960). The unreasonable effectiveness of mathematics in the natural sciences. *Communications in Pure & Applied Mathematics, 13,* 1–14.
Wright, C. (1992). *Truth and objectivity*. Cambridge: Harvard University Press.
Yablo, S. (2001). Go figure: A path through fictionalism. *Midwest Studies in Philosophy, 25,* 72–102.

Chapter 10
Complexity and Hierarchy in Truth Predicates

Michael Glanzberg

Abstract In this paper, I speak in favor of hierarchies in the theory of truth. I argue that hierarchies are more well-motivated and can provide better and more workable theories than is often assumed. Along the way, I sketch the sort of hierarchy I believe is plausible and defensible. My defense of hierarchies assumes an 'inflationary' view of truth that sees truth as a substantial semantic concept. I argue that if one adopts this view of truth, hierarchies arise naturally. I also show that this approach to truth makes it a very complex concept. I argue that this complexity helps motivate hierarchies. Complexity and hierarchy go together, if you adopt the right view of truth.

Since seminal work of Tarski (e.g. Tarski 1935) hierarchies have been much discussed in the literature on truth and paradox. Especially in recent years, this discussion has been decidedly negative. Tarski's hierarchy of languages is sometimes described as the "orthodox" response to the Liar paradox (e.g. Kripke 1975), but it is an orthodoxy many authors have gone to great lengths to avoid. Frequently, the unpalatability of hierarchies is taken for granted, and the main task is taken to be developing theories of truth that avoid them. In this paper, I shall speak in favor of hierarchies. I shall argue that hierarchies are more well-motivated and can provide better and more workable theories than is often assumed. I shall not argue here that hierarchies are inevitable (though I have argued that elsewhere); rather, I shall argue that if we wind up with one, that is not by itself a bad result. Along the way, I shall sketch the sort of hierarchy I believe is plausible and defensible, which is different in important respects from the orthodox Tarskian one.

Versions of this material were presented at the Seventh Barcelona Workshop on Issues in the Theory of Reference, University of Barcelona, June 2011, the Truth at Work Conference, Institut d'Histoire et de Philosophie des Sciences et des Techniques, Université Paris-1 and Ecole Normal Supérieure, June 2011, and Birkbeck College, University of London, February 2012. Thanks to all the participants at those events for very helpful discussions. I am especially grateful to Eduardo Barrio, Jc Beall, Alexi Burgess, Marcus Giaquinto, Øystein Linnebo, and two anonymous referees for comments and discussion of previous versions of this paper.

M. Glanzberg
Northwestern University, Chicago/Evanston, USA
e-mail: m-glanzberg@northwestern.edu

T. Achourioti et al. (Eds.), *Unifying the Philosophy of Truth,* Logic, Epistemology, and the Unity of Science 36, DOI 10.1007/978-94-017-9673-6_10

My defense of hierarchies will assume a particular view of the nature of truth that is fundamentally 'inflationary' and sees truth as a substantial semantic concept. My main thesis will be that if one adopts this view of truth, hierarchies arise naturally. In contrast, if you adopt a deflationist line, hierarchies are much less plausible, and certainly lack motivation. As a corollary of these claims, we will see an important way in which theorizing about the nature of truth affects how we proceed with the task of addressing the paradoxes. We will also see along the way that the approach to truth I shall advocate makes truth a complex concept, and that in the presence of self-applicative truth and the Liar, truth becomes a very complex concept. As I shall show, this complexity helps motivate hierarchies. Complexity and hierarchy go together, if you adopt the right view of truth.

The plan for this paper is as follows. In Sect. 10.1, I shall introduce the semantic view of truth I shall suppose throughout the paper, and contrast it with deflationist views. In Sect. 10.2, I shall introduce the notion of reflection, which is a process by which we can make our implicit grasp of concepts like semantic truth explicit. I shall go on to argue that reflection is an engine that generates hierarchies, especially when combined with the complexity of truth according to the semantic view. Examples of reflection, and how they indicate that truth is a complex concept will be discussed in Sect. 10.3. In Sect. 10.4, I shall argue that approaching truth through reflection motivates hierarchies. I shall also sketch the form I think a good hierarchical theory of truth should take in this section. I shall continue my defense of hierarchies of this sort in Sect. 10.5. I shall argue there that my favored form of hierarchy offers a plausible theory that works well, and is not vulnerable to some standard objections. The hierarchy is not without costs, but they are not nearly so high as it is often assumed. I shall conclude in Sect. 10.6 by returning to the issue of the nature of truth with which the paper began. I shall argue there that the defense of hierarchies I offer on the basis of a semantic view of truth is not available to deflationists.

10.1 The Nature and Complexity of Truth

My discussion of hierarchies will be informed by some background ideas about the nature of truth, of the sort discussed in the more traditional literature on the metaphysics of truth but not typically applied to the paradoxes. I shall isolate two very broad approaches to truth: one deflationary, the other inflationary and motivated by the role of truth in semantics. I shall go on to argue that the inflationary view of truth can support and motivate hierarchies.[1]

Let us begin by noting some general features of *deflationary* theories of truth. Though there are many such theories, different in important respects, they share the idea that in some ways truth is not a substantial property with an interesting underlying

[1] This way of thinking about the nature of truth and its role in solutions to the paradoxes comes from joint work with Jc Beall (e.g. Beall and Glanzberg 2008), though Beall himself prefers a very different set of options to the ones I endorse here.

nature. I shall take as my representative deflationary position the *transparency* view of truth advocated by Beall (e.g. Beall 2009) and Field (e.g. Field 1994); not for least of reasons, they are explicit about the logical properties they attribute to truth, and are concerned with the paradoxes. In the spirit of disquotational view of truth, Beall and Field hold the main feature of truth to be the intersubstitutability of ϕ and $Tr(\ulcorner\phi\urcorner)$ in all non-opaque contexts. This in turn supports the logical role of truth in enabling such functions as simulating infinite conjunctions or disjunctions. Thus, the nature of truth is exhausted by its simple logical properties, which in turn endow it with a specific measure of logical utility. We can of course say these features are interesting, especially to logicians, but with the broader deflationist tradition, Beall and Field will insist there is no underlying metaphysical nature of truth which determines its logical properties. Most importantly for what follows, the rules of substitution are essentially all there is to the nature of truth, if indeed that is a nature at all. They are simple, obvious rules, and mastering them is complete mastery of the concept of truth.[2]

My main focus here will be on an opposing, inflationary, view of truth. To see how it works, we might start with important work of Field (1972). In that early work (he has drastically changed his view since), Field envisages a substantial theory of truth combining two components. One is the familiar Tarskian apparatus, which provides an inductive characterization of truth for a language. The other is an account of 'primitive denotation', which he imagines will be along the lines of the causal theory of reference and provide a reductive account of the basic relations of reference and satisfaction from which the Tarskian theory is constructed.

There are some features of Field's view that I shall ignore, especially, the specific role of the causal theory of reference in providing a reductive analysis of primitive denotation. Abstracting from such details, we get a picture of truth with several distinctive features. First, the truth of any sentence is determined by substantial facts about reference and satisfaction, whatever underwrites those facts. These are basic *word-to-world* relations. Second, those facts are facts about the semantic properties of parts of a language, like the referents of singular terms and satisfaction for predicates. Third, truth is determined compositionally: the truth of a complex sentence is determined by the semantic properties of its parts, including the truth of embedded sentences, in a compositional way. Again, it is the language which determines how such composition is carried out.

One way to bring all these points together is to capture them under the idea that truth is a fundamental semantic property of sentences of a language.[3] We might start with the widespread idea that to understand a sentence is to a great extent to understand its truth conditions.[4] When we articulate a theory of truth, along the lines

[2] The transparency view is a descendent of the disquotational view of truth associated with Leeds (1978) and Quine (1970). Of course, there are a number of other versions of deflationism, which are importantly different, but this sample view will suffice for our purposes here.

[3] The usual provisos apply here, about sentences in contexts.

[4] This view of meaning is closely associated with Davidson (e.g. Davidson 1967), but it is also part of the long tradition in philosophy of language from Frege to Carnap to Montague and beyond.

of Tarski plus Field, we are articulating the way this fundamental semantic property works for a particular language. Yet at the same time, as Davidson (1990) reminds us, we can see the fundamental properties of truth illustrated by the application to a specific language. We can see, for instance, the role of word-to-world relations and semantic composition in fixing truth. Truth, on this view, is a fundamental semantic property whose features we can see at work in particular languages, like the ones we speak.

A theory like this can be seen as a descendent of the traditional correspondence theory of truth, at least in a limited way. It bears out the traditional idea of truth as relying on substantial truth-making relations to the world; but it does not rely on a metaphysics of facts, or a structural correspondence relation between truth bearers and facts, as the traditional correspondence theories of the late nineteenth and early twentieth centuries did.[5] The structural correspondence relation to a single fact is replaced by multiple relations of reference and satisfaction for parts of sentence. These are brought together by semantic properties of composition, rather than metaphysical ones of fact creation.

I do think this approach to truth is appealing, and captures a lot of what seemed right about the idea of correspondence. However, it is not my goal to defend this particular view of truth here. Instead, I shall try to argue that it offers motivation and support for hierarchies of truth predicates. For these purposes, one feature of the approach is especially important. In sharp contrast to deflationist approaches, this approach implies that truth has substantial internal structure. This structure is shown in the division of labor between facts about reference and the compositional determination of truth value. These combine to fix the truth values of sentences by way of the internal structure of the sentences and the nature of the things their constituents refer to. Whatever the logical behavior of truth is, it is determined by this internal structure of reference and satisfaction and semantic composition. This is a fundamentally different picture than the transparency view offers, according to which truth is at heart a simple property, fully captured by simple substitution rules, with effectively no internal structure and nothing more fundamental to determine them. Furthermore, Beall and I have argued (2008) that the approach I am endorsing does not yield full transparency. Moreover, whatever degree of transparency truth enjoys is not its basic feature, but a consequence of its more basic properties. Truth, on the approach I prefer, does have an underlying nature.

One consequence of this is that, according to the view I endorse, there is a sense in which truth is *complex*. The internal structure of truth shows possibly complex ways that the truth of sentences of a language are determined by reference and satisfaction.

Of course, like all philosophical views, it is controversial, and conceptual role or inferentialist approaches to meaning deny it. Indeed, deflationism of the sort described by Field (1986) also denies it. Though Davidson endorses the close connection between truth and meaning, he holds a very different view of the place of reference in semantics, as we see in Davidson (1977, 1990).

[5] I have in mind the correspondence theory in something like the form it appeared in work of Russell (e.g. Russell 1912) and Moore (e.g. Moore 1953). The sort of theory I advocate here perhaps has more in common with the sort of correspondence discussed by David (1994).

The underlying nature of truth reveals a specific kind of complexity, and it is not one the transparency view finds. As we will see in a moment, the mathematics of truth bears out this complexity in precise ways. Truth, seen from this view, is a complex property.

To fix some terminology, let us call the view that truth is a fundamental semantic property with complex internal structure the *semantic view* of truth.[6] I shall argue in what follows that the semantic view of truth provides motivation and justification for hierarchies. I shall also argue that it indicates a special role for considerations of complexity in the response to the paradoxes, as we will see in the following sections.

10.2 Implicit Grasp and Reflection

Now that we have a basic view of truth in hand, what does it tell us about the paradoxes or hierarchies? In this section, I shall introduce the notion of what I call *reflection*. Reflection enters the picture when we ask if and how we understand the complex property of truth. In virtue of our grasp of our languages, I shall argue, we are in a position to engage in a form of reflection which reveals some of its features. I shall go on to argue in Sects. 10.3 and 10.4 that reflection, together with the complexity of truth (according to the semantic view) are the engines that generate hierarchies. In Sect. 10.4 I shall explain what those hierarchies are like. But first, we need to see what reflection is, and how it works.

The notion of reflection may be introduced by asking what is involved in understanding the nature of the truth predicate? It follows from the semantic view that understanding truth is simultaneously incredibly easy and very hard. First, we might say, understanding truth is easy: in virtue of having competence with our languages, we already understand it. Or more precisely, you implicitly *grasp* it. Truth is, according to the view in question, a fundamental semantic property of your language. In virtue of understanding your language, you implicitly make use of the concept of truth. You thus have that concept in your cognitive repertoire. This is a form of implicit grasp, as you have and make use of the concept. It is implicit, as the concepts that figure into the basic functioning of your language need not be overly accessible to you. But regardless, you do, according to the semantic view, already possess substantial implicit grasp of the concept of truth.

On the other hand, coming to understand the nature of truth is, according to the semantic view, extremely hard. Truth has a substantial and complex underlying nature, including aspects of recursion, and complicated notions like reference. Coming to understand that can be, and experience shows is, an extremely difficult challenge.

[6] The terminology is somewhat unfortunate, as Tarski already appropriated the term 'semantic' for his semantic conception of truth. Alas, it is not easy to say just what Tarski had in mind by that. Depending on what he did have in mind, my use of the term may or may not overlap with his. In previous versions of this work, I used the term 'complex view' of truth, but that proves confusing when we come to discuss complexity results below.

Indeed, according to the semantic view, it is hard in just the same way that coming to fully understand the semantics of a human language is hard. Anyone who has dipped so much as a toe into the field of semantics knows just how hard that is! Indeed, it may be harder, as we need to understand not merely the semantics of one language, but a fundamental concept that is common across the semantics of many languages.

How can understanding truth be both hard and easy at the same time? It is not really so mysterious. That is just what one would expect from implicit grasp. What is easy is the state of having such implicit grasp (well, easy in virtue of whatever enables us to learn our languages!). Great efforts have gone into explaining what that sort of state is, and I shall not delve into the issue here.[7] All that matters for us is that whatever this implicit grasp consists in, it is not *explicit*. When we see understanding truth as hard, we are asking for an explicit articulation of the concept, let us say, by offering a theory of it in the appropriate setting. For the semantic view, this setting will include the semantics of some language, and ultimately more than that; but regardless, we are asking for an explicit articulation of the complex nature of the concept. For concepts we grasp implicitly, it is making them explicit that can be hard.

When we encounter some concept with a complex underlying nature, that we also enjoy an implicit grasp of, we have at our disposal a unique way to study that concept. We can *reflect* on our own abilities that are underwritten by the implicit grasp, and thereby come to learn about the concept. We can reflect on our linguistic abilities, and thereby learn about the languages we speak, and the concept of truth which plays a fundamental semantic role in them. In doing so, we can begin to articulate the nature of the concept explicitly.[8]

Reflection as I understand it is an activity we can engage in, when we encounter concepts of which we have some implicit grasp. In the case of truth, we have assumed that our understanding of the terms and sentences of our languages includes an implicit understanding of their properties related to truth and reference. This understanding, though highly tacit, guides our linguistic uses, our comprehension of sentences, and other manifestations of our linguistic competence. This provides us with some evidence we can use to try to make the nature of the concept explicit. The evidence is available through introspection of the contents of our sentences, and self-directed observation of our linguistic practices. It is thus generally evidence we can access by focusing our attention on ourselves, our thoughts, and our activities. Based on such evidence, we can begin to articulate the semantic properties of our languages, presumably in the form of some semantic theory. This, we assume, will include articulating a body of semantic facts about truth and reference, and how they

[7] See the large literature in the philosophy of language on tacit knowledge, including such contributions as Chomsky (1980), Davies (1987), and Higginbotham (1989).

[8] Notions of reflection have appeared in the literature on truth, though often with relatively little discussion. For instance, Kripke's famous remark about "some later stage in the development of natural language, one in which speakers reflect on the generation process leading to the minimal fixed point" (Kripke 1975, p. 80) seems to be gesturing towards the sort of reflection I have in mind. I have discussed this idea in my (2006), and in a somewhat different form in my (2004c).

combine compositionally. This, in turn, forms the basis of a theory of truth. Though the concept of truth is implicit in our languages, we can put ourselves in a position to offer an explicit articulation of it by reflection.

Reflection is a complicated and demanding process. In what follows, I shall focus mainly on articulating the semantic properties of our own languages or languages like them by reflection, or even highly simplified formal languages. Strictly speaking, this is only a part of the process of reflection that would have to go on to fully articulate the nature of the concept of truth, as that concept applies across languages. But it is already a complicated enough task, and it is the one that is of special interest when it comes to the paradoxes and hierarchies, so I shall focus on it.

With a task of this difficulty, there is no guarantee that we will in any one instance of reflection produce a particularly complete theory of the underlying concepts of our semantics. There is not even any guarantee we will get those properties right. We have a complex concept and only some indirect sources of evidence from which to try to characterize it. Even for a gifted linguist or logician, there may well be limits on how much can be accomplished in any one exercise of reflection. Indeed, as I shall argue, the paradoxes or various incompleteness phenomena show that in some cases, even the most gifted logicians will fail to capture the entirety of certain concepts in any one instance of reflection. In these cases, I claim, hierarchies ensue. But we can already see why that might be an unsurprising result. The difficulty in producing comprehensive theories in reflection is not one that stems only from incompleteness or paradox. It stems from the difficulties of reflection—the complexity of the task, and the limited resources we have to do it—as well.

10.3 Models of Reflection and Complexity

So far, we have seen that according to the semantic view of truth, we have implicit grasp of the complex concept of truth. That grasp is something we can try to make explicit by reflection, and in particular, we can try to make the semantic properties of the languages we speak or related ones explicit by reflection. In this section, I shall explore some examples of how that process might go, and what the results might be. I shall focus on formal languages, following the tradition in work on truth predicates. They will show us enough to see how complex the results of reflection must be. As we will see in Sect. 10.4, this in turn will show us how reflection can generate hierarchies.

I shall begin with a very Tarskian case of a language with no semantic predicates. Though it will only occasionally matter, let us take the language $\mathcal{L}$ to be the language of arithmetic.[9] For a language like this, classic work of Tarski (1935) provides a very good illustration of what reflection should yield. Indeed, for this case, we can

[9] I will be moving back and forth between proof-theoretic and definability-theoretic perspectives. For proof theory, it will sometimes matter that $\mathcal{L}$ is the language of arithmetic, though we will rarely get into enough technical detail to see this. Definability theory often prefers to work with purely

articulate what a fully successful exercise in reflection should yield, without having to face paradox problems.

The task of reflection is to make explicit the semantic properties of a language, especially, those properties relating to truth and reference. For a language like $\mathcal{L}$, that is in effect to write out the definition of truth in the way Tarski showed us. Hence, a Tarskian theory of truth for $\mathcal{L}$ is good representation of what successful reflection should look like for such a language.

We should be a little more specific about what will count as a 'Tarskian theory of truth', and we will see that there are several different ways to describe it formally. One way, oriented around model theory, is to see the Tarskian definition as the definition of truth in a model. Hence, to display the Tarskian truth theory, we need to provide a model $\mathfrak{M}$ of $\mathcal{L}$, and the definition of truth in a model $\mathfrak{M} \models \phi$ for sentences of $\mathcal{L}$. If we do this, we should not lose sight of the inductive nature of the definition of truth in a model and its route through satisfaction. Moreover, as reflection asks us to make the concepts at work in the semantics explicit, we should follow Tarski in displaying the truth predicate over the structure explicitly. This is in effect to define the Tarski truth predicate for $\mathcal{L}$ over $\mathfrak{M}$, which provides an interpretation for a language extending $\mathcal{L}$ with a Tarskian truth predicate Tr.[10]

We can also take a more proof-theoretic approach, and ask for an axiomatic theory of truth. A good proof-theoretic representation of the way truth works in a language like $\mathcal{L}$ is provided by directly axiomatizing the compositional definition of truth Tarski provides. To do this, we add the truth predicate Tr to $\mathcal{L}$, and the following axioms:[11]

1. $\forall s \forall t(Tr(s \dot{=} t) \leftrightarrow s^\circ = t^\circ)$
2. $\forall x(Sent(x) \rightarrow (Tr(\dot{\neg} x) \leftrightarrow \neg Tr(x)))$
3. $\forall x \forall y(Sent(x \dot{\wedge} y) \rightarrow (Tr(x \dot{\wedge} y) \leftrightarrow Tr(x) \wedge Tr(y)))$
4. $\forall x \forall y(Sent(x \dot{\vee} y) \rightarrow (Tr(x \dot{\vee} y) \leftrightarrow Tr(x) \vee Tr(y)))$
5. $\forall v \forall x(Sent(\dot{\forall} vx) \rightarrow (Tr(\dot{\forall} vx) \leftrightarrow \forall t Tr(x(t/v))))$
6. $\forall v \forall x(Sent(\dot{\exists} vx) \rightarrow (Tr(\dot{\exists} vx) \leftrightarrow \exists t Tr(x(t/v))))$

relational structures and replace functions with relations; but again, we will not get into enough details to see this.

[10] When thinking about the semantic properties of a language like the ones we speak, we should probably focus on the intended interpretation, and so perhaps for $\mathcal{L}$ we should be working with $\mathbb{N}$ rather than an arbitrary model. Occasionally, we will need to know the model is reasonably nice, but for the most part, we will not be concerned with which structure it is. We should also note that the mathematical definition of truth is a mathematical representation of a concept with empirical applications (as Etchemendy (1988) and Soames (1984) reminded us).

[11] I follow the notational conventions of Halbach (2011). They are mostly standard. $^\circ$ is the evaluation function for terms, which is definable in PA. Recall that the language of PA has no predicates other than identity, and hence the form of the axioms below is specific to PA. Minor changes can accomodate other sorts of languages.

We always think of axioms like these as added to some base theory. One good representative starts with PA, but with induction extended to the expanded language including the truth predicate Tr. The theory which adds the truth axioms to this base theory is known as the *compositional truth theory* or CT (following the terminology of Halbach 2011).[12]

Both the model-theoretic definition of truth over a structure and CT represent successful reflection for $\mathcal{L}$. Both provide essentially complete accounts of truth for $\mathcal{L}$. The truth predicate we define model-theoretically provides an extensionally correct truth predicate for $\mathcal{L}$ over the base structure. Likewise, CT proves each instance of the T-schema, and so is extensionally adequate. But we have more than an extensionally correct truth predicate. Both approaches generate that predicate by describing correctly the semantic workings of the language, just as reflection asks. Both yield useful results. For instance, CT can prove the consistency of PA, while the model-theoretic definition of truth is the basis for pretty much everything else that happens in model theory. In a less mathematical vein, the way the theories illustrate the compositional determination of truth via satisfaction shows something important about how the semantics of a language can work. We thus have two good examples, for a highly idealized case, of how we may present a truth predicate, and how we can do so via reflection on the semantic properties of a language.[13]

In Sect. 10.1, I discussed how the semantic view of truth makes truth a property with a complex underlying nature. Both our examples of reflection give us a way to make this mathematically more precise, as we can apply complexity measures to model-theoretic and proof-theoretic truth predicates. Let us begin with the proof-theoretic CT. We can measure its proof-theoretic strength in a few ways. It is stronger than PA, as it proves the global reflection principle for PA, and in fact, it is slightly stronger than $PA + RFN_{PA}$. Another measure is that CT is proof-theoretically equivalent to the second-order theory ACA, which is second-order PA with the full induction schema and a comprehension axiom for arithmetic formulas (i.e. ones with no second-order quantifiers).[14]

Definability theory also provides measures of the complexity of the model-theoretic truth predicate. Over reasonably nice models (including the standard model of arithmetic), the truth predicate is not elementary, but Δ^1_1.[15]

---

[12] In many cases, I shall talk about formal theories without going into full details of their expositions, but this case is central enough, and illustrative enough, that the details seem to be worth mentioning.

[13] There are some limits to what these sorts of models of reflection capture. For one thing, we may well learn about $\mathcal{L}$ and PA more explicitly than we learn our natural languages. Neither approach fully addresses the question of how reference and satisfaction are fixed for a language. This is illustrated by the fact that CT does not rule out non-standard models. Generally, these are good theories of how truth works, but by no means complete theories of intentionality.

[14] For discussion of these sorts of results, see for instance Feferman (1991) or Halbach (2011). These results are proof-theoretically somewhat delicate; for instance, as is well-known, if we weaken the induction schema of CT we get back a conservative extension of PA.

[15] This result is quite general, and not really specific to arithmetic. Moschovakis (1974) proves a general version only assuming what he calls an 'acceptable structure'. Some assumptions are needed; for instance, the result fails for recursively saturated structures (cf. Barwise 1975).

Both results are formal versions of a general point: truth is complex, and complex enough to be more complex than whatever we start with. If we start with PA, we get a stronger theory. If we start with a model, we get a predicate that is not elementary over that model. Reflection thus can, when successful, yield something markedly more complex than what we had when reflection began. We can capture this in terms of proof-theoretic strength of theories, or in definability terms for structures of certain sorts, but the idea of added complexity from reflection on truth stands out.[16]

The jump up in complexity is significant, but also limited. In proof-theoretic terms, for instance, we are able to prove facts about soundness we could not before, as well as some corresponding statements of arithmetic. Much more becomes elementary, in definability-theoretic terms. But at the same time, the compositional theory reveals a fairly modest jump in complexity. ACA is a weak second-order theory, building in a substantial amount of predicativity, while a Δ^1_1 predicate is just above elementary by standard definability-theoretic measures. We see in these examples the complexity that goes with the inner workings of truth, but only a limited amount of it.

In addition to the way it can yield complex results, this example shows one other feature that often goes with reflection. In many cases, reflection involves a change of topic or subject-matter of investigation. Suppose we start with the language of arithmetic, and a theory like PA. The subject-matter we investigate with these resources is clear enough: it is arithmetic. We do so in an interpreted language which has semantic properties, but those are not what the sentences of PA talk about or the subject-matter of arithmetic. When we engage in reflection on the semantics of a language, we make its semantic properties a topic of investigation. If it was not part of the subject-matter we were investigating before, it becomes so. Reflection can overtly change the topic. Once something becomes part of the subject-matter of investigation, we can start to build up explicit theories of it, and so, carry out the task of rendering something explicit that was previously merely implicit.

The extent to which reflection changes the topic is often a matter of degree, and it is not always trivial to tell how much takes place. To make this vivid, recall that PA can define lots of truth predicates, such as the Σ_n-truth predicates. It can also define proof predicates, and all sorts of other things that might not have transparently seemed to be part of the subject-matter of arithmetic. But the Tarskian example does show that in some cases, a fairly pronounced change of topic occurs. The complexity results in effect confirm this, by showing that we cannot take our original theory to implicitly characterize our new subject-matter. The way reflection can go with changes of topic will become important as we explore hierarchies in Sects. 10.4 and 10.5.[17]

So far, I have presented a view of truth which makes it a complex concept, and discussed the process of reflection which can make aspects of that concept explicit.

[16] This is what Horsten (2011) calls the power of the compositional theory of truth. As he notes, it is a surprising fact that we gain in arithmetic strength simply by adding semantic axioms.

[17] I have argued (Glanzberg 2002, 2004a, 2006) that in the kinds of cases at issue for the paradox, reflection invariably does change the topic.

Our example of a language like $\mathcal{L}$ of arithmetic is, of course, highly idealized. Though there is nothing wrong with working with idealized examples, there is one feature of this idealization which we must remove. By insisting that $\mathcal{L}$ contain no truth prediate, we avoid issues of paradox and self-applicative truth predicates that have been the focus of much of the logical work on truth. To address the issue of hierarchies, and to get a more useful model of reflection, we need to remove this restriction and work with languages with a self-applicative truth predicate.

In the case of a language with a self-applicative truth predicate, the basic task of reflection remains the same: to capture the semantic properties of the language. But the semantic properties include the word 'true', and in turn, the word 'true' is supposed to mean something which relates closely to the semantics of our language. We need the semantics and the word 'true' to properly relate. From the perspective I am taking, we should not assume they will be identical. The semantics of the language is part of its underlying workings, and those need not coincide with the meaning of any expression. But all the same, our insights into the nature of truth and into the meaning of the word 'true' are related, and we would clearly miss something about the meaning of the word 'true' if the two had nothing to do with each other. So, even though the semantics of a language and its truth predicate need not be treated in exactly the same ways, they do interact.

One thing we have learned from the study of the Liar paradox is that this interaction is in fact quite complex. This complexity is the basis for a general strategy for dealing with self-applicative truth and paradox. The basic idea is that we can model self-applicative truth as the result of iterating a more Tarskian truth construction, where the iterations increasingly well capture the interactions between the truth predicate and the semantic properties of the language it is in. When a self-applicative truth predicate figures into the language, these iterations can be very long indeed, but they do reach stages where a reasonable snapshot of the semantics of the language, including its truth predicate, is reached.

This is one way of thinking about the Kripke construction (Kripke 1975), but I shall suggest, it is a common feature of a number of the leading approaches to the paradoxes. To flesh out the idea, I shall review a couple of ways of modeling it formally. However, the mathematics involved gets quite complex very quickly, and in many cases, the mathematical details will not really be important for the argument I am making here. So, I shall try to give a very rough indication of what the formal models might look like, but I shall often skip a great deal of substantial and interesting detail, and I shall occasionally simply cite results.

Let us start with the Kripke construction as an example. Recall the main features of the Kripke construction. We start with a language $\mathcal{L}^+$ that adds a truth predicate Tr to $\mathcal{L}$, and has no Tarskian syntactic restrictions on the application of Tr. We fix a model $\mathfrak{M}$ for $\mathcal{L}$, and our job is to extend it to a model for $\mathcal{L}^+$ by defining a truth predicate, as in the Tarskian case above. Unlike the Tarskian case, the truth predicate is interpreted partially, by an extension and anti-extension $\mathcal{I} = \langle E, A\rangle$. So a model of $\mathcal{L}^+$ looks like $\mathfrak{M}^+ = \langle \mathfrak{M}, \mathcal{I}\rangle$. We can think of each $\mathfrak{M}^+$ as summarizing an exercise in reflection for $\mathcal{L}^+$, and so reporting an account of its semantics. As before, we should remember that it is the structure and the definition of truth in it

that provides our explanation of the semantics of the language. As we now are using partial predicates, this will involve a choice of valuation scheme—like the Strong Kleene or supervaluation scheme. Aside from the treatment of partial predicates the exercise goes much the way we learned from Tarski.

Part of the exercise in reflection will be the interpretation $\mathcal{I}$ of the truth predicate. Typically, this will not be all that good, as it will not come close enough to the semantics provided by $\mathfrak{M}^+$. Thus, our attempt at reflection represented by such structures may be only partially successful. But one of the insights of the Kripke construction (over and above the partial treatment of the truth predicate[18]) is that we can define a process for constructing a sequence of models of the form $\mathfrak{M}^+$ that provide better and better approximations of the semantics of $\mathcal{L}^+$ and the interpretation of Tr. If we iterate the process long enough, we can get something very good. Very good here means reaching a fixed point, where further iteration does not improve our model. In fixed point models, the semantics provided by $\mathfrak{M}^+$ and the interpretation $\mathcal{I}$ of Tr come very close to coinciding, as we get the fixed point property:

$$\mathfrak{M}^+ \models Tr(\ulcorner \phi \urcorner) \leftrightarrow \mathfrak{M}^+ \models \phi.$$

My notation here masks that $\mathfrak{M}^+$ is a partial model, and there will be may sentences, including Liar sentences, which are treated as gaps in fixed points. A classical version, known as the 'closed-off Kripke fixed point', simply replaces the partial interpretation with its extension. Doing so gives us:

$$\langle \mathfrak{M}, E \rangle \models ((Tr(\ulcorner \phi \urcorner) \vee Tr(\ulcorner \neg \phi \urcorner)) \rightarrow (Tr(\ulcorner \phi \urcorner) \leftrightarrow \phi)).$$

Either way, we see that at least for non-pathological sentences (those where we have $Tr(\ulcorner \phi \urcorner) \vee Tr(\ulcorner \neg \phi \urcorner)$), our semantics for $\mathcal{L}^+$ and the interpretation of Tr agree.

There are many different ways of developing this idea formally, and many more ways of interpreting it. I shall put it in the setting of reflection. If we think of each stage in the process as the results of reflection on the semantics of a language, we can see the whole process as an extended iteration of reflection. Our process of reflection becomes an extended one, involving repeated reflection on the semantics of the language, relative to some hypothesis[19] about the interpretation of Tr in it, and repeated refinement of the results until we reach something that is a plausible interpretation of the whole language $\mathcal{L}^+$. To give this idea a name, call it the *long iteration* strategy for reflection on languages with self-applicative truth predicates.

The Kripkean implementation of the long iteration strategy comes close to simply a long iteration of the Tarskian sort of reflection we discussed above. As it is usually presented, it is not quite exactly that, as the use of a partial truth predicate at least as an intermediate step is not Tarskian. However, with some effort, the Kripke process and a transfinite Tarskian hierarchy of languages can be shown to be equivalent in important respects (Halbach 1997).

---

[18] Which admittedly had precursors in the literature, such as van Fraassen (1968, 1970).

[19] The role of hypotheses in the process is highlighted by the revision theory of truth (Gupta and Belnap 1993).

I have emphasized that reflection is something we engage in; something we do. However, it is best not to think of long iteration as something we will carry out step by step. Reaching a fixed point often requires iterating well into the transfinite ordinals, and we cannot do that step by step. (The strong Kleene valuation reaches a fixed point at ω^1_{CK}, the first non-recursive ordinal.) Rather, we should think of the long iteration strategy as being used the very way that it is presented by Kripke and others. Typically, we are shown features of the process of building interpretations, like monotonicity, and then we see that a process is triggered which we can prove reaches a fixed point. This is a very complicated story, no doubt; but from the perspective on truth we are adopting here, such complication is no problem. We already observed that truth is a complex property, with a substantial underlying 'nature'. Even in the simple Tarskian case, we saw that reflection is a fairly complex endeavor, producing results of complexity measurably higher than we started with. What we learn here is that truth is *very complex*. We see this in the nature of the long iteration process. We also see it in the results. Whereas a Tarski truth predicate is Δ^1_1 over the ground model (with the right assumptions), the Kripke minimal fixed point is Π^1_1-complete. As I described it above, the Tarskian truth predicate provides only slightly more complexity, while the Kripke one provides a great deal more.

To illustrate the long iteration strategy, I have used the familiar Kripke construction as an example. This is not essential to my main point, which is that a more complicated form of reflection can provide an articulation of the semantics of a language containing its own truth predicate. Other approaches might provide slightly different results than the Kripkean one (e.g. Gaifman 1992). Actually, long iteration is a feature of practically every modern approach to the Liar. It is clearly on display in the revision theory of truth (Gupta and Belnap 1993), in recent work on paracomplete theories of Field (2008) and paraconsistent ones of Beall (2009), and as we will see in a moment, in influential proof-theoretic approaches too. Of course these theories differ in many respects, but they all make use of the general strategy of a long iteration process, where each stage shows some of the features we saw in the simple case of Tarskian reflection. Even if we have not yet fully understood all its details, I believe that long iteration shows us a fundamental aspect of the nature of (complex, semantic) truth.

In discussing reflection, I noted we can think of it in proof-theoretic rather than model-theoretic terms if we like. The same is true for the long iteration strategy, though the mathematical situation is not yet fully understood. We might think about long iteration simply in terms of iterating the theory CT. Indeed, if we take a fully Tarskian approach, with a hierarchy of truth predicates, we can do just that. The proof theory of hierarchies of Tarskian truth predicates and CT theories (as usual, up to appropriate proof-theoretic ordinals) has been explored by Halbach (1995, 2011). The correlation between CT and ACA continues, and levels of this hierarchy match-up to levels of ramified analysis. Yet just as with the Kripkean approach, this will not really be an adequate analysis of a language with self-applicative truth, and some care needs to be taken to build a consistent theory that does do a reasonable job of capturing self-applicative truth.

There are a number of axiomatic theories of self-applicative truth which have been explored in recent years. Perhaps the two most widely discussed are the Friedman-Sheard theory FS (Friedman and Sheard 1987) and the Kripke-Feferman theory KF (Feferman 1991).[20] I shall not attempt to go into much detail about either of these two theories, nor shall I advocate one over another as a theory of truth, but I shall discuss enough of their features to give some sense of how we can think of them as falling under the long iteration strategy.

First, just as I noted that the Kripke process relates closely to the Tarskian hierarchy of languages, we can find proof-theoretic connections between hierarchies of CT-theories and both FS and KF. KF defines each Tarskian CT-truth predicate up to the ordinal ϵ_0, and is arithmetically equivalent to ramified analysis up to that same ordinal.[21] Thus, KF encodes a long iteration, much like the Kripke minimal fixed point does. Indeed, KF models are precisely fixed point models (though not only minimal ones). Related but weaker properties hold for FS, which is arithmetically equivalent to the hierarchy of CT-theories up to ω. FS does not axiomatize anything like a fixed point property, but at least partially reflects the finite stages of the revision process of the revision theory of truth, and its notion of nearly stable truth. Thus, both FS and KF display features which connect them to iteration of Tarski-like theories.

As with the Kripke construction, we see reflections of iterated Tarskian theories, but have to make some significant modifications to preserve consistency and generate reasonably good theories. FS in a way modifies CT the least, except for allowing a self-applicative truth predicate. FS extends the compositional axioms of CT to all sentences of $\mathcal{L}^+$, but keeps the axiom for atomic sentences only for $\mathcal{L}$. The theory thus contains CT, but proves far too little about iterated applications of the truth predicate. To make up for this, rule forms of the two directions of the T-schema (known as necessitation and co-necessitation) are added. The result is a very classical theory, and as we saw, one as strong as ramified analysis up to ω. Unfortunately, it is also known to be ω-inconsistent, though it is arithmetically sound.[22]

The other approach, KF, follows Kripke's lead in building in some partiality for the truth predicate. To do so, the compositional axioms need to be changed to reflect the characteristics of negation for partial predicates (or the theory can be formulated with both truth and falsity predicates). The resulting theory is formulated in a classical metalanguage, but reflects the partiality of Kripke's approach. Thus, like Kripke's theory, it builds on Tarskian ideas, but implements them with care about partiality in the truth predicate.

Both theories offer us proof-theoretic ways of thinking about the long iteration strategy. Both show us ways to start with the basic idea of reflection on the semantics of a language we saw with the Tarskian CT, and modify it in ways to make room

[20] For extensive discussion of these and other theories, see Cantini (1996), Halbach (2011), and Horsten (2011). Feferman's work was circulated well before publication, and was reported in part by McGee (1991) and Reinhardt (1986).

[21] Feferman (1991) also presents an alternative version of KF which employs a different, and stronger, way of treating schemas. The result is equivalent to ramified analysis up to Γ_0.

[22] Many find ω-inconsistency a reason to reject FS. For an interesting discussion, see Barrio (2006).

for self-applicative truth. When we do, we get theories which capture the idea of iterating a Tarskian construction up to a suitable proof-theoretic ordinal. Thus, like the Kripke fixed point models, they capture the idea of a complex reflection on the semantics of a language with a self-applicative truth predicate, involving the core Tarskian insights, modified suitably, and iterated far enough to get a good theory. The naturalness of FS and KF help to substantiate the idea that the iteration was far enough to reach a good stopping place. Thus, though there are a great number of outstanding issues here, both technical and philosophical, I believe it is plausible enough to count these proof-theoretic options as falling within the long iteration strategy.

In either form, the long iteration strategy shows that self-applicative truth is very complex. We already saw that the complexity of the Kripke minimal fixed point is quite great, and both FS and KF are much stronger than CT, going up quite high in the levels of ramified analysis (especially KF). As I said above, I suspect this complexity is a genuine feature of self-applicative truth, and one that bears out the idea from the semantic theory that truth has a complex underlying nature.[23]

The long iteration strategy, in either model-theoretic or proof-theoretic form, gives a way to think about the kind of reflection that would be involved in reflecting on languages with self-applicative truth predicates. It is a complex process, involving iteration of basic semantic insights about truth to properly relate them to the semantics of the truth predicate within the language.

In the Tarskian case, we saw that reflection involved a marked change in topic, as it adds an entirely distinct truth predicate to the language. In the self-applicative case, that does not happen. As I mentioned, the kind of change of topic or subject-matter we see in reflection comes in degrees, and one of the features of the long iteration strategy is that it involves only a much more modest change of topic. After all, our language already contains a self-applicative truth predicate, so we are already able to talk about the semantic properties of the language.

Even so, I believe we see some rather modest aspects of change of subject-matter in the long iteration strategy. Though we have a truth predicate in the language, the task of reflection is to describe the semantics of the whole language, and that is not quite the same as talking about the truth of some sentences using the truth predicate. In model-theoretic terms, we see this in reflection providing a model and definition of truth in the model for the whole language, not merely a truth predicate. In proof-theoretic terms, we see it in the added strength of our theories. Though we have a truth predicate in the language, we begin with PA formulated in that

[23] This raises the question of whether the concept of truth is too complex to be grasped implicitly by all speakers, as the semantic view of truth requires. I do not have space to pursue this issue in depth, but let me quickly note that a great deal of work in cognitive science suggests we do have implicit grasp of complex concepts. A nice example is the concept of causation, which children have in some form starting as young as 6 months. The pressing question as I see it is not whether we can have implicit grasp of complex concepts, but how we can. The literature on perception of cause raises interesting questions about modularity and innateness of this concept. See, for instance Carey (2009) and Scholl and Tremoulet (2000).

language. We switch to a stronger theory of truth, which tries to capture the semantic properties of the language compositionally. Though it is not a switch in topic marked by the introduction of a novel predicate, we see change in subject-matter either way. The difference between these two ways topics can be changed will become more important as we discuss hierarchies in Sects. 10.4 and 10.5.

We now have some idea what reflection on the concept of truth should look like in the presence of self-applicative truth predicates. We have seen that the long iteration strategy gives us a reasonable way to take the basic Tarskian insights into the nature of semantic truth and apply them in this complex setting. And, we have seen, the results are indeed very complex. My main contention all along is that reflection, and especially reflection of this very complex sort, can lead to hierarchies. It is now time to explain why.

10.4 Complex Truth and Hierarchies

Why might we expect hierarchies in our theory of truth? Here is the general idea. We began with the proposal that truth is a fundamental semantic property. This makes truth a complex property, but one whose nature can be studied by reflection. We saw that reflection genuinely indicates complexity, in mathematically measurable ways. When we take into account the self-applicative nature of truth predicates, and the interactions between the word 'true' and the underlying semantics of a language, we see that in fact truth is very complex. Again, this complexity can be measured mathematically in various ways, depending on the formal setting.

The complexity of truth should not come as a surprise, according to the semantic view. Reflection requires stepping out of the language you are speaking, and reflecting on its semantic properties as a whole. That the results turn out to be complex, measured against things you could do in the language, is not surprising (thought just what the degrees of complexity are might be surprising). What is special about the case of self-applicative truth is the additional complexity it creates. We handle that complexity, I suggested, by some form of the long iteration strategy, which provides a complex process of reflection suitable for the task of capturing the semantics of languages with self-applicative truth predicates.

If the long iteration strategy were to be fully successful, we would have a perfectly good theory of truth by lights of the semantic view of truth.[24] Nothing like a hierarchy would ensue. On the other hand, as I stressed in Sect. 10.2, there is no guarantee that any exercise of reflection will be fully successful, and no guarantee such an exercise will produce a complete or otherwise good theory. I argued that the long iteration strategy can produce reasonably good theories, but even so, for something as complex as the long iteration strategy, we might especially wonder if a complete

[24] Or at least, almost. As I mentioned in Sect. 10.1, we would have a perfect theory of truth as applied to one language, which has been the main concern of work on the paradoxes. We still might like to understand better the way truth works across languages.

theory of truth will be the result. We might wonder why such reflection should be able to return a fully correct theory of the semantics of a complex language all at once. If it does not, then the result is a hierarchy. We would have to restart the process of reflection, and generate a further truth predicate. The process could well be open-ended, as we have no guarantee we will ever reach a completely finished product. We would then have a hierarchy of accounts of the semantics of the language, each with a distinct truth predicate. We would indeed have a hierarchy.

Thus, I claim, if we see our theories of truth as the results of complex reflection on the semantics of a language, we should not be surprised if we encounter hierarchies, any more than we should be surprised that such complex tasks can yield incomplete results. Hierarchies, on this view, should not be surprising.

Actually, I believe something stronger. I maintain that reflection, even very good instances which generate plausible theories of truth, must be incomplete, and so, hierarchies are not just unsurprising, they are inevitable. This is, of course, a highly contentious claim. Many of the theories we reviewed in the last section are offered as non-hierarchical theories of self-applicative truth, and spirited defenses of them as such have been offered. This is so for the defenders of theories like KF or FS, such as again Halbach (2011) and Horsten (2011), and those who develop model-theoretic approaches in non-classical settings like the paraconsistent theories of Beall (2009) or Priest (2006) and the paracomplete theory of Field (2008) (all of whom, in one way or another, rely on the long iteration strategy).

My own view is that we cannot avoid hierarchies. I have argued this at length elsewhere, and I shall not try to mount a full defense of the claim here.[25] But it will help make clearer the nature of the hierarchy I think we are stuck with to roughly sketch why I think hierarchies are unavoidable. The main idea is that once we have our semantics in hand, it turns out we can engage in further reflection on how it works, and that leads us to more inclusive truth predicates. Each such reflection introduces a new level of the hierarchy, and the process is open-ended and does not terminate.

Of course, the Liar is doing a lot of work in showing that this further reflection really generates something new. Let us consider how this process of stepping back and seeing how the semantics that we generated in reflection works might go. Suppose we take our semantics to be a Kripke model $\mathfrak{M}^+$, understood as the result of a long iteration process of reflection. Above we noted that this is a pretty plausible semantics, in virtue of the fixed point property. Hence, it seemed that the long iteration strategy produced a successful exercise in reflection. But now we come to the Liar. Observing how the semantics works, we can observe that the Liar sentence is not assigned the value true or false in this model (it gets the third or gap value), and it is not in the extension or anti-extension of Tr. But then, we can observe according to the semantics, the Liar is not true. But of course, this is just what the Liar says, so it appears we have used the semantics to show that the Liar sentence is in fact true.

[25] I discussed this in fairly general semantic terms in Glanzberg (2001), in model-theoretic terms in Glanzberg (2004a), and in proof-theoretic terms, focusing on FS, in Glanzberg (2004c).

But then, we have found our semantics to be inadequate. The results of our initial exercise in reflection was not so successful after all. Likewise, our interpretation of Tr appears to be off, as we have convinced ourselves that the Liar is true, and so should be in the extension of Tr. If we take the closed-off model $\langle \mathfrak{M}, E \rangle$, we get similar results, though in a somewhat different way. The Liar is not in E, hence, the semantics simply says that the Liar sentence is true. But then we have an inadequate semantics, as our interpretation of Tr and the semantics fail to match-up adequately after all. Alternatively, we could note that though the Liar sentence is true, it is also true that $\neg Tr(\ulcorner L \urcorner)$ for the Liar sentence L, and so the semantics in effect denies the truth of the Liar sentence. There are other ways to illustrate the inadequacy of the semantics. For instance, it fails to define negation or some conditionals we need to describe the semantic properties of the Liar fully. In proof-theoretic terms, we get results like $KF \vdash L \wedge \neg Tr(\ulcorner L \urcorner)$ for Liar sentence L. Different ways of implementing the details here will capture the inadequacy differently, but I hope to have made clear why one way or another, our semantics and our account of Tr are not good enough.

Of course, this is just the Strengthened Liar. As I said, the standing of this and other 'revenge' arguments is hotly contested.[26] Also as I said, I have tried to defend a more carefully worked out version of it in other work, and I shall just accept it here. My main point now is to put it in the context of reflection. We have engaged in a process of reflection, which provides us with a seemingly plausible semantics for our language. But the exercise of the Strengthened Liar shows that we can observe how the semantics works, and come to see that it is not fully adequate after all. Just how that inadequacy manifests itself depends on the details, but one way or another, we can see it.

How are we to respond to this sort of problem? From the point of view we have taken here, where the task is one of reflection, the answer is easy. We found that our exercise in reflection yielded good, but we now see not good enough, results. As I have stressed, even without the force of paradoxes, this kind of situation is hardly surprising. And we know what to do when we find ourselves with reasonably good but not good enough theories. We should simply re-start the process of reflection to try to build a better theory. Building one will involve producing a wider truth predicate, and thereby a wider semantics, which will do better than the one we had. To see how this might work, let us again look at the closed-off Kripke model $\langle \mathfrak{M}, E \rangle$. This failed to accurately capture the properties of the Liar sentence, by failing to report the truth of the Liar sentence, which follows from the very semantics. We need to build a new truth predicate, which does better. But we also need to be careful, as simply throwing the Liar sentence into the extension of Tr gets us back into trouble. (Again, just what trouble depends on the details, but in the fully classical setting, we make $Tr(\ulcorner L \urcorner)$ true, which makes the Liar sentence false even though it is reported as true.) One way or another, we cannot simply adjust for the Liar sentence without getting a new paradox back. So, the basic task is clear—we need to re-start the process of reflection

[26] See the papers in Beall (2008) for many different perspectives on revenge paradoxes.

and build a better theory—but just how to do so in a productive way is a difficult task.

My own approach to this task is to rely on our observations about the semantics, as described in the earlier exercise of reflection, as we go forward. Thus, we start with the observation that the Liar comes out true *according to the semantics as we worked it out in the previous exercise of reflection*. What we need is an expanded semantics, which reports this, and so does a better job of making the semantic properties of the language explicit. Here we see the change of topic feature of reflection which was evident for the purely Tarskian case we explored in Sect. 10.2. We are making the product of the previous round of reflection—the (reasonably good) approximation of the semantics—part of the subject-matter of our next round of reflection. In doing so, of course, we shift its role from the being the underlying mechanism of the language we are speaking—the real semantics—to something we are talking about. We can thus build accounts of the semantics of the language which take it into account, but depart from it. Of course, the basic approach to this further exercise of reflection will be the same: it will continue to use the long iteration strategy. But it will do so based on our new subject-matter, including the semantics produced by the previous round of reflection.

It is, unfortunately, somewhat complicated to capture this process formally. I developed a model-theoretic version of it in Glanzberg (2004a), but it relies on some fairly heavy use of definability theory. I shall here try to sketch the main idea with a minimum of formal apparatus. If we take $\langle \mathfrak{M}, E \rangle$ to represent the results of the first round of reflection, we want to repeat the long iteration strategy, taking it as part of the subject-matter. We thus want to re-do the Kripke construction, over the expanded structure $\langle \mathfrak{M}, E \rangle$. Of course, E is no longer interpreting the truth predicate, it is just another predicate. When we re-do the Kripke construction over this expanded model, we get an expanded truth predicate. The new interpretation of Tr includes facts like $Tr(\ulcorner \neg \dot{E}(\ulcorner L \urcorner) \urcorner)$. The new interpretation reports the facts about the semantics as it was before our new round of reflection. It is also much more complex than E, as we might expect. One reason the machinery gets complicated is it is not completely trivial to keep track of such iterated inductive definitions and their complexities, and so we wind up working with the machinery of next admissible ordinals.

The picture that emerges has some Tarskian aspects. As I loosely described it, we have the old truth predicate $\dot{E}$ and a new one Tr. This will help to fix ideas, but in fact, the model of Glanzberg (2004a) is slightly less Tarskian than that. The model does not simply add a new predicate to the language, and the main role of the prior interpretation E is to add complexity to the ground model. The expanded truth predicate allows us to reconstruct the semantics of the prior stage, and define the old truth predicate. So, $\dot{E}$ is definable, and we do not really have to outright change the vocabulary. Contextualism also enters the picture, as I take the ground structures to represent contextually salient elements, and work with domains of truth

conditions relative to contexts. That helps to model the way in which we make the prior semantics a new topic.[27]

With all those details, we get a less Tarskian theory, but one that is still decidedly hierarchical. We can define the old truth predicate via the new one, and so, the old truth predicate is present in the language, though because of definability-theoretic strength rather than outright change of vocabulary. The response to the Strengthened Liar is likewise fundamentally hierarchical, as we say that the Liar sentence is true relative to the semantics as it was at the prior stage. It is true at the lower level of the hierarchy. As I mentioned, this is an effect of the topic-changing nature of reflection, though one that seems to be necessitated by the Strengthened Liar. I shall return to the comparisons between my preferred form of the hierarchy and Tarski's in Sect. 10.5.[28]

Once we go down this road, an open-ended hierarchy ensues, for familiar reasons. The very reasons we found to be unsatisfied with the results of the first round of reflection can be applied again to the new results. They too will not be completely adequate, and we will trigger a new round of reflection. The process is open-ended. (Indeed, if it reaches a genuine top, we get paradoxes back.)

In thinking about reflection and the long iteration strategy, I grouped model-theoretic and proof-theoretic versions together as various ways to represent results of reflection. When it comes to capturing the kind of open-ended hierarchy generated by reflection we have been discussing proof-theoretically, there are few results available, but work of Fujimoto (2011) and Jäger et al. (1999) might be pressed into service. As with the model-theoretic approach, things can get quite complex very quickly, so I shall not explore this idea in any depth. One point is worth mentioning. The extant theories explore iterating KF through appropriate proof-theoretic ordinals (and develop relations with theories of iterated inductive definitions), and hence might be plausible proof-theoretic representations of the kind of hierarchy that I claim we get. Interestingly, they also display the Tarskian features I remarked on in discussing the model-theoretic version. If anything, they do so more starkly. The known theories rely on a binary truth predicate, which in effect indexes the truth predicate to levels of a well-ordering (via some notation system). Thus, it seems that the only ways we know to develop the kind of iteration the Strengthened Liar reveals are at least somewhat Tarskian.

---

[27] I have defended the contextualist aspects of my view in Glanzberg (2001, 2004a, 2006). As I mentioned, contextualism helps with the development of the particular sort of hierarchical view I prefer, but generally, the step from any hierarchical view to contextualism is very small. One need only accept that reflection (or whatever else generates the hierarchy) takes place in real time as we work with and reason about our concepts, and so takes place within contexts. Contexts thus serve to index the stages of reflection. This is the core of the view I have defended, and I believe underlies other contextualist views such as those of Parsons (1974).

[28] A theory close in spirit to mine, but using very different resources, is developed by Barwise and Etchemendy (1987). Iterated Kripke constructions are also discussed by Field (2008), and briefly in the older discussions of Burge (1979) and the postscript to Parsons (1974).

We now have at least a hint of what sort of hierarchy I think emerges for truth, and why. I also hope to have made clear why the hierarchy has some sound motivations and is not absurd on its face. As I have presented things, the hierarchy is a hierarchy of results of reflection. As truth, according to the semantic view, is a complex concept, we should not have particularly expected such reflection to yield fully complete theories in any one exercise. The extreme complexity of self-applicative truth only reinforces this expectation. That we find, after doing a good job of reflection, that we need to do more, is just not surprising or problematic. That is all there is to the hierarchy, and so, I claim, the mere fact that we get some sort of hierarchy is not problematic.

This is a partial defense of the hierarchy, and there remain some important questions. Just because some sort of hierarchy is not repugnant does not mean the one we have is plausible. The Tarskian aspects of the hierarchy I have noted might make us cautious about accepting it, even given the kind of motivation I have offered. I shall go on to discuss these issues in the next section.

There is one final point to make about the general idea that less-than-complete exercises of reflection are to be expected. Even if that is true, it does not address the fact that according to the kind of Strengthened Liar or 'revenge' reasoning I am relying on, such incompleteness is necessary. This shows that the hierarchy is not merely the result of our being fallible beings, with limited abilities to reflect on our own languages. Motivating and defending the hierarchy, as I am doing here, does not explain everything we might want to know about its source and nature.[29]

10.5 Varieties of Stratification

So far, I have argued that if we adopt the semantic view of truth, we naturally get a hierarchy (at least in the face of the Strengthened Liar), and the hierarchy is generally well-motivated. But as I mentioned, the very general kind of motivation I have offered does not show the specific hierarchy we get is plausible. To further defend the hierarchical approach I have been developing, I shall now explore in more detail how well it functions, and how vulnerable it is to objections. I shall argue it is quite successful on these counts, though it is not entirely without costs.

To do this, I shall begin by reviewing some common objections to hierarchies. I shall then discuss a somewhat broader notion of stratification, which is the core feature of hierarchies, but also includes some instances which are not usually labeled 'hierarchies'. I shall show that different forms of stratification make the objections more or less compelling. I shall argue that though the truth hierarchy is not at the completely innocuous end of stratification, it is close enough to evade the objections.

[29] There is something special about 'stepping back' and reflecting about the semantics of your own language that triggers hierarchies. I investigated some aspects of this in a number of papers (e.g. Glanzberg 2006). A very different view is presented in Gauker (2006).

Thus, I hope to offer a more detailed defense of the hierarchy, beyond very general motivations.

One preliminary point is needed. I shall in many cases contrast the sort of hierarchy I proposed in Sect. 10.4 with an orthodox Tarskian one. I take it that the orthodox Tarskian hierarchy is well-known, and I shall not review its structure in detail.[30] But there are a couple of features of it that will become salient. First, the orthodox Tarskian hierarchy introduces new truth predicates at each level. The distinct predicates are indexed by the hierarchy. There is a syntactic restriction that each predicate can only apply to sentences of lower levels of the hierarchy, so each truth predicate can only apply to sentences containing only lower-level truth predicates. The orthodox hierarchy is thus syntactically driven, in the way it separates levels and restricts truth predicates.

Many have found hierarchies of truth predicates so obviously objectionable as to be dismissed without discussion. One reason behind this, I suspect, is what I shall call the *one concept* objection. There certainly seems to be one unified concept of truth, while a hierarchical theory might seem to present us with many distinct concepts of 'truth at some level'. Surely any such analysis must be wrong, the objection goes.

Another set of objections to hierarchies is that they are unnatural or unworkable in some way or another. I shall group these together as the *clumsiness* objection. There are a couple of salient instances of this sort of objection. Perhaps the most important is Kripke's well-known 'Nixon-Dean' objection. Recall, Kripke presented an example where Nixon says 'Everything Dean says about Watergate is false' while Dean says 'Everything Nixon says about Watergate is false'.[31] Kripke rightly points out that the orthodox Tarskian hierarchy simply cannot capture these statements, at any level. The problem is with the fixed syntactic indexing of levels, which precludes reasonable assignments of levels to truth predicates in Nixon's or Dean's sentences. There are other versions of the clumsiness objection. For instance, Halbach (2011) and McGee (1991) complain that we cannot express in a hierarchy certain generalizations that we find theoretically interesting, especially semantic ones. Each version points out things we would like to do or say with truth predicates which seem problematic on hierarchical accounts. The theory is, it seems, too clumsy to work the way it should.

Another objection is the *weakness* objection, which says that whether or not they are clumsy or artificial, the theories we get from hierarchical approaches are just too weak. For instance, they might well be too weak to serve mathematical purposes, as discussed by Halbach (2011).

I shall discuss the one concept objection at some length in a moment. First, I want to address the clumsiness and weakness objections. As careful commentators such as Field (2008) and Halbach (2011) have noted, these objections may have some force against the orthodox Tarskian hierarchy, but they apply to more liberal

[30] See McGee (1991) for a nice presentation of the details, and Halbach (1995) for an in-depth exploration.

[31] For those who find this example dated, John Dean was White House Counsel to Richard Nixon, and went on to testify against Nixon at the Senate Watergate Hearings.

hierarchies, like the one I sketched, only to a very limited extent. Actually, the weakness objection turns out not to be specifically one against hierarchical theories. We have already seen that in terms of pure mathematical power, versions of the Tarski hierarchy and various other non-hierarchical theories (like the Kripke construction or KF) turn out to be equivalent. So, there is no special weakness issue even for the orthodox Tarskian hierarchy. There are, of course, a number of more specific issues. The Tarski hierarchy itself has levels that are weaker in strength than some untyped theories. We see this, for instance, from the fact that CT is much weaker than KF. But this does not pose a problem for the hierarchy I endorse, or any hierarchy that uses the long iteration strategy to build much stronger theories at each level. If anything, the kind of hierarchy I propose indicates much stronger theories than the familiar non-hierarchical ones. All of those (with the notable exception of the revision theory) mathematically correspond to specific levels of the Tarski hierarchy or ramified analysis, while the higher levels of my hierarchy will go beyond them. This is shown, for instance, in the impredicative nature of the iterated KF theory of Fujimoto (2011) and Jäger et al. (1999). So, there is nothing that makes hierarchies themselves a source of mathematical weakness. Rather, I believe the real worry behind the weakness objection is a general one about the state of our current theories of truth, which for the most part turn out to be mathematically weaker than we might require for some applications. The weakness problem is thus, as I see it, not a problem for hierarchies. It is a general problem for all of us working in theories of truth.

When we turn to the clumsiness objections, I believe that the objections are good as applied to the orthodox Tarskian hierarchy, but not to mine. Let us begin with the Nixon-Dean objection, which many (myself included) take to be decisive against the orthodox Tarskian approach. Though I think it is good against the Tarskian hierarchy, it does not apply to mine. The kind of hierarchy I endorse is very coarsely stratified, and so, it makes plenty of room for Nixon-Dean cases. This is made especially clear if we take my model-theoretic version, which builds on the Kripke construction. Thus, my hierarchy has no more trouble with the Nixon-Dean case than Kripke's own theory, and it solves it just the same way he does. Within the fixed point, we have lots of room for the kinds of complicated self-applications of truth the case presents.

Not only is the stratification I propose very course-grained, it does not rely on syntactically defined levels. Thus, the Nixon and Dean sentences are well-formed. Assuming there are no Liar cycles in the Nixon-Dean discourse, they will turn out to be grounded, and so get truth values according to the fixed point. Truth does find its own 'level' (as Kripke says), in the long iteration strategy, and so, the kind of failing Kripke illustrates for the orthodox Tarskian hierarchy are not problems for hierarchies based on long iteration.

There is one way in which we might see a variant Nixon-Dean problem, but I am not sure it is really a problem. If Nixon and Dean were to go in for some Strengthened Liar discourse, we might have to hold there is a shift of levels within their discourse. There are two reasons I am not sure this is a problem. First, setting up this sort of discourse would be very unnatural compared to the original Nixon-Dean case, which though a little contrived, is an extension of a situation that occurs all the time. (In the Watergate scandal, lots of people were called liars!) So, the degree of clumsiness we

might encounter would be at worst very small. But second, in the kind of case we are imagining, I am not sure the hierarchy would not be the right answer anyway. If Nixon and Dean were to create a complex Strengthened Liar, and then walk through the Strengthened Liar reasoning together, the hierarchy gives a very natural explanation of what is happening. Hierarchies may well show some clumsiness, but they also show theoretical naturalness in some cases too. So, as objections to theories go, the clumsiness one has relatively little force against my proposal.

The same can be said for the generalization version of the clumsiness objection. Because the hierarchy is not syntactically based, there is no barrier to forming all the sentence involving truth we might want, including generalizations about truth. The question becomes whether the right ones come out true in our model, or are provable by our theory. Here again, hierarchies can achieve reasonable but limited success. Lots of generalizations about truth come out true, or are provable, in the sorts of hierarchies I discussed in Sect. 10.4. For instance, the closure of truth under modus ponens is true in the minimal fixed point and provable in KF. Now, the kinds of theories we have been discussing all fail to prove every generalization we might find plausible, and some of them prove things we might find implausible, as discussed at length by Field (2008). But notice, whatever problems these limitations really pose, they are problems about the theories we might adopt at various levels, not problems for the kind of hierarchy I offer. We can observe again that, from the perspective of reflection, our theories turning out to be defective in various ways, including missing or misdescribing some generalizations, is not a surprise. It is what our discussion of reflection showed we should expect. The response from the reflection viewpoint is simply to go in for further reflection and try to correct the defects. That is just what hierarchies do, on my view. The current non-hierarchical theories of truth all fail to secure some generalizations, or generate seemingly incorrect ones. Hierarchies give us a natural way to fix this problem. Thus, if anything, in this respect the hierarchical view can secure more generalizations over the long run than the non-hierarchical theories that have been developed to date.

As before, there is still a sense in which the objection applies. There is an absolute complete picture of truth, be it a theory or a semantics, which the hierarchical view says we cannot ever achieve. The points I made above all really point out that if we make the hierarchy coarse-grained and liberal in other respects, each level does as well as many other (non-hierarchical) theories of truth on offer. Of course, those other theories are presumably not the one complete theory of truth or semantics either. When compared to what we actually have, the hierarchy looks just fine. The objection is that the hierarchy makes a further objectionable step, by declaring that this situation ultimately cannot be remedied, by denying there can be one complete final theory of truth. Proponents of the non-hierarchical theories will presumably say their offerings may be incomplete or inaccurate, but they are merely the best that can be had so far, and they are striving for one single absolute theory of truth they have not yet achieved. Hierarchical approaches, including mine, say there is no such thing. My main point all along has been that from the point of view of reflection, this

is not so objectionable after all. But it must be admitted that it is a limitation, and it might be disappointing. If I am right, then it is not really much more than that.[32]

On the weakness and clumsiness objections, I conclude that hierarchies do not fare particularly badly, and might, from the reflection perspective, actually fare reasonably well. They are not perfect, but the role of reflection helps to show why we can live with the problems.[33] Now, let us consider the one concept objection. I am going to argue that as a general point about hierarchies, the one concept objection fails. Not everything that counts as a hierarchy falls prey to the one concept objection. Even so, some do, including the orthodox Tarskian hierarchy. I shall suggest that my own proposal does reasonably well by lights of the one concept problem, but is not without drawbacks. Along the way, we will see how the ways topics can change in reflection can raise questions about one versus many concepts.

To discuss the general one concept problem, I shall introduce a general notion of stratification. A concept is *stratified* if we cannot provide a single theory or definition for it. Instead, we provide a family of related theories or definitions, each of which is systematically connected to others. In effect, a concept is stratified if when we try to analyze it, we wind up with a hierarchy. However, we will see examples of stratification which look rather different than the kinds of hierarchies we have been discussing so far, so a different term seems in order.

The orthodox Tarskian hierarchy of languages, and the kind of hierarchy I have endorsed, are both examples of stratified theories of the concept of truth. The Tarskian hierarchy offers a family of truth predicates, though they are systematically related. (For instance, except for indexing of levels, we have the same definition or theory at each level.) My own suggestion here is also a family of definitions or theories of truth (thought of as the result of reflection on the semantics of the language). Again, they show the same systematic relations, and we move from level to level by stepping back and reflecting on the results at the previous stage.

The hierarchies we have considered, both my own variant and the Tarskian original, present truth as a stratified concept. But this is not the only way stratification can arise. For comparison purposes, let me mention two other cases. First, the case of mathematical proof, which I have discussed at length in Glanzberg (2004c). Recall that the incompleteness phenomena show us that concepts of mathematical proof are typically stratified. Suppose we begin with some formal theory for some reasonable piece of mathematics, say PA as a theory of arithmetic. This gives us an articulation of a concept of mathematical proof, in the specific domain of arithmetic. Furthermore, we might think of it as the result of reflection upon our mathematical practices, restricted to arithmetic. We know from the incompleteness theorems that neither PA

[32] In the background here is the issue of absolute generality, as hierarchies one way or another deny the possibility of expressing absolute generality. The attitude I am taking here about generalizations is much the same as the one I took about absolute generality in Glanzberg (2004b, 2006). The other papers in Rayo and Uzquiano (2006) will give a good indication of the state of that debate.

[33] The careful discussions of hierarchies in Field (2008) and Halbach (2011) come to conclusions about these objections somewhat similar to mine, but disagree sharply on how well we can live with the problems we do find.

nor any other reasonably good, recursively axiomatizable, theory of arithmetic will be a complete theory of arithmetic. Insofar as we have an implicit concept of proof in arithmetic and articulate it by reflection as a theory like PA, we have not completely articulated our concept. As we have come to expect, the task of reflection does not always yield complete theories. But for proof, we can say more about what is left out. We also know from the incompleteness theorems that statements of consistency or soundness are not provable in the theory. What is left out, at least, is the fact that the theory is *correct*.

Kreisel (1970) observed that this sort of fact is really implicit in our accepting or using the theory. He writes (Kreisel 1970, p. 489), "What principles of proof do we recognize as valid once we have understood (or, as one sometimes says, 'accepted') certain given concepts?" His answer to the question is that statements of soundness are so recognized. Transposing Kreisel's proposal into the setting we are working in here, we start with an implicit concept of proof for arithmetic, and reflect upon it and produce an articulation, yielding a theory like PA. But we do not merely reflect as an exercise in theory construction. We make something that was implicit explicit. Insofar as we have a practice of doing arithmetic, we make its features explicit by articulating PA, and we also step into a practice of doing arithmetic formally, in PA. We thus use the theory, in a practice much like the one we had before but now more formally articulated. As Kreisel notes, doing so implicitly commits us to the soundness—the correctness—of our articulation. We are thus implicitly committed to the soundness of PA. That is not explicit, as it is not a consequence of PA, but it is implicit.

This further implicit content is available to reflection, which produces a further theory to capture something like PA together with its soundness. For theories like PA, in fact, we can make the soundness of the theory explicit as a *reflection principle*:[34]

$$Prov_{PA}(\ulcorner \phi \dot{\overline{x}} \urcorner) \rightarrow \phi x.$$

Our new articulation is the theory PA plus the reflection principle RFN_{PA} for PA. This is a stronger theory.

We know not only that the result is a stronger theory, but also that the process we just started is open-ended. The new theory $PA + RFN_{PA}$ is (of course) incomplete, and we can again engage in just the same kind of reflection and produce a theory which adds an additional reflection principle for the new theory. The result is an open-ended family of theories, each a stronger articulation of the concept of proof in arithmetic with which we began.[35]

[34] The connection between 'reflection' and 'reflection principle' is probably no accident, and I am certainly exploiting it. This is the 'uniform reflection principle'. For more on reflection principles, see Feferman (1962), and Kreisel and Lévy (1968).

[35] It is natural to ask if we can get a single complete theory by iterating this process to a suitable end-point. Results in this area are quite subtle, but substantially negative. See the papers of Feferman (1962), Feferman and Spector (1962), and Visser (1981).

There are a number of ways one can extract morals from this case.[36] For our purposes, let me highlight what I take to be the important aspects of the case. Reflection generates a natural stratified analysis of the concept of proof (mathematical proof, e.g. for arithmetic). Further results show that such analyses are unavoidable, and so, the concept really is stratified; but the role of reflection helps us to see how that situation can emerge and is natural. The kind of stratification we see here, where a single concept is only analyzable by a related family of theories, is one way stratification can occur. To give it a name, let us call it *Kreiselian stratification*, to highlight Kreisel's insight about what is implicit in accepting a theory.

The case of Kreiselian stratification is important for turning aside the one concept objection. Concepts of mathematical proof are stratified, and the Kreiselian view makes them in important respects similar to the case of truth. Not only are they stratified, but the stratification is evident in a process of reflectively articulating stronger and stronger theories. Moreover, as with the truth case, some deep underlying phenomenon, in this case Gödelian incompleteness and other results, guarantee that stratification is unavoidable.

Though we find proof to be stratified, it does not appear to be vulnerable to the one concept objection. What we have is a family of *theories*, all of which are recognizably articulations of the concept of proof in arithmetic. They differ in strength, not subject-matter. Now admittedly, I am not offering a worked-out criterion for when we have distinct concepts versus distinct theories of the same concept. We can, of course, identify specific concepts going with our theories, like proof in PA, proof in $PA + RFN_{PA}$, etc. But these all seem clearly to be sub-concepts which simply map to theories. At the risk of appealing to a brute intuition, I think we can count this as a case of one concept with multiple theories. If that is right, then the mere presence of stratification does not indicate multiple concepts; at least, Kreiselian stratification does not. So, the one concept objection does not succeed as a general objection to stratification.

Once one goes looking for stratification like the Kreiselian variety, it is not that hard to find. Set theory, for instance, provides another example. If we start with a good set theory (say ZFC), we can build up stronger and stronger theories. One way to do this is to add more and more large cardinal axioms, which in many cases also flow from what are called reflection principles. These are different from proof-theoretic reflection principles, and express the idea that the universe of sets is maximally large. We might even tell a story about how this is implicit in the concept of set. I shall not explore this is detail, but simply note the appeal of the same intuition as in the Kreiselian case, that we have stratification but one concept.[37]

The one concept objection fails in general, as not all instances of stratification are unacceptable multiplication of concepts. At the same time, it appears the orthodox

[36] I discussed the connection with Hilbert's program, and some specific issues about revenge paradoxes, at greater length in Glanzberg (2004c).

[37] Actually, I think a good story can be told here. Insofar as accounts like the iterative conception of set help to identify our concept of set, then we can observe how the multiple theories all express the iterative conception. Thus, I believe we can explain why we really have one concept.

Tarskian hierarchy looks rather bad by lights of the one concept objection. After all, each new language with a new truth predicate seems to offer a new concept, and the sense in which they are all related is one which is highly implicit in the resulting hierarchy. On further inspection, however, it turns out the difference between the Kreiselian and Tarskian cases is not all that easy to specify. In the Kreiselian case, we also have distinct predicates. There are distinct predicates $Prov_{PA}$, $Prov_{PA+RFN_{PA}}$, etc., and these figure crucially in the theories we write down. So, what is the important difference? One is that for arithmetic, all these predicates are definable, and we do not need to jump to a new language altogether, even if we define and use new predicates which were not definable in the old theory.

It appears one crucial aspect of the difference between Kreiselian and Tarskian cases is that in the Tarskian one, we have to jump to a new language. That does not 'feel' like just articulating a new theory, which generally does not involve shifts in the signature of the language. The difference is sometimes hard to specify, as with increases of strength of theory, we can get ability to define things in the old language that we could not before, which is a lot like expanding the language. I believe that the phenomenon we are observing is the same one we talked about in terms of changing the topic or subject-matter in reflection. Reflection tends to do this in some way, as witnessed by the availability of new predicates, definable or not. But some exercises of reflection require a much more pronounced change of topic, perhaps going with a wholesale change of language. I do not have a full analysis of this difference, but I think we can use it, and the general difference between the Kreiselian and Tarskian cases, to try to measure the force of the one concept objection against the kind of hierarchy I sketched in Sect. 10.4.

Let us now return to this sort of hierarchy. I shall argue that it is much less vulnerable to the one concept objection than the orthodox Tarskian one. But unfortunately, I doubt it will be possible to see the truth hierarchy, even in the form I prefer, as purely Kreiselian. There are a number of reasons for this. In the proof-theoretic variant, we might have hoped that we could get something very Kreiselian, by capturing the levels in the hierarchy by iterating a genuine proof-theoretic reflection principle. In some cases, this is simply not possible. It is known, for instance, that $FS + RFN_{FS}$ is *inconsistent* (Halbach 1994, 2011). This is due to the ω-inconsistency of FS, and so the result is very specific to FS. It is not known to apply to KF, for instance. But all the same, it shows that adding proof-theoretic reflection principles is not an automatically available route. As I mentioned in Sect. 10.4, the developments of iterated KF theories take a very different route, relying on a parameterized truth predicate.

The proof-theoretic situation shows that technically, we cannot take genuine Kreiselian structure for granted, and it might not be available. But beyond the technicalities, it also illustrates a genuinely non-Kreiselian feature of the hierarchy. The kind of reflection involved in moving between levels of the hierarchy is more complicated than merely accepting the correctness of the theory we had. If it were not, then adding a proof-theoretic reflection principle would not be able to make the trouble it does for FS.

We see this in the semantic versions of the hierarchy as well, and in the general motivation for the hierarchy I offered. We get the hierarchy because even when we are done with long iteration, we can step back and observe features of our semantics; especially, we can observe Strengthened Liar effects, and other inadequacies of the semantics. That leads us to take the semantics as we developed it as part of the subject-matter for further reflection. I discussed, in very general terms, how that might look in a model theory, and we also glanced at proof-theoretic versions. These steps, which make the semantics itself a topic of investigation, seem to be unlike what we saw in the Kreiselian story. They are not accepting the correctness of our theory, they are rather noting inadequacies of it, and modifying it.

Not only is this unlike the Kreiselian case, it has some important features in common with the Tarskian one. We do not literally have to expand the language, but we do something near to it. In the model-theoretic variant, we treat the prior truth predicate as a distinct predicate, not representing the semantics of the language. Even if the signature of the language does not change, this seems to be allowing it to be a distinct topic from the semantics of the language.

Thus, in terms of shifting the topic, the hierarchy I have proposed is not fully Kreiselian, and shows some Tarskian tendencies. But it still is not fully Tarskian stratification. Most importantly, at any level, we have only one genuine truth predicate. Though we can define other predicates that capture truth in the semantics of prior levels, we never have multiple genuine truth predicates at any level. It thus appears that the hierarchy I have proposed falls in-between Kreiselian and Tarskian hierarchies in the kind of stratification it proposes.[38]

It is worth mentioning that this is a point of contrast between the steps of reflection where we genuinely move up in level of the hierarchy, and the kind of internal steps of reflection that are part of the long iteration strategy. The latter are much more Kreiselian. In the steps of long iteration, we continually refine the one semantic predicate we are working with, and build better and better theories. (The monotonicity of the Kripke jump confirms this, for instance.) But the long iteration strategy does not generate the hierarchy. It is the less Kreiselian steps of reflection that mark the significant levels of the truth hierarchy.

Where does this leave my proposal with respect to the one concept objection? If Kreiselian stratification is completely immune to the objection, and Tarskian stratification vulnerable, then my own proposal falls somewhere in-between. The fact that it does not use multiple truth predicates is a reason to think that it models one concept of truth, and the processes we use to generate the levels is uniformly one of reflecting on the semantic properties of the language. Insofar as the semantic view says that is where the nature of truth is to be found, nothing in the apparatus or the

[38] I thus depart from the position I took in Glanzberg (2004c). I do still hold most of what I said there; particularly, that the comparison with Kreiselian stratification helps to show why the hierarchical nature of truth is unobjectionable. But, in that paper, I was more optimistic about how close the analogy between the Kreiselian case and truth could be drawn.

way it is deployed really suggests it offers multiple concepts of truth. To that extent, it is not as badly off as the traditional Tarskian approach.

Where it is vulnerable is that even so, it does recognize distinct semantics at different stages. We may not have multiple concepts of truth, but we do have multiple representations of the semantic properties of the language we speak. We have, as it were, distinct concepts of 'truth as it appeared at a give stage of reflection'. Those are multiple concepts. My own view is that when we think about the semantic view of truth and the process of reflection, we should not be too unhappy about having those multiple concepts, as they arise from the kind of constrained processes of reflection we can engage in. They do not, strictly speaking, show truth proper to be anything other than one concept—it is our attempts at reflection which fragment, not truth. But all the same, this shows a way in which the hierarchical nature of my proposal is substantial.

I conclude that a hierarchical theory of the sort I have sketched is not without costs, but it also has benefits. On objections like the weakness and clumsiness ones, it actually fares quite well. It is no more vulnerable to them than any other theories currently on the market, and when it comes to Strengthened Liar sorts of cases, it actually looks better and more natural than other non-hierarchical options. On the one concept objection, it does show some non-trivial Tarski-like effects of stratification, but not so much as to undermine its status as the results of reflection on one single concept of truth. Though I do grant this is a cost, it also has the benefit of allowing the kind of explanation of Strengthened Liar cases I think is appealing and natural. When combined with the kind of motivation for the hierarchy the semantic view of truth provides, it seems to me the benefits outweigh the costs.

The main cost of the hierarchy is its failing to provide a single complete theory of truth. Again, I grant this is a cost, as such a theory would be very nice to have. But as the discussion of the Kreiselian stratification shows, in many cases we find not achieving that goal not to be such a high cost after all. More importantly, when we measure the hierarchical approach not against our desires for complete theories—a desire that often cannot be fulfilled—but rather against other real options, hierarchies come out looking surprisingly good. I know of no approach to the paradoxes which does not have some costs that are hard to swallow (otherwise we would hardly call them paradoxes!), but I claim that hierarchies have fewer costs, and enjoy more solid motivations, than it is often supposed. We should accept the hierarchical theory.

10.6 Comparison with Deflationism

To conclude, I shall return briefly to the issue of the nature of truth with which I began. I have argued that if we assume a semantic view of truth, hierarchies are a defensible approach to the paradoxes. Most importantly, as I have argued throughout this paper, the semantic view of truth provides a solid motivation for hierarchies. The semantic view of truth makes the process of reflection substantial, and allows for a complex concept of truth. That, in turn, was the engine that produces hierarchies according to

my proposal. Moreover, I have argued that the kind of hierarchy reflection generates is one that is not vulnerable to a number of objections to the orthodox Tarskian hierarchy, and generally, hierarchies of my preferred sort provide useful and workable theories.

All this assumes the semantic view. If you adopt a deflationist view, then the results are very different. I shall not make any claim about whether deflationists can accept hierarchies in general, but virtually everything I proposed here about how hierarchies are generated, what they describe, and why they are plausible, is unavailable to a deflationist. My defense of the hierarchy is not one a deflationist can use.

First and foremost, if you adopt a deflationist view, then there is no underlying nature of truth, which we might implicitly grasp, and which we might make explicit via reflection. Truth is not a substantial semantic concept. As I noted in Sect. 10.1, it is a simple one with transparent logical properties which are all there is to the nature of the concept. If there is no room for reflection, then the story I told about how hierarchies are generated does not get off the ground. Moreover, the defense of hierarchies I provided does not either. I repeatedly noted how for a very complex concept like truth, we should not expect reflection to generate complete theories, and hence hierarchies are a natural result. But for many deflationists, like the transparency theorists I mentioned in Sect. 10.1, truth is a simple property, and the intersubstitutability of $Tr(\ulcorner\phi\urcorner)$ and ϕ in non-opaque contexts virtually exhausts what we need to say about it. The defense of hierarchies via complexity is thus not available.[39] There is little reason to think such a concept should show even Kreiselian stratification, much less the sort of I described for truth. Thus, the defense against the one concept objection is not available either. If you are a deflationist, none of the important parts of the defense of hierarchies I have offered here will be available.[40]

If one starts with deflationist views of truth, then perhaps the hierarchy just looks unacceptable. At least, the defense of it I offered here is not available. But, as I have tried to show, if one starts with the semantic view, then the hierarchy is a reasonably workable, natural, and well-motivated approach to truth and paradox.

[39] Again, there are formal results to back this up. A natural deflationist analog to CT simply uses the Tarski biconditionals rather than the CT axioms. This theory is a conservative extension of PA, as Halbach (2011) discusses.

[40] Though I do not have the space to pursue the matter here, let me mention briefly that I suspect this is why Field (2008) endorses a hierarchy of determinacy operators but not a hierarchy of truth predicates. There are some important issues here, but in very crude terms, if you adopt the semantic perspective, not only can you offer the defense of truth hierarchies I have, you will also find determinacy operators to look a lot like they try to capture a notion of truth. If you adopt Field's own transparency view of truth, on the other hand, thy are clearly distinct conceptually, and the hierarchy seems unnatural and unmotivated for truth.

References

Barrio, E. A. (2006). Theories of truth without standard models and Yablo's sequences. *Studia Logica, 82,* 1–17.

Barwise, J. (1975). *Admissible sets and structures*. Berlin: Springer-Verlag.

Barwise, J., & Etchemendy, J. (1987). *The Liar*. Oxford: Oxford University Press.

Beall, Jc. (Ed.). (2008). *Revenge of the Liar*. Oxford: Oxford University Press.

Beall, Jc. (2009). *Spandrels of truth*. Oxford: Oxford University Press.

Beall, Jc., & Glanzberg, M. (2008). Where the paths meet: Remarks on truth and paradox. In P. A. French & H. K. Wettstein (Eds.), *Midwest studies in philosophy volume XXXII: Truth and its deformities* (pp. 169–198). Boston: Wiley-Blackwell.

Burge, T. (1979). Semantical paradox. *Journal of Philosophy, 76,* 169–198 (Reprinted in Martin 1984).

Cantini, A. (1996). *Logical frameworks for truth and abstraction: An axiomatic study*. Amsterdam: Elsevier.

Carey, S. (2009). *The origin of concepts*. Oxford: Oxford University Press.

Chomsky, N. (1980). *Rules and representations*. New York: Columbia University Press.

David, M. (1994). *Correspondence and disquotation*. Oxford: Oxford University Press.

Davidson, D. (1967). Truth and meaning. *Synthese, 17,* 304–323 (Reprinted in Davidson 2001).

Davidson, D. (1977). Reality without reference. *Dialectica, 31,* 247–253 (Reprinted in Davidson 2001).

Davidson, D. (1990). The structure and content of truth. *Journal of Philosophy, 87,* 279–328 (Reprinted in revised form in Davidson 2005).

Davidson, D. (2001). *Inquiries into truth and interpretation (2nd ed).* Oxford: Oxford University Press.

Davidson, D. (2005). *Truth and predication.* Cambridge: Harvard University Press.

Davies, M. (1987). Tacit knowledge and semantic theory: Can a five percent difference matter? *Mind, 96,* 441–462.

Etchemendy, J. (1988). Tarski on truth and logical consequence. *Journal of Symbolic Logic, 53,* 51–79.

Feferman, S. (1962). Transfinite recursive progressions of axiomatic theories. *Journal of Symbolic Logic, 27,* 259–316.

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

Feferman, S., & Spector, C. (1962). Incompleteness along paths in progressions of theories. *Journal of Symbolic Logic, 27,* 383–390.

Field, H. (1972). Tarski's theory of truth. *Journal of Philosophy, 69,* 347–375.

Field, H. (1986). The deflationary conception of truth. In C. Wright & G. MacDonald (Eds.), *Fact, science and value* (pp. 55–117). Oxford: Basil Blackwell.

Field, H. (1994). Deflationist views of meaning and content. *Mind, 103,* 249–285.

Field, H. (2008). *Saving truth from paradox*. Oxford: Oxford University Press.

Friedman, H., & Sheard, M. (1987). An axiomatic approach to self-referential truth. *Annals of Pure and Applied Logic, 33,* 1–21.

Fujimoto, K. (2011). Autonomous progression and transfinite iteration of self-applicable truth. *Journal of Symbolic Logic, 76,* 914–945.

Gaifman, H. (1992). Pointers to truth. *Journal of Philosophy, 89,* 223–261.

Gauker, C. (2006). Against stepping back: A critique of contextualist approaches to the semantic paradoxes. *Journal of Philosophical Logic, 35,* 393–422.

Glanzberg, M. (2001). The Liar in context. *Philosophical Studies, 103,* 217–251.

Glanzberg, M. (2002). Topic and discourse. *Mind and Language, 17,* 333–375.

Glanzberg, M. (2004a). A contextual-hierarchical approach to truth and the Liar paradox. *Journal of Philosophical Logic, 33,* 27–88.

Glanzberg, M. (2004b). Quantification and realism. *Philosophy and Phenomenological Research, 69,* 541–572.

Glanzberg, M. (2004c). Truth, reflection, and hierarchies. *Synthese, 142,* 289–315.
Glanzberg, M. (2006). Context and unrestricted quantification. In A. Rayo & G. Uzquiano (Eds.), *Absolute generality* (pp. 45–74). Oxford: Oxford University Press.
Gupta, A., & Belnap, N. (1993). *The revision theory of truth*. Cambridge: MIT Press.
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. (1997). Tarskian and Kripean truth. *Journal of Philosophical Logic, 26,* 69–80.
Halbach, V. (2011). *Axiomatic theories of truth*. Cambridge: Cambridge University Press.
Higginbotham, J. (1989). Knowledge of reference. In A. George (Ed.), *Reflections on chomsky* (pp. 153–174). Oxford: Basil Blackwell.
Horsten, L. (2011). *The Tarskian turn*. Cambridge: MIT Press.
Jäger, G., Kahle, R., Setzer, A., & Strahm, T. (1999). The proof-theoretic analysis of transfinitely iterated fixed point theories. *Journal of Symbolic Logic, 64,* 53–67.
Kreisel, G. (1970). Principles of proof and ordinals implicit in given concepts. In A. Kino, J. Myhill, & R. E. Vesley (Eds.), *Intuitionism and proof theory* (pp. 489–516). Amsterdam: North-Holland.
Kreisel, G., & Lévy, A. (1968). Reflection principles and their use for establishing the complexity of axiomatic systems. *Zeitschrift für mathematische Logik und Grundlagen der Mathematik,* 14*,* 97–142.
Kripke, S. (1975). Outline of a theory of truth. *Journal of Philosophy, 72,* 690–716 (Reprinted in Martin 1984).
Leeds, S. (1978). Theories of reference and truth. *Erkenntnis, 13,* 111–129.
Martin, R. M. (Ed.). (1984). *Recent essays on truth and the Liar paradox.* Oxford: Oxford University Press.
McGee, V. (1991). *Truth, vagueness, and paradox*. Indianapolis: Hackett.
Moore, G. E. (1953). *Some main problems of philosophy*. London: George Allen and Unwin.
Moschovakis, Y. N. (1974). *Elementary induction on abstract structures*. Amsterdam: North-Holland.
Parsons, C. (1974). The Liar paradox. *Journal of Philosophical Logic, 3,* 381–412 (Reprinted in Parsons 1983).
Parsons, C. (1983). *Mathematics in philosophy.* Ithaca: Cornell University Press.
Priest, G. (2006). *In contradiction* (2nd ed). Oxford: Oxford University Press.
Quine, W. V. (1970). *Philosophy of logic*. Cambridge: Harvard University Press.
Rayo, A., & Uzquiano, G. (Eds.). (2006). *Absolute generality*. Oxford: Oxford University Press.
Reinhardt, W. N. (1986). Some remarks on extending and interpreting theories with a partial predicate for truth. *Journal of Philosophical Logic, 15,* 219–251.
Russell, B. (1912). *The problems of philosophy*. London: Oxford University Press.
Scholl, B. J., & Tremoulet, P. D. (2000). Perceptual causality and animacy. *Trends in Cognitive Science, 4,* 299–309.
Soames, S. (1984). What is a theory of truth? *Journal of Philosophy, 81,* 411–429.
Tarski, A. (1935). Der Wahrheitsbegriff in den formalizierten Sprachen. *Studia Philosophica*, 1:261–405. References are to the translation by J. H. Woodger as "The concept of truth in formalized languages" in Tarski 1983. Original Polish version published in 1933.
Tarski, A. (1983). *Logic, semantics, metamathematics (2nd ed).* Indianapolis: Hackett. Edited by J. Corcoran with translations by J. H. Woodger.
van Fraassen, B. C. (1968). Presupposition, implication, and self-reference. *Journal of Philosophy, 65,* 136–152.
van Fraassen, B. C. (1970). Truth and paradoxical consequence. In R. L. Martin (Ed.), *Paradox of the Liar* (pp. 13–23). Atascadero: Ridgeview.
Visser, A. (1981). An incompleteness result for paths through or within $\mathcal{O}$. Nederlandse Akademie van Wetenschappen. Proceedings. Series A. Mathematical Sciences. 43:237–243.

Chapter 11
Can Deflationism Account for the Norm of Truth?

Pascal Engel

Abstract Deflationism about truth has to deny that there is a norm of truth which governs assertion and belief. This article examines two strategies that the deflationist can take. The first is a form of error-theory: there is no such thing as a norm for assertion and belief. Against this argue that if the deflationist accepts that there is no more to a belief or an assertion being correct than the belief or assertion being true, the deflationist has no account of the correctness of belief or of assertion. The second strategy for the deflationist consists in accepting the correctness feature, but in denying that this feature carries any weight. I argue that this strategy too fails. Although my defense of this claim is here purely negative, truth has a normative import, and the norm of truth is a substantive property attached to truth.

11.1 Introduction

It is a common place in contemporary philosophy that truth involves a normative dimension, at least in the sense that truth is the criterion of correctness of belief and assertion, perhaps in other ways related to our epistemic goals and values. The exact nature of this normative involvement is a matter of dispute, but it is *prima facie* in conflict with the idea, promoted by deflationism, that there is no more to truth than the standard equivalence principle

(E) $< p >$ is true if and only if p [1]

If deflationism is correct, neither truth nor the concept of truth contain, or involve, such a normative dimension in any essential way. Conversely, if there is such a

[1] Horwich 1990/1998 takes $< p >$ to range over propositions. Disquotational versions take it to range over sentences. The difference will not matter here, although I shall stick to Horwich's version.

P. Engel
EHESS, Paris, France
e-mail: pascal.engel@ehess.fr

T. Achourioti et al. (Eds.), *Unifying the Philosophy of Truth,* Logic, Epistemology, and the Unity of Science 36, DOI 10.1007/978-94-017-9673-6_11

dimension, and if it is in some sense essential or intrinsic to the nature or to the concept of truth, then the deflationist's claim to be able to explain all features of truth will be defeated.

Deflationism comes in various forms. I shall here leave aside its different versions, and shall take it to be the conjunction of the three following claims:

(D1) there is complete cognitive equivalence between $< p >$ *is true* and p

(D2) conforming to that equivalence is all that is required to manifest complete understanding of the truth predicate.

(D2) the truth predicate is purely logical: it is a mere syntactical device for indirect reference and generalization.[2]

I shall not here deal with the familiar issues raised by this view: whether it can account for all of our linguistic uses of the truth predicate, whether it can account for the usual sorts of substantive properties associated to truth, such as correspondence or objectivity. I shall only here be concerned with the deflationist's thesis that the *prima facie* normative import of truth is either inexistent or shallow, and therefore not significant. Faced with the claim that "true" entails "correct" for our beliefs and assertions and that correctness is a normative notion, a deflationist can react in two ways: either (s)he can deny that the feature in question tracks any genuine property (let us call this, for reasons to be made clear in § 1, the strategy from error) or (s)he can accept that truth or our concept of truth have normative implications, but deny that these carry any weight (let us call this the concessive strategy). I shall examine each strategy in turn, and shall argue that each of them fails. My conclusion here will be limited to this negative point, and to the defense of the claim that truth *has* a normative import, hence that the norm of truth is a substantive property attached to truth. There are different versions of this claim, but the one which I defend is that there is a *constitutive* norm of truth for belief and assertion. I shall not, however, try to articulate this view here in detail[3]. My main objective here is to argue that the deflationist has not given us any good ground to reject this view.

11.2 Correctness and the Argument from Error

The deflationist says that there is no more to truth than (E), which exhausts all that there is to say about truth. Truth is merely a syntactic device of assertion allowing us to generate a potentially infinite set of sentences of the form $< p >$ *is true*. A striking consequence of this claim is that there is no common property of truth which these sentences share, and no explanation of what distinguishes truths for falsehoods (David 1994, pp. 65–66). *A fortiori* any other property than this purely formal or syntactic property is either inexistent—a pseudo property—or can in some way be reduced to (E).

---

[2] I borrow this formulation from Blackburn 2012. In substance, these claims are those defended with an exemplary clarity by Horwich 1990/1998. For reviews of various versions of deflationism, see David 1994, Hawhorne and Oppy 1997, Field 2001, Engel 2002, Hill 2002, Stoljar 2010.

[3] For more, see Engel 2005, 2007, 2013, 2013a, b.

There is indeed nothing normative in (E). Neither is there if one holds a conception of truth as correspondence to the facts. That *p* is true if and only if *p* corresponds to the facts is a purely descriptive property of truth. There is, however, a meaning of "true" as "correct" in the sense in which truth is predicated of an object, as when one talks of "a true friend" or "a true work of art". This meaning is traditionally associated with the medieval notion of *veritas in rebus*. Aquinas thus says that when our thought agrees with reality, the *thing (res)* serves as the *mensura* and *regula* (measure and rule) of the thought. Hegel even says that in his sense, that

> Truth consists in the conformity of objectivity with the notion. . . . It is in this deeper sense of truth that we speak of a true State, or of a true work of art. Objects are true, if they are as they ought to be, i.e. if their reality conforms to their notion. . . .. In common life correctness and truth are very often taken to be synonyms.[4]

According to Hegel an F is true in the objectual sense if and only if something satisfies the highest standards for being an F or if the F is as it ought to be. Indeed in this sense "true" means "correct" (*richtig*). But that can hardly apply to the ordinary notion of *propositional* truth, the truth of a proposition of a thought. Deflationists and their critics can agree on the fact that the truth or falsity of a proposition is a descriptive property. What is in question in the debate between the deflationist and his critic is not that truth is a property of propositions, but that it is associated to assertion and to belief.

Let us start with assertion. To assert that *p* is to present *p* as true and to assert that *p* correctly is to assert *p* only if *p* is true. In other words, truth is the correctness norm of assertion. Even Frege, who can be considered as the contemporary father of deflationism, when he said that " It is really by using the form of an assertoric sentence that we assert truth, and to do this we do not need to use the word 'true"' (1897, p. 129), established the connection between truth and the correctness of assertion, through the following reasoning (Rozenkranz 2010, pp. 225–226). To assert (or to judge) that *p* is correct only if *p* has the value *true*. So we can derive from the truth-value of a proposition whether an assertion which has that proposition as its content is correct. Similarly, we can derive from the correctness of the joint assertion of P and $P \rightarrow Q$ that an assertion of $\sim Q$ will be incorrect: since the joint assertion of P and $P \rightarrow Q$ is correct only if both P and $P \rightarrow Q$ have the truth-value *True*, and the latter requires that Q also have the truth-value *True*, and since Q and $\sim Q$ cannot both have that truth-value, anyone who asserts $\sim Q$ will present $\sim Q$ as having a truth-value that it does not have. In this way, the laws of truth can be used to derive norms of judgment and assertion (cf. Frege 1897, pp. 38–39, 64, 69). Truth is the correctness norm of assertion, and since the laws of logic are the laws of truth, the laws of truth are the laws of the correctness of assertion.

Take now belief. Beliefs can be correct or incorrect, and the standard of correctness for a belief is the truth of the belief's propositional content: a belief whose content is

[4] Aquinas, *Summa theol. Ia*, q. 16, a. 1; q. 21, a. 2, Hegel, *Enzyklopädie, Wissenschaft der Logik* [1830, Suhrkamp Verlag 1986, Band 8], § 213 and n.; § 172; § 24 n. 2, quoted by Künne 2003, 10-3-104. Thanks to Davide Fassio for the reference.

true is a correct belief, and a belief whose content is false is an incorrect belief. The standard of correctness for belief is *normative*: correct belief has a positive normative status, and incorrect belief has a negative normative status. So truth is the correctness norm of belief.

Truth is indeed not the only norm of belief and of assertions. Assertions and beliefs are also supposed to be correct only if they are asserted or believed for good reasons, and because there is sufficient evidence for them. Perhaps there are other norms as well, but the point is that there is at least one norm, and that truth seems to be the main norm or standard. We thus have the following norms for belief and for assertion:

(CC) A belief (assertion) is correct if it is true
(NT) One ought to believe (assert) a proposition if it is true
(NE) One ought to believe (assert) a proposition on the basis of sufficient evidence[5]

Should we, however take these normative liaisons at face value? For deflationism these liaisons are dangerous, since they mean that there is at least one property common to all our ascriptions of truth above the syntactic feature encapsulated in (E).

The most radical answer that the deflationist can give is simply to deny the claim that beliefs or assertions are correct when true and incorrect when false, or that truth is the standard of correctness of belief or of assertion. On this view, which is analogous to the one defended by irrealists about moral values in ethics, such as Mackie (1977), it is simply an *error* to see anything normative in such claims. They are actually only descriptive of what beliefs and assertions are, and carry no normative weight at all.

The deflationist can simply try to reduce the *prima facie* property that beliefs and assertions are correct or not to the fact that they are true or false. On this view "correct" simply means "true", and "incorrect" simply means "false". Beliefs have contents which can be true or false. The same can be said about assertions. The schoolmen used to say that truth is the "formal" object of belief and other intentional states[6]. In Anscombe's adaptation of this old idea beliefs have a "mind-to-world" direction of fit, unlike desires. Beliefs are supposed to be true but can be false, desires are supposed to be realized but can remain unrealized. But a "direction of fit" is not a normative property; it is a purely descriptive one. It is just another way of saying that beliefs are true or false. As Fred Dretske says:

> I agree that beliefs are necessarily true or false. If I didn't understand what it was to be true or false, I could hardly understand what it was to be a belief. But I do not see that I need go further than this. This seems like enough to distinguish beliefs from other mental states like wishes, desires, hopes, doubts, and pains [. . .]. Why, in order to understand what a belief is, do I also have to think of a belief as something that is supposed to be true? If I deliberately deceive you, is the resulting belief supposed to be true? (Dretske 2001, p. 248)

[5] These principles could be strengthened by making them biconditionals, but the present formulation will suffice for our present purposes.
[6] See Kenny 1963, p. 189.

So the deflationist is denouncing here an inference of an *ought* from an *is*. He tells us that from

(i) John's belief that *p* is true

we cannot infer

(ii) John belief's is correct

in any other sense than a purely redundant one. (i) is indeed non normative. (ii) is supposed to be normative. But for the deflationist it is not. The move is familiar from the moral case. It is just an error to read "correct" as a normative predicate. If we understand "correct" in (ii) as just meaning "true", then everything is in place.

But this is strange. For (ii) does seem to mean something distinct from (i). If "correct" meant the same thing as "true", then we could replace "true" by "correct" with all intentional states to which we can predicate "true". Thus in many contexts, to imagine that *p* is to imagine that *p* is true. If "true" were synonym of "correct" we could say such things as "He imagined that he had been elected to the Waynflete chair at Oxford, but his imagining was incorrect". Or we can say "He hopes to be elected, but his hope will turn out false" but it sounds odd to say "He hopes to be elected but his hope is incorrect". "Correct" is not used in these contexts[7]. So it cannot mean the same as "true".

Moreover if the deflationist intends to be an error theorist about normativity, he or she must defend the claim that

> There are no normative propositions about belief

And that is clearly incompatible with accepting propositions like

> John's belief that J.K. Rowling is poor is incorrect

if in these "correct" is understood as a normative predicate.[8] What holds for belief also holds for assertion. Assertions are correct when they are true, but to say that an assertion is correct is to say more than saying that it is true. Truth, as Dummett (1959) reminded us a long time ago, is not merely a property of thoughts or propositions. It is also the point of assertion, as winning is the point of playing a certain game. In this sense "it is part of the concept of truth that we aim at making true statements" (*ibid*). And Dummett precisely opposed on that basis the "redundancy conception of truth", which is but a version of the deflationary view (*ibid*. 7).

The argument from error thus fails if it just amounts to denying these normative liaisons of truth and of assertion. "Correct" is a normative predicate, and in so far as truth is the correctness condition of belief and assertion it is at least part of the concept of truth that it has this role. But where does this correctness feature come from? If we consider belief, it seems to come from the fact that beliefs, like a number of other mental attitudes, such as judgments, guesses, conjectures or suppositions,

[7] Wedgwood 2007, p. 158.
[8] Shah 2011, p. 101.

is such that it can miss the mark or fail, precisely when they are false. This is not the case with mental states such as knowing, seeing, hearing, perceiving, noticing or realizing. These states are factive in that they cannot fail to hit the mark. It is tempting to conclude that the normativity of belief and of other doxastic attitudes comes from the fact that one can succeed or fail in hitting the target. From there comes the familiar metaphor that truth is the "aim" of belief, in the sense of its goal. Thus the normativity which attaches to belief is a kind of teleological normativity: belief succeeds to reach its goal when it is true, and fails otherwise[9]. The same holds of assertion, which is an action, for which there is such a thing as carrying it out correctly or incorrectly. And truth carries with it its normative weight through this feature.

A deflationist, however, can resist this argument in the following way. S(h)e can agree that assertion is a kind of achievement, something that one does well or not, in the right or in the wrong way. Thus a speaker who uses an ambiguous expression can fail to carry the message that he intends to convey. Alternatively, he may succeed to convey another message, if he intends to be ironical, and takes his use of an ambiguous expression as a means to that end. So in this sense, his assertion will be correct, since he has conveyed the message that he intended to convey, although it is incorrect because false or ambiguous. So there are two distinct senses of "correctness" here. In so far as the propositional content of his assertion is concerned, correctness in the first sense does not reside in any achievement or success of the speaker in performing a certain action. It just consists, when this content is evaluable for truth or falsity, in its being true. In so far as correctness resides in the success with which the speaker has conveyed her message, it is independent of its correctness in the first sense. So we can distinguish, with Judith Jarvis Thomson (2008, p. 109) two kinds of normativity. *Internal normativity*, on the one hand, pertains to the performing of a certain kind of activity, or the exercise of a certain capacity of skill, in reaching a certain result. Internal normativity need not be intentional in the sense that the agent intends consciously to reach a goal—it may be the result of a habit—but it is a certain kind of achievement. In this sense actions, such as tying one's shoe laces are correct or not. *External normativity*, on the other hand, is the fact that the agent's goal is reached or not, independently of the performance of the agent or of the working of a given function. For assertion and for belief, external normativity or correctness just consists in the truth of their content. Typically this applies to doxastic and cognitive attitudes, which have an intentional content, which is semantically evaluable, but not to actions, although there is such a thing as the external correctness of, say, tying one's shoe laces: it is the state of affairs that one's shoe laces are tied.

Although it is clear in the case of assertion, the distinction between internal and external normativity is not so clear in the case of belief. For beliefs are not things that we do, or performances or achievements. There may be such a thing as guessing well, conjecturing well, judging well, but it is odd to say that there is such a thing

[9] This view has been defended at one stage by David Velleman 2000. It is explicitly defended by Sosa 2011.

as believing *well* in the internal sense. What one means then has to do not with the content of the belief but with the *believing* (Thomson 2008, p. 110). The internal connection has to be spelled out in another way. But whatever this internal correctness can be in the case of belief, the distinction between internal and external correctness can be exploited by the deflationist.

For the latter can argue that whenever we talk about truth as the correctness condition or standard of belief and of assertion, we only refer to *external* correctness, and not to internal correctness. But the former, unlike the latter is not normative. It is merely a descriptive condition. Thomson (2008, p. 83) reminds us that correctness of an item is relative to a kind: a correctness fixing kind K is what fixes the standards that a K has to meet in order to be correct *qua* K. For example the kind *map of England* fixes the standards that have to be met for something to be a correct map of England, and the kind *spelling of chiaroscuro* sets the standard that a spelling of that word has to meet in order to be correct qua spelling of that word. The features which make a map a *correct* map of England are purely descriptive: these are properties of scale, similarity and representation. Similarly the features which make correct a spelling of a word are purely orthographic or phonetic. The correctness fixing kind for belief is truth. But the feature which makes a standard normative is not itself normative. When we say that a belief is correct in the external sense, we just say that its propositional content is true. Thus the initial deflationist claim seems vindicated.

The extent to which this deflationist rejoinder works will depend on how much one can consider internal correctness as separate from external correctness and upon whether it is right to say that external correctness is a merely descriptive condition. I shall come back to this in Sect. 3 below. But in so far as he accepts that external correctness is a genuine feature of belief and of assertion, the deflationist cannot simply stick to the error argument.

11.3 Wright's Inflationary Argument

The most well-known argument against deflationism is Wright's (1992) and concerns assertion. It is a descendent of Dummett's argument. It purports to show that deflationism is inconsistent with the idea that truth is but a device of endorsement of a proposition. The point of the concept of truth, for the deflationist, is to endorse assertions. We can put this, as before, in terms of the notion of correctness. A proposition is correct if it can be asserted, hence if it is warrantedly assertible. So the only "norm", if one wants to express oneself in this way, which the deflationist can grant is that of warranted assertibility. Truth, however, registers a different norm than warranted assertibility. In brief form, Wright's argument is the following. Warrant and truth are intimately related in our assertoric practice: whenever I believe I am warranted in asserting some proposition, I also believe that it is true, and whenever I believe that some proposition is true, I also believe that I have warrant for it. So truth and warrant coincide in positive normative force. Now according to deflationism, the schema (E) has an equivalent for negation

(Neg) It is true that *not p* iff it is not true that *p*

But the corresponding instance of (Neg) is wrong if we substitute in it "warranted" for "true". For any proposition which is for us neither warranted nor unwarranted (such as, say, *that the Loch Ness monster does not exist*) the conditional

It is warrantedly assertible that *not p* if it is not warrantedly assertible that *p*

does not hold. So truth is a norm of correctness *distinct* from warranted assertibility. They diverge in extension (Wright 1992, p. 18, 2001, p. 756).

Truth is normative, in the sense that it is "a property the possession or lack of which determines which assertions are acceptable and which are not" (Wright 2001, p. 775). Truth is the correctness condition for assertions, and, more fundamentally, for beliefs. It is, moreover, a substantive and essential property of these. If one were to describe the assertoric practices of a population without mentioning that truth is what these assertoric practices are *for*, and that it is what makes them *correct*, one would have failed to describe these practices. We could not *explain* these practices if we took truth and warranted assertibility to coincide in extension. So we need to "inflate" the concept or the property of truth rather than to deflate it.

One way to formulate Wright's point would be to say that the deflationist's dissolution of the normative character truth does not capture the distinction between a *subjective* or *prima facie* reason to assert or to believe something.

a. Believing that *p* is true is a *prima facie* reason for asserting that *p*

and an *objective* reason

b. The fact that *p* is true provides a good reason to assert it

Wright's inflationary argument involves this distinction between merely having *prima facie* reason to assert (believe) that *p* and having a *warranted reason* to assert (believe that *p)*. But it also involves the distinction between *having a warrant* and *being true*. The difference is clearly brought out by Huw Price's (1998, 2011) distinction between three norms of assertibility:

1. *Subjective*: it is *prima facie* correct to assert that *p* if one believes that *p*
   or: one is incorrect to assert that p if one does not believe that *p*
2. *Objective*: *p* is objectively assertible if S's belief that *p* is justified
   or: one is incorrect to assert that *p* if though one believes that *p* one does not have adequate grounds for believing that *p*
3. *Hyper-objective*: if *p* is true one should assert that *p*
   or: one is incorrect to assert that *p* if, in fact it is not the case that *p*.

To spell out these differences, Price imagines a tribe, the MOA (Merely Opinionated Asserters), who criticize assertions for flouting the principles of subjective assertibility and objective assertibility but not for flouting that of hyper-objective assertibility. These speakers "express their beliefs—i.e., the kind of behavioural dispositions which we would characterize as beliefs—by means of a speech act that we might call *merely opinionated assertion*" They criticize one another for making insincere or inadequately justified assertions, but not for asserting what's false. We can

also imagine these speakers being fully competent in applying the deflationist truth concept. They fully understand the deflationist truth concept, but not the concept of truth. Thus, the former can't be the same as the latter.

The MOA's concept of truth is limited to the deflationist one and to the warranted assertibility one. But they become extinct because they lack the capacity to express genuine disagreements. They can only express faultless disagreements. They would be relativists of sorts[10].

To arguments such as Wright's and Price's, deflationists can answer by biting the bullet, and by accepting that there is a norm of truth, or a normative import of truth, hence following the second strategy mentioned above—the one I have called concessive—which accepts the idea that there is a normative dimension in belief and in assertion, and thus grants that in so far as belief and assertion have correctness conditions, this dimension belongs to our concept of truth as well. But the deflationists deny that this normativity is an *essential* feature of truth. So the deflationist (Horwich 1990, 2001; Dodd 1999) can perfectly accept the idea that truth carries with it a normative load and he can say that his account does take care of the norms

1. One ought to assert (believe) only what is true

   (if one takes normativity to be of a deontic kind), or

2. Truth is what it is (good) valuable to assert (believe)

He denies, however, that this comes down to more than

3. One ought to believe that *p* iff *p* (NT)

Where there is no mention of truth, and from which in turn one can derive a (potentially infinite) disjunction of sentences of the form

4. One ought to assert that snow is white only if snow is white; one ought to assert that grass is green only if grass is green, etc.

Or again, if we want to cash out normativity in terms of what it is good or valuable to believe or to assert

5. It is good valuable to assert that snow is white only if snow is white; one should assert that grass is green only if grass is green, etc.

Indeed (i) or (ii) allow us to generalise over the conjunctions of claims (iv) and (v), but that does not mean that there is a general norm of truth for assertion independently of subject matter and of the particular assertions listed, and truth is neither mentioned in (iii) nor in conjunctions like (iii) and (iv)[11]. In other words there is no normative substantive property which would underlie the uses of "true". So the deflationist can claim that Wright's or Price's arguments do not show that truth is normative in any

[10] In a sense not far from the one which is advocated by Kölbel 2002 and other contemporary versions of relativism about truth (MacFarlane 2005).

[11] Horwich 1990, Dodd 1999, p. 297,see also Blackburn 2013.

robust sense. The essentials of truth still remain that it is a purely logical device for quoting propositions and generalising over them. The deflationist can perfectly accept infinitely many inferences of the following form:

1. < p > is true iff p,
2. One is incorrect to assert that p if one does not have adequate grounds for believing that p.
3. One is incorrect to assert that p if one does not have adequate grounds for believing that < p > is true.[12]

Moreover, the deflationist claims that one can explain the value which it is attached to truth in instrumentalist terms, on the basis of the familiar idea that true beliefs conjoined with desires lead to actions: for any action A resulting in reaching a goal G, there are beliefs of the form "If I do A I shall get G", which are true. Hence the truth of these beliefs is explained in instrumentalist terms. If one objects that truth is not simply instrumentally valuable but that it is also intrinsically valuable, the deflationist can also grant this and claim that the intrinsic value of truth does not amount to more than an infinite list of the form (iv) or (v) (Horwich 1999, pp. 256–258, 2006). As Blackburn says, summarizing this line:

> 'You must take care that what you say is true' is a schema for collecting individual pieces of advice: 'you must take care that if you say that aardvarks amble, then it is true that aardvarks amble'. Again, the truth predicate can be knocked out of these individual statements with no change, for the same norm or aim is put by saying 'you must take care that if you say that aardvarks amble, then aardvarks amble', and again the generalization or schema of normative advice (or obligation or aspiration) introduces nothing more. (Blackburn 2013, p. 264)

But this answer is certainly unsatisfactory. First, because, as Wright points out, one cannot accept that what it is to satisfy the norm of truth for belief or assertion amounts to nothing more than knowing that series of conjunctions like (iv) and (v) unless one *already* understands the difference between the proposition that snow is white or that grass is green and the proposition that those propositions are warranted (2001, p. 757). In other words the normative character of assertion or belief expressed by conjunctions like (iv) and (v) does not capture the generality of the norm or value of truth[13]. Moreover it seems to lead us to a form of particularism about norms or values according to which there are as many cognitive norms as there are particular true sentences that we could assert or beliefs that we could entertain about particular subject matters. On the deflationist reading of the norms of truth for assertion and belief, we can only acknowledge the existence of particular norms or values attached to each sentence or belief, not the existence of a general norm such as (NT). The deflationist agrees that we ought (or that it is good) to believe that snow is white when snow is white, and that we ought (or it is good) to believe that grass is green when grass is green, and agrees that these judgments are normative, but he denies that there

[12] Mc Grath 2003.
[13] See also Engel 2008 for a similar criticism.

is a general property of being governed by the general norm that one ought (or it is good) to believe what is true, except as a mere generalization on a variety of such particular normative statements (Lynch 2009, pp. 512–513). But it is implausible, in particular if one adopts the deontic version of the norm of truth for belief, that one ought to believe what is true. We do not learn norms by being presented with a variety of instances of statements of the form "One ought to φ", but by grasping that there is a common property behind these instances. Cain cannot find relief from his accusing conscience in saying to himself "I ought not to have killed Abel, but this deontic principle would be distinct if I had killed Jonah, if I had killed Elijah, etc". One is forced to accept that the norm is not only a mere generalization over particular cases, but also that it explains our uses of the truth predicate, hence that it is a substantive norm.

One other way of seeing this substantive character is to compare the deflationist position with relativism about norms. Deflationism has often been said to have relativist implications. For if all there is in common about such statements as "< Penguins waddle > is true if and only if penguins waddle, or "< aardvarks amble > is true if and only if aardvarks amble", etc. is the existence of the truth-predicate which functions as a device of quotation and of generalization, then there is nothing in common between these statements. They have as many subject matters as there are topics about which they are asserted, and as many reasons to assert them as there are particular subject matters. As Heal (1987/1988, p. 107) remarks, there is no general property which can account for the success and—hence correctness—of an assertion since this success depends in each case upon the subject matter. My assertion that aardvarks amble has different correctness conditions from those of your assertion that penguins waddle. This thought also invites a form of pluralism about truth, which has been associated to Wright's work. But I shall leave aside here the territory where deflationism meets, at least in part, varieties of pluralism[14], to consider only one of its relativist implications. Suppose that we grant, as the concessive deflationist does, that truth is a norm or standard of correctness for belief and assertion. Why should we accept that there is only one such norm? After all there are plenty of reasons to assert various statements, some cognitive, some practical, and others of a different nature. As Stuart Cohen says discussing contextualist views of knowledge:

> Contexts shift depending on variations in speaker's interests, intentions and depending on various standards that sentences used to attribute knowledge encode. *No such standard is simply correct or simply incorrect. . . And there is no context independent standard*. (Cohen 2000, p. 97)

A number of writers, like Kölbel (2002) and MacFarlane (2005), have argued, and elaborated in considerable detail the idea that truth is, a least in a number of cases, relative to contexts of assessment or of evaluation. The classical question is, of course, whether this relativism is coherent. Can the relativist accept the very idea that truth is the norm (or aim) of assertion if there are as many norms as there are

[14] See Wright 1992, Lynch 2008, and for an examination of how pluralism can deal (and in my view fails to deal) with the norm of truth, see Engel 2013.

contexts of evaluation? The relativist has to say that even if truth were *an* internal normative aim of assertion, it is certainly not the only such aim: it is also part of the practice of assertion that we strive to say what is relevant to the conversation at hand, and to say things that are appropriately justified (or on some accounts, known). For a context of assessment A and a context of assessment, B, the relativist cannot validate both

P is *A*-true↔ *P*.
P is *B*-true↔ *P*

He must validate each of these statements only within a perspective. It then seems hard to accept Frege's reasoning above and the claim that truth is the correctness norm of assertion, since it is not clear how speakers coming from different perspectives or contexts of assessment but who would assert respectively P and not-P in their respective context could be described as disagreeing at all, and in this sense we find again the problem of how to conciliate the stance of a "mere opinionated asserter" with the objective norm of truth (Rozenkranz 2010, pp. 232–234). It is not clear that a deflationist would like to pay this price by embracing a form of relativism about norms of assertion

11.4 The Normative Import of Truth

If the deflationist can adopt neither the irrealist conception of the norm of truth nor the concessive view, what are the options left for him? It seems to me that he has at least three other options. But none of these is very attractive.

The first one, in line with the concessive strategy, would be to accept that truth is the external correctness condition for belief. This external condition is, as we saw only a criterion of rightness of belief or of assertion. It tells us that the formal object of belief is truth. But in what sense could we say that it vindicates the thesis that belief is "normative"? A norm, to be a norm, has to be able to explain the practice that it governs and to guide us in our actions or in our thoughts. But the norm of truth as criterion of rightness has no "normative force" (Railton 2003): it just tells us what a belief is. So we seem to be brought back to the view expressed by Dretske in the quote given above: the external condition of correctness amount to nothing else than a definition of belief. And from that we get no normative force or power. Indeed, as I suggested above, this conclusion, if it were the only one to which we could be led, would not displease the deflationist, since his point has always been that the normative import of truth is but a very shallow one. But it would fail to meet the point which the critics of deflationism have emphasized: that the norm of truth, in order to be a norm, has to make a difference. What is it for a norm to make a difference? I would suggest that it involves giving an account of the external norm or correctness condition *and* of the internal norm. Remember that the internal norm is meant to tell us how the external norm can regulate and govern the actions of agents or their thoughts. The external norm—or correctness condition of a guess is

the same as the external norm for belief, but the internal norms are not the same, since guessing is a distinct kind of state (or perhaps activity) than belief (which is not an activity). In other words the conditions of success of guessings are distinct from those of believings, even though the external conditions are the same (in both cases one is supposed to reach true beliefs and true guesses). In the case of the norm of truth for belief it would have to explain in virtue of what believers can be motivated by the external norm. It must give us, or so it seems, prescriptions about what we ought to believe. Now a number of critics have argued that it is an impossible task, since the norm of truth leads to impossible predictions. (NT) for instance seems to entail that one has to believe any truth whatsoever, a consequence which is equally absurd for assertions (Bykvist and Hattiangadi 2007; Glüer and Wikforss 2009).

I can only sketch my own conception of the norm of truth here, concentrating on belief only[15]. It seems clear that we cannot account for the normative force of the norm of truth without investigating its relationship with the norm of evidence (NE). The primary reason why we believe anything is that we have evidence for our beliefs. There must be some conditions such that certain amount of evidence is for us a reason, not only in the normative sense of "reason" (good grounds) but also in the motivating sense. Truth here seems to play no causal or psychological role. But at this point the external condition of correctness must enter the picture. A belief which is based on sufficient evidence can manifest the believer's aptitude at the internal correctness of the norm, but it might also be false, hence not satisfy external correctness. So the external norm is the ultimate arbiter, although the internal norm—here the norm of evidence (NE) is the one which motivates our believing. Critics of (NT) conclude that it is inert or useless. But it is not. I propose to construe it not as a prescription or a directive which believers have to obey, but as an *ideal* condition which a believer has to meet in order to conform to the norm. But the ideal has to be in place whatever the epistemic condition of the believer can be and whatever his psychological condition. It is just silent upon these. In Boghossian's terms, which I largely share," The truth is what one ought to believe, whether or not one ought to go about it, and whether or not you know whether you have attained it" (Boghossian 2003, pp. 38–39). This ideal norm is objective, in the sense that it holds independently of believers. It is constitutive of belief, and of the concept of belief. The evidential norm, in contrast, is subjective, and does not hold independently of the epistemic state of the believers. Much more indeed needs to be said to explain how the objective and the subjective norms are related, and in what sense the norm can be said to be essential and constitutive of the state of belief. But it is clear that the nature of this norm makes it substantive, in a sense which the deflationist cannot accept.[16]

[15] I have developped it elsewhere in Engel 2007, 2013 and 2013b.

[16] The deflationist could here argue, once again, that if the norm of truth is supposed to *define* belief, it is not a norm in any interesting sense. He can also argue that it is no constitutive of the concept of belief either. I shall not here deal with these objections.

The second option which the deflationist can take is to stick to a deflationist account of the truth-predicate—hence to reject the idea that he could be associated to an external norm of assertion or to belief—while giving a separate account of the norms of assertion. This is the line taken by Brandom (1994), who gives a purely deflationist account of the truth predicate (actually a version of the "prosentential conception of truth", according to which it is not a predicate but a component of prosentences such as: "*p; if that is true, then q"*), but gives an account of assertions in terms of commitments and norms which are in the last instance social. For instance if I assert that *p* I commit myself to *p*, and am liable to criticism and censures if *not p* turns out to be the case, and I commit myself to give reasons for *p*, and hold myself responsible for giving these reasons. Brandom's program consists in "starting with an antecedent notion of assertional significance and then moving via that principle to an understanding of what is involved in talk of truth" (Brandom 1994, p. 232). This is a version of the view that truth is but a device of assertion, although the notion of assertion is itself characterized in terms of "deontic statuses" and social norms. Given that Brandom's account is clearly normative with respect to assertion—but not to truth—does it count as an answer to the kind of objections voiced by Wright or Price? Brandom's proposal seems to be caught up in a dilemma. On the one hand, if he aims at understanding "true" in terms of the rules of assertion or of our "taking-true", then truth becomes a normative notion and one cannot just get rid of it. But then the deflationist thesis according to which it does not amounts to nothing more than (E) cannot be sustained, and Brandom's view is not any more a deflationist one. On the other hand, if the accounts meant to keep separate the deflationist analysis of the truth predicate and the social analysis of norms of assertion, then we no longer get an explanation of the former in terms of the latter (Bar-On and Simmons 2007).

The third option open to a deflationist could be to accept the normative liaisons of truth hence that norms such as (NT) are genuine properties associated to truth, but adopt an expressivist account it. Expressivists about truth take our use of "true" not as descriptive, but as expressive of our mental states, and registering our approbation. There are various versions of this view, from the thesis, sometimes attributed to P.F. Strawson, that truth plays a mere performative role to the idea, voiced by Rorty (1986), that truth is but a "compliment" that we pay to our assertions. On a more sophisticated version truth expressivism one must analyse our truth ascriptions as expressions of our acceptance of the norms associated to truth, namely those of correctness of assertions and belief. The view as a counterpart for epistemic norms (Field 2009). Expressivism about truth, far from playing down the normative features of truth, put them forward. "True" becomes a normative notion through and through, expressing our *valuing* the statements that we assert, and describing nothing. "True" means indeed, correct, and is associated to the norm of correctness, but this norm is understood in the expressive way, just as the *oughts* for belief which figure in (NT) and (NE) are to be understood in this expressive way too. I shall not here discuss the various objections to his view, and shall raise only one of these (Lynch 2009, p. 515). On the expressivist view, the claim (NT) that one ought to believe only what is true cannot mean that one ought to have beliefs which have the property of being truth. It must mean that one ought to have the belief towards which one adopts a certain

stance or attitude. But what makes this attitude the right one when snow is white and the wrong one when snow is black? It seems that we go back to the difficulty already encountered with the mere opinionated assertors, of failing to get a norm which could be objective.

11.5 Conclusion

I have reviewed here some of the main lines of argument that a deflationary conception of truth can give in order to account for the normative features of truth. All illustrate a dilemma: either deflationism simply eliminates the normative features, in which case he is unable to account for the correctness of assertions and beliefs, or he grants the *prima facie* normative character of these features, but is unable to account for the objective and substantive character of the norms associated to truth. The deflationist, however, is right on this point: the norm of truth for belief is not a property of truth itself, and it is neither part of the essence of truth nor analytically contained in its concept. If the norm of truth is essential or constitutive, it is of belief and of assertion, not of truth. So truth can remain as norm-free as is it is on any correspondentist theory. In contrast, truth cannot remain as simple and austere as the deflationist aspires it to be.[17]

References

Bar-On, D., & Simmons, K. (2007). The use of force against deflationism: Assertion and truth. In D. Greimann & G. Siegwart (Eds.), *Truth and speech act: Studies in philosophy of language*. London: Routledge.

Blackburn, S. (2013). Deflationism, pluralism, expressivism, pragmatism. In Pedersen N. J. L. & Wright C. D (Eds.), *Truth pluralism* (pp. 263–277). Oxford: Oxford University Press.

Boghossian, P. (2003). The normativity of content. *Philosophical Issues*, *13*(1), 31–45.

Brandom, R. (1994). *Making it explicit*. Harvard: Harvard University Press.

Bykvist, K., & Hattiangadi, A. (2007). Does thought implies ought? *Analysis, 67,* 277–285.

Cohen, S. (2000). Contextualism and skepticism. *Philosophical Studies, 10,* 90–107.

David, M. (1994). *Correspondence and disquotation*. Oxford: Oxford University Press.

Dodd, J. (1999). There is no norm of truth, a minimalist reply to wright. *Analysis, 59,* 291–299.

Dretske, F. (2001). *"Norms, history, and the mental". in perception, knowledge and belief.* Cambridge: Cambridge University Press (2000).

Dummett, M. (1959). Truth. *Proceedings of the Aristotelian Society, 59,* 141–162. (repr in *Truth and Other Enigmas,* London: Duckworth 1978).

---

[17] Related, although distinct, versions of this article have been read in various versions in Amsterdam in March 2011at the conference "Truth be told" organized by Dora Achouriotti and Peter van Ormondt, and then at the Paris conference organized by Henri Galinon and Dennis Bonnay in June 2011. I thank them all for their invitations and hospitality, and my commentator at Amsterdam Timothy Chan for his excellent remarks. I thank José Martinez and Henri Galinon for their patience and indulgence, and anonymous referees for their comments.

Engel, P. (2002). *Truth*. Bucks: Acumen.
Engel, P. (2005). Truth and the aim of belief. In D. Gillies (Ed.), *Models in science* (pp. 77–97). London: King's College Publications.
Engel, P. (2007). Belief and normativity. *Disputatio, 23,* 153–177.
Engel, P. (2008). Pragmatic encroachments and epistemic value. In A. Haddock, A. Millar, & D. Prichard (Eds.), *Epistemic value* (pp. 183–203). Oxford: Oxford University Press.
Engel, P. (2013). Alethic functionalism and the norm of belief. In Pedersen N. J. L. & Wright C. D (Eds.), *Truth pluralism* (pp. 69–86). Oxford: Oxford University Press.
Engel, P. (2013a). In defense of normativism. In T. Chan (Ed), *The aim of belief*. Oxford: Oxford University Press.
Engel, P. (2013b). Doxastic correctness. *Proceedings of the Aristotelian Society*, *87*(1), 199–216.
Field, H. (2001). *Truth and the absence of fact*. Oxford: Oxford University Press.
Field, H. (2009). Epistemology without metaphysics. *Philosophical Studies, 143*(2), 249–290.
Frege, G. (1897). Logik. In G. Gabriel (Ed.), *Schriften zur Logik und Sprachphilosophie* (pp. 35–73). Hamburg: Meiner Verlag (1990).
Glüer, K., & Wikforss, A. (2009). Against content normativity. *Mind,* 118(469), 31–68.
Heal, J. (1987/1988). The disinterested search for truth. *Procedings of the Aristotelian Society*, 97–108.
Hill, C. (2002). *Thought and World*. Cambridge: Cambridge University Press.
Horwich, P. (1990). *Truth*. Oxford: Oxford University Press (second ed. 1998).
Horwich, P. (1999). The minimalist conception of truth. In S. Blackburn & K. Simmons (Eds), *Truth*. Oxford: Oxford University Press.
Horwich, P. (2001). Norms of truth and meaning. In R. Shanz (Ed), *What is truth?* (pp. 133–145). Berlin: De Gruyter.
Horwich, P. (2006). The value of truth. *Nous,* 40(2)*,* 347–360.
Kenny, A. (1963). Action, emotion and will. Oxford: Oxford University Press.
Kölbel, M. (2002). *Truth without objectivity*. London: Routledge.
Künne, W. (2003). *Conceptions of truth.* Oxford: Oxford University Press.
Lynch, M. (2009). *Truth as many and one*. Oxford: Oxford University Press.
MacFarlane, J. (2005). Making sense of relative truth. *Proceedings of the Aristotelian Society, 105,* 321–339c.
Mackie, J. L. (1977). *Ethics, inventing right and wrong.* London: Penguin Books.
Mc Grath, M. (2003). Deflationism and the normativity of truth. *Philosophical Studies, 112*(1), 47–67.
O'Leary-Hawthorne, J. & Oppy, J. (1997). Minimalism and truth. *Noûs*, *31*(2), 170–196.
Price, H. (1998). Three norms of assertibility. *Philosophical Perspectives, 12,* 41–54 (in Tomberlin, J., ed.).
Price, H. (2011). *Naturalism without mirrors*. Oxford: Oxford University Press.
Railton, P. (2003). *Facts, values and norms*. Cambridge: Cambridge University Press.
Rorty, R. (1986). Pragmatism, Davidson and truth. In Le Pore (Ed.), *Davidson, truth and interpretation* (pp. 333–355). Oxford: Blackwell.
Rozenkranz, S. (2010). Frege, relativism and faultless disagreement. In M. Garcia- Carpintero & M. Kolbel (Eds.), *Relative truth* (pp. 225–238). Oxford : Oxford University Press.
Shah, N. (2011). Can reasons for belief be debunked. In A. Reisner & A. Steglich-Petersen (Eds.), *Reasons for belief* (pp. 94–107). Cambridge: Cambridge University Press.
Sosa, E. (2011). *Knowing full well*. Princeton: Princeton University Press.
Stoljar, D. (2010). "The deflationist theory of Truth", *Stanford encyclopedia of philosophy*.
Thomson, J. J. (2008). *Normativity*. La Salle; Open court.
Velleman, D. (2000). On the aim of belief. *The possibility of practical reason*. Oxford: Oxford University Press.
Wedgwood, R. (2007). *The nature of normativity*. Oxford: Oxford University Press.
Wright, C. (1992). *Truth and objectivity*. Cambridge: Harvard University Press.
Wright, C. (2001). Minimalism, deflationism, pragmatism, pluralism. In M. Lynch (Ed.), *The nature of truth* (pp. 751–789). Cambridge: MIT.

Part IV
Deflationism and Conservativity

Chapter 12
Norms for Theories of Reflexive Truth

Volker Halbach and Leon Horsten

Abstract In the past two decades we have witnessed a shift to axiomatic theories of truth. But in this tradition there has been a proliferation of truth theories. In this article we carry out a meta-theoretical reflection on the conditions that we should want axiomatic truth theories to satisfy.

12.1 Introduction

In the past two decades we have witnessed a shift from semantic—such as Kripke's or the revision theory—to axiomatic theories of truth. But in this line of research there has been a proliferation of truth theories. How should we adjudicate between them? [1]

We list some desiderata or norms for axiomatic theories of truth and explain how they can be used to discriminate between theories. To some extent these norms will also be useful for explaining what is driving the work on axiomatic theories of truth, as in many cases authors—including the authors of this article—have not been very explicit about their motivations, but rather concentrated on analysing the formal properties of the theories. So to a limited extent the norms we are going to list should be understood to be not only normative but also descriptive in the sense that they are intended to make explicit the norms that have be applied by various authors in the field.

Volker Halbach's work is part of the research project AH/H039791/1 that is supported by the Arts & Humanities Research Council.

[1] The title and some of the content of this paper is inspired by Giulia Terzian's (2012) PhD work on norms for theories of truth.

V. Halbach
New College, University of Oxford, Oxford, UK
e-mail: volker.halbach@philosophy.ox.ac.uk

L. Horsten
University of Bristol, Bristol, UK
e-mail: leon.horsten@bristol.ac.uk

T. Achourioti et al. (Eds.), *Unifying the Philosophy of Truth,* Logic, Epistemology, and the Unity of Science 36, DOI 10.1007/978-94-017-9673-6_12

Of course the formal analysis of a theory will be helpful in assessing whether a theory satisfies the desiderata and the formal analysis cannot be separated from a philosophical judgement. We don't see the reflections on the norms as a stage that should precede formal work on the systems. Formal results have helped logicians to choose and formulate the norms, and the norms have motivated the formal work on theories. So we do not think of the norms as a *prima philosophia* that is conceptually prior to the logical analysis.

Some efforts have been made to carry out exercises similar to that of the present article: these include Sheard 1994 and Leitgeb 2007. But, for reasons that we hope to make clear, these efforts have remained less than fully satisfactory.

Let us start by listing some aspects of axiomatic truth theories that we require all axiomatic truth theories to accept wholesale. They will be treated as background assumptions in the sequel. This is done mainly to focus the discussion. We do not claim that they are unproblematic and cannot be challenged.

The aim is to explicate the meaning of the truth predicate without presupposing the distinction between object- and metalanguage. We will investigate systems in which the truth predicate is contained in the object language. So we mainly aim to unfold the notion of type-free or *reflexive* truth. The distinction between typed and type-free notions of truth is not unproblematic.[2] But all the theories that we are considering here prove the truth of sentences containing the truth predicate and are thus type-free in this sense.

We will treat the notion of truth as a primitive predicate T and reflect on how truth can and should be axiomatised. The treatment of truth as a primitive predicate in itself does not rule out the possibility that truth is a definable concept. Whether truth is definable or reducible by other means depends on the chosen axioms. However, only very weak theories will escape Tarski's theorem on the undefinability of truth. Hence the undefinability of truth in the case of all interesting truth theories will be a result that is arrived at, not a presupposition. So in contrast to semantic approaches and "substantial" theories, where the definability of truth has to be assumed from the outset, the axiomatic approach is neutral with respect to the question whether truth is definable or not.

It is commonly acknowledged that the axioms of truth should be studied in conjunction with axioms for the objects to which truth is ascribed. These may be sentence types or tokens, propositions, or still other objects. Many authors working on the axiomatic approach to truth think of truth as a predicate applying to (codes of) sentence types and we will follow them here without justifying this assumption.

The languages and theories we are interested in are formulated in first-order predicate logic. We hope that many of the results shed light on the use of the truth predicate in philosophical discussions, but we do not claim that our setting is the best starting point for an analysis of truth in natural language.

[2] For more on the distinction between typed and type-free theories of truth, (see Halbach 2011, Sect. 10).

The language in which the axiomatic theories are formulated will be taken to be $\mathcal{L}_T$, which is the language $\mathcal{L}_{PA}$ of first-order arithmetic expanded with the primitive truth predicate T. For again familiar reasons, we need in our axiomatic systems to be able to reason about finite sequences of symbols, or, equivalently, about finite sequences of numbers (conceived of as codes of symbols). We insist that all the axiomatic truth theories that we consider can prove all theorems of PA restricted to $\mathcal{L}_{PA}$. So we shall call PA (restricted to $\mathcal{L}_{PA}$) the *base theory*, and we shall call $\mathcal{L}_{PA}$ the *ground language*.[3]

Many of our comments will apply to many other base theories *mutatis mutandis*. Using theories weaker than PA will complicate the considerations in many cases. For instance, if a theory such as $\mathrm{I}\Sigma_1$ employed, then it is far from being clear how the induction principle should be generalised to the expanded language with the truth predicate. We prefer to steer clear of these delicate issues in the present paper. Generally it is easier to apply our considerations to theories with unrestricted schemata like Zermelo–Fraenkel set theory. But also in this case some delicate issues arise, as is shown by Fujimoto 2012.

12.2 Imprecise and Contestable

We aim to formulate informative criteria that apply directly to axiomatic systems. Nonetheless, all our criteria are imprecise and fuzzy.

Fulfillment of our criteria is not going to be an all-or-nothing affair.[4] The criteria will be such that they can be satisfied to a lesser or to a greater extent. It should not be assumed that the degree of satisfying individual desiderata can be quantitatively measured: perhaps a partial ordering relation for truth theories is the best we can get.

Given that we are plagued by the semantic paradoxes, it will come as no surprise that the desiderata on our list will not be independent: they cannot all jointly be satisfied, as we will see. But given that fulfillment of the criteria will be a matter of degree, we may hope to be able to satisfy all desiderata to a reasonable degree. While the criteria are not independent, we do want a significant amount of independence between them.

Even if meaningful quantitative degrees could be obtained, evaluating the adequacy of an axiomatic theory of reflexive truth will not be a simple matter of adding degrees. We should not even assume that there is one "formula" that assigns the correct weight to each of the desiderata that we will propose.

We will not go as far as stating one single norm for truth theories. Some researchers believe that any list of norms for axiomatic truth theories should derive from a single purpose of truth, such as expressing generality, or from a single property such as "transparency". But we cannot see how the desiderata actually guiding the search for

[3] For a discussion of the ground language and the base theory see Halbach 2011 and Horsten 2011.
[4] This is also the case for most of Sheard's criteria, which we will discuss later. See Sheard 2002, p. 173.

attractive theories of truth can be derived from such a single purpose or property.[5] At least some non-triviality condition is driving the quest for theories of truth as well.

In fact we think the usability of truth as a device of generalisation is derived from more basic properties such as semantic ascent and perhaps also from compositionality. We do not deny that truth is a tool serving a specific purpose, but we do not think that whatever fits the purpose should be called truth. Rather we start with a predicate satisfying our or similar desiderata and then observe what purposes it can serve.

12.3 Five Desiderata

We shall now propose and discuss a series of desiderata for axiomatic theories of truth. We list them in no particular order of priority.

When authors appraise and advocate their theories of truth, they usually employ criteria of highly diverse kinds. Presumably the desiderata of the simplest kind are those that require the theory to prove certain theorems. For instance, the theory may be required to prove all T-sentences for all sentences from a certain class or the claim that truth commutes with conjunction.

Another kind of desideratum is formulated in more metatheoretic terms. For instance, the theory may be required to be ω-consistent (see Leitgeb 2007) or "symmetric" concerning its inner and outer logic, where the outer logic of a theory S is the set of S-provable sentences and the inner logic is the set of sentences φ with $S \vdash T\ulcorner\varphi\urcorner$ (see Halbach 1994; Leitgeb 2007).

Some desiderata formulated in metatheoretic terms are problematic for our purposes, because we aim to assess the strengths and weaknesses of a truth theory from the perspective of the base theory of that theory. This is especially important when we use our overall theory—this may be our most general theory containing set theory—, because then no standpoint external to this most comprehensive theory will be available to us.

This problem is not specific to truth theories but to overarching frameworks in general. For instance, postulating that our set theory, when used as a general foundations for mathematics, is consistent, cannot be used as a desideratum for the theories themselves: none of the theories can prove the existence of such a model (unless it is inconsistent). This does not imply that the desideratum cannot be used as a guide for extensions of the theory, but we will never be able to see from the standpoint of our best theory that it has a nice and well behaved model. What we can do in certain cases, is to prove in the base theory that if the base theory is consistent, then extending it with certain truth axioms will still yield a consistent theory.

However, in many cases the metatheoretic desiderata are provable in the axiomatic truth theory itself. For instance, the claim that the inner and outer logic coincide has

[5] In fact, we suspect that it has not been clarified what it means to "express generalisations". We are not really satisfied by Halbach's proposal (1999), for instance. Also transparency in itself does not seem to be the full story.

been advocated as a norm, and in most cases this claim that $T\ulcorner\varphi\urcorner$ is provable if and only if φ is provable is itself provable in the truth theory, or, in fact, already in weak arithmetical theories.

For truth theories even the postulate that there should be well behaved models can be reformulated and internalized to *some* extent in the object theory: We might expect that adding the truth axioms to a weaker base theory yields a theory with well behaved models. For instance, we might aim at a truth theory based on Zermelo–Fraenkel and consider whether adding the truth axioms to small fragments of ZF or Peano arithmetic results is a theory for which nice models can be constructed in ZF.

12.3.1 Coherence

Coherence is a notoriously underdetermined notion. Somehow the axioms and rules of a theory should be in harmony with each other. Coherence should not be understood in the way it is often understood in coherence theories of knowledge as supporting or even implying each other. We expect that the axioms and rules for truth do not clash with the base theory and the other axioms and rules for truth. But more may be said. If say the connective $\wedge$ and its interaction with the truth predicate is axiomatised in a certain way, then this would "cohere" with an analogous treatment of a connective like $\vee$.

If the truth theory contradicts the base theory, then the truth theory completely fails on the coherence norm. So if the truth theory proves the negation of an arithmetical theorem of Peano arithmetic, the theory does not cohere with the base theory. We adopt this requirement even for theories of truth in nonclassical logics, such as paraconsistent logics. Contradictions may not be lethal if the contradictions arise from the liar paradox and the theory proves both the liar sentence and its contradiction, but even the paraconsistent logicians will admit that a theory contradicting its own base theory is hopeless, because then the contradiction will be located in the non-semantic part of the theory that ought to be unproblematic. Perhaps exceptions may be made for extremely rich base theories that contain already problematic notions, but even then the addition of a truth predicate should not create any new contradictions in the ground language.

However, coherence is more than just plain consistency with the base theory. The axioms of a truth theory can also be incoherent in other ways.

Another example of a theory that clashes with the base theory is the Friedman–Sheard theory FS introduced under another name by Friedman and Sheard (1987) and further studied by Halbach (1994). The theory is internally consistent, but it is ω-inconsistent, that is, for some formula $\varphi(x)$ the system FS proves $\varphi(\overline{n})$ for each number n but it also proves $\exists x\, \neg\varphi(x)$. This fact can be proved in FS itself and thus FS is inconsistent with the uniform reflection principle $\forall x\, (\mathrm{Bew}_{FS}(\ulcorner\varphi(\dot{x})\urcorner) \rightarrow \varphi(x))$ for

FS.[6] It is also PA-provably inconsistent with the stronger global reflection principle stating that all closed theorems of FS are true. Hence the theory FS refutes its own soundness. We take this to be a form of incoherence.

Other authors such as Leitgeb (2007) have listed the existence of a standard interpretation as one of their norms for truth theories.

Certainly ω-consistency is an important requirement, but we do not see it as a very fundamental requirement: hardly anyone would have thought of imposing ω-consistency as a requirement before McGee (1985) proved the ω-inconsistency of a vast class of truth theories that looked otherwise very attractive. There are many different ways a theory can be incoherent with its base theory; ω-inconsistency is just one of them and McGee's proof showed that this form of incoherence can easily arise for theories of truth.

There may be other forms of incoherence with the base theory. For instance, if the induction schema cannot be extended to the language with the truth predicate on pain of inconsistency, then the theory in question is incoherent with the base theory. We are not aware of a natural theory of truth that violates this requirement, but if it should arise in the future, then the theory would have to be rejected as incoherent, even though nobody had imposed this requirement (except for us). The rejection of the theory would be justified, because of the incoherence with the base theory.

So far we have focused on the coherence of the axioms and rules for truth with the base theory. But a theory of truth should also be coherent in its truth-theoretic part. For instance, the truth theory may be internally inconsistent in the sense that the theory proves $T\ulcorner\varphi\urcorner$ for all sentences φ. Such a theory is not coherent and almost as bad as an inconsistent theory in classical logic.

We do not go into the discussion of the coherence of nonclassical systems. Paraconsistent logicians may want to say that their theories are coherent in some way. At least these theories can be non-trivial, but we are not quite sure whether this should be counted as evidence for coherence. The formulation of criteria for coherence may depend on the particular logic that is chosen.

12.3.2 Disquotation and Ascent

Deflationists do not tire of reminding us that truth is a disquotational device and a device for performing semantic ascent. Hence a sentence φ and the claim $T(\ulcorner\varphi\urcorner)$ should be equivalent in *some* sense. Some disquotation principles will also apply to formulae φ with free variables, but we will not consider them here.

A strong way of giving this slogan precise content is to desire that for a sentence φ of $\mathcal{L}_T$, φ and $T(\ulcorner\varphi\urcorner)$ can be substituted *salva demonstrabilitate* in any formula of $\mathcal{L}_T$. Field 2008 advocates this requirement under the label *transparency*. Others

[6] See Halbach 2011 for details.

focus on the Tarski-biconditionals, that is, equivalences of the form $T(\ulcorner\varphi\urcorner) \leftrightarrow \varphi$. Depending on the chosen logic, both requirements need not coincide.

The Tarski-biconditionals figure prominently in the theory of the paradoxes. Under fairly weak assumptions both the transparency principle and the full Tarski-biconditionals lead to inconsistency. Some philosophers, among them Tarski and more recently Burgess and Burgess 2011, do not flinch and accept the inconsistency view of truth. We have some sympathies with this view, but we do not pursue the project of giving an analysis of the truth predicate as found in everyday English. We see the quest for a good theory of truth as a revisionary enterprise, which involves the rejection of certain features of the truth predicate that may form part of a full analysis of the pretheoretic notion of truth.

There are many different ways to weaken the transparency principle and the Tarski-biconditionals to obtain a nontrivial theory or truth, and there are many different ways to classify these weakenings. We distinguish two possible ways to a non-trivial version of disquotation requirement: Either the class of instances of the T-schema is restricted or the connective $\leftrightarrow$ is replaced with some other (weaker and nonclassical) operator. Both methods can be combined.

On the first account, the set of instantiating sentences is restricted. Often the guiding principle seems to be to retain as many instances of the T-schema as possible. The qualification "as possible" has proved to be less tractable than expected. In particular, the restriction to *consistent* instances does not suffice, as McGee 1992 has shown by using a variant of Curry's paradox. So one might try to confine the set of permissible instances to those that do not prove any new theorems in the ground language. But Cieśliński 2007 showed that this policy does not fare much better than the bolder restriction to maximal consistent sets of instances.

In particular, using an argument reminiscent of Curry's paradox, McGee 1992 showed that in the presence of the diagonal lemma every sentence of $\mathcal{L}_T$ is equivalent to a Tarski-biconditional. Thus any sentence independent from the base theory can be decided using a consistent instance of the T-schema. The reasoning is very simple: Given a sentence φ one obtains a sentence γ_φ by the diagonal lemma:

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

This equivalence logically implies that φ is equivalent to $T\ulcorner\gamma_\varphi\urcorner \leftrightarrow \gamma_\varphi$.

These equivalences are puzzling. If φ is the negation of a theorem of Peano arithmetic or the claim that PA is inconsistent, then the Tarski-biconditional $T\ulcorner\gamma_\varphi\urcorner \leftrightarrow \gamma_\varphi$ is incoherent with the base theory and ruled out by our first norm. In contrast, if φ is a theorem of PA, then $T\ulcorner\gamma_\varphi\urcorner \leftrightarrow \gamma_\varphi$ is provable at any rate without any further truth-theoretic assumptions and thus coherent. If φ is a sentence such as the consistency statement Con_{PA} of PA, the situation is more involved: It will be hard to argue that $T\ulcorner\gamma_{\mathrm{Con}\urcorner_{PA}} \leftrightarrow \gamma_{\mathrm{Con}_{PA}}$ is incoherent; after all it is equivalent to Con_{PA} and that sentence has been argued to be implicit in the acceptance of PA. In contrast, if φ is the consistency statement for a strong theory such as Zermelo–Fraenkel set theory the Tarski-biconditional $T\ulcorner\gamma_\varphi\urcorner \leftrightarrow \gamma_\varphi$ is at least not obviously coherent with any theory of truth with PA as base theory, because the consistency of such as strong theory is,

so to speak, beyond the horizon of an arithmetical theory (assuming that it actually is consistent). So we suspect that there is no simple uniform policy concerning these McGee–equivalences as theorems of a truth theory and they seem to come in many degrees of implausibility, which makes any policy on restricting the disquotation schema tricky. We leave the treatment of these sentences to a future paper where also other restrictions on permissible instances of the T-schema will be discussed.

At any rate, there is no proposal to restrict the admissible instances of the T-schema that is commonly accepted. Even the restriction to T-free instances may be too liberal (Halbach 2006), although it is commonly rejected as too restrictive.

Hence the second option to render the Tarski-biconditionals nontrivial may look more attractive: The material biconditional between $T(\ulcorner\varphi\urcorner)$ and φ in the Tarski-biconditionals is replaced with some other operator or relation. The use of some binary predicate to this end has not become very popular. Expressing the claim that $T(\ulcorner\varphi\urcorner)$ and φ are materially equivalent in a single sentence in a straightforward way by means of a binary predicate expressing material equivalence requires a theory of this binary relation. But material equivalence between sentences φ and ψ is usually analysed as the truth of $\varphi \leftrightarrow \psi$. In general it does not seem attractive to base the theory of truth on a theory of some other relation.

More popular is the substitution of the biconditional by metatheoretic predicates. The transparency principle is an example; but it is too strong if classical logic is to be preserved. However, one can consistently demand that a truth theory S be closed under the inference rules $S \vdash \varphi \Rightarrow S \vdash T(\ulcorner\varphi\urcorner)$ (Necessitation) and $S \vdash T(\ulcorner\varphi\urcorner) \Rightarrow S \vdash \varphi$ (Co-necessitation). This is often paraphrased by saying that the outer logic of S coincides with the inner logic of S.

If classical logic is given up, the possibilities are multiplied. All kinds of so-called biconditionals may be used instead of the classical biconditional. Also metatheoretic principles that collapse into the Tarski-biconditionals in classical logic can be employed, even when the full Tarski-biconditionals are rejected. For instance, one can demand the truth theory S be closed under the inference rules $\varphi \Rightarrow T(\ulcorner\varphi\urcorner)$ and $T(\ulcorner\varphi\urcorner) \Rightarrow \varphi$. Of course in the presence of a deduction theorem, this demand yields the Tarski-biconditionals.

The satisfaction of the disquotation or semantic ascent criterion cannot even be ordered in a linear fashion. However, we do not think that the criterion can be abandoned, not even in favour of "compositional" axioms for truth.

In many cases the Tarski-biconditionals are strengthened by admitting free variables in the instantiating formulae. How this can be done depends on the chosen framework. In a purely arithmetic framework one can formalise the following statement for any formula $\varphi(x)$:

> For any number n: $\varphi(\dot{n})$ is true iff $\varphi(n)$.

The dot above n indicates that the numeral of n is substituted for x in $\varphi(x)$. This can be expressed in a language with the resources of arithmetic as the substitution function is formally expressible in arithmetic.

There are versions with more than one free variable. In a more general setting, when not just numbers are admitted into the ontology, the use of a satisfaction

predicate may be commendable. These versions of the Tarski-biconditionals with free variables are known as *uniform* or *parametrized* Tarski-biconditionals. We think that they flow from the disquotational feature of truth or satisfaction.

12.3.3 Compositionality

Davidson famously defended the view that truth is a compositional notion. In the present context with its very restricted ground language this means that the truth predicate commutes with the logical connectives and quantifiers such as $\wedge$, $\neg$, and $\forall$. This entails that truth of logically complicated sentences φ is determined by the truth value of logically "simpler" components of φ.

The compositional feature of truth does not contain its disquotational feature, at least not in an obvious way. Just demanding that truth commutes with all connectives will not suffice as long as the truth of atomic sentences is not regulated in any way. Usually "compositional" theories also contain the Tarski-biconditionals for T-free atomic sentences (often in a uniform, that is, parametrized version). This is the case for a Davidsonian conception. Once these are added, at least the Tarski-biconditionals for all T-free instances will be derivable under fairly general conditions.

Davidson and most of his followers have applied Tarski's solution to the paradoxes or have not cared much about the paradoxes in general. In particular, they have not said much about the truth of sentences that themselves contain the truth predicate and whether compositional semantics is possible for a language containing its own truth predicate. There are different ways of applying the compositionality requirement to sentences containing the truth predicate.

In the best case, truth should be expected to commute with the connectives and quantifiers independently of whether the sentences contain the truth predicate or not. Truth theories such as the Friedman–Sheard theory FS contain axioms stating this feature (see Halbach 2011).

However, the coherence requirement together with the disquotation desideratum may clash with full compositionality: FS contains all compositional axioms; it thus scores highly on compositionality. Moreover, it contains the rule version of the T-sentences, i.e., necessitation and co-necessitation. This yields an ω-inconsistency.

Therefore some truth theorists explicitly reject the full compositional axioms. In particular, many reject the axiom stating that truth commutes with negation while still retaining other compositional axioms. There are various motives for this restriction of compositionality. According to the view that rejects commutation of truth with negation, truth is fully compositional, but truth is a partial concept and does not apply to all sentences. In particular, the liar sentence may be said to be meaningless or the like and thus neither true nor false. Thus we should not expect the negation of the liar sentence to be true if the liar sentences is not true, because the negation of the liar sentence is as indeterminate as the liar sentence itself.

Thus in theories such as the Kripke–Feferman theory the compositional axioms are weakened to *positive* compositionality capturing the compositionality of a partial

concept. The axioms for truth then no longer describe classical compositionality but rather the compositional principles of some nonclassical logic.

In the Kripke–Feferman theory this is combined with truth iteration axioms that have a strongly compositional flavour. Especially in Burgess' (2009) strengthening of Kripke–Feferman compositionality requirements seem to lead to a theory of *grounded* truth where the truth of each single truth depends on the truth of nonsemantical sentences, that is, sentences without the truth predicate. In this sense the groundedness requirement, which is occasionally seen as a desideratum for a truth theory, can perhaps be seen as a consequence of a strong version of the compositionality desideratum. However, the axiom that Burgess adds to the Kripke–Feferman principles can hardly be seen as expressing a form of compositionality.

Compositionality is not uncontroversial. In particular, supervaluationists argue that for languages containing vague expressions, truth does not distribute over all the connectives. We do not want to take a stance in this discussion. So the most we can claim here is that the compositional axioms should presumably be provable for the classical connectives and quantifiers as far as the vagueness-free fragments of first-order formalisations of vagueness-free fragments of natural language are concerned. Of course, if one goes beyond that, then there are serious worries concerning the possibility of a compositional semantics for natural languages.

12.3.4 Sustaining Ordinary Reasoning

Feferman famously rejected the possibility of withdrawing from full classical logic for $\mathcal{L}_T$ to partial logic as a way of avoiding the liar paradox, on the ground that "nothing like sustained ordinary reasoning" can be carried out in partial logic (Feferman 1984, p. 264). Thus a desideratum for axiomatic truth theories is that they should sustain ordinary reasoning.

Ordinary reasoning should be taken to include schematic mathematical or syntactic reasoning. So, for instance, in mathematics we are used to subjecting every predicate to mathematical induction. This means that axiomatic truth theories where the truth predicate is not allowed in the induction schema, do not receive a maximal score on this desideratum.

It is important to note that this demand extends not just to reasoning concerning sentences of the ground language but to sentences of the entire language $\mathcal{L}_T$. Our ordinary and mathematical reasoning is carried out in classical logic. Reasoning in intuitionistic logic does not come as natural to most of us, but it can be learned without too much difficulty. Reasoning in partial or in paraconsistent logic is a lot less natural still: it is very difficult to learn. Reasoning fluently in even more artificial logic, such as a logic in which certain structural rules are restricted (such as contraction, perhaps), might well be practically impossible.

It should also be mentioned that sustaining ordinary reasoning is not just a matter of the underlying logic and mathematics. If the truth laws themselves form a motley and scattered bunch, then even if the logical system containing it is fully classical,

it will lose points on this desideratum. Also if for instance the inner logic is not classical, points are lost on this criterion. And this again underscores the fact that the criteria proposed here are not fully independent.

12.3.5 A Philosophical Account

Suppose that a theory of truth T is just given as a list of truth axioms in $\mathcal{L}_T$ added to a list of principles and rules of some logic. And suppose furthermore that T scores reasonably well against the norms discussed above: let us assume that T scores better against some of these desiderata than against others. Then some truth theorists would still find T thoroughly unsatisfactory as it stands. These truth theorists want in addition an explanation of why these norms are reasonable desiderata for a truth theory in the first place, and a justification of why it is acceptable that T does not satisfy each of the desiderata to the maximum extent. In sum, they request a philosophical account that justifies the norms, explains them, and ties them together. This request therefore is of a different nature than the other norms: it can be seen as a meta-norm.

The situation may be compared with that in set theory. If the naive theory of comprehension had proved useful (and consistent), then probably it could have passed more easily as a logical or almost logical principle that requires as little a philosophical story for its justification as the rules for the connectives.[7] But because a more sophisticated system such as Zermelo–Fraenkel with fairly complex axioms is required, philosophers felt that some philosophical account—such as Boolos' (1971) story of the cumulative hierarchy—is needed to motivate the axioms of set theory.

In the case of truth, the unrestricted Tarski-biconditionals are inconsistent under fairly general circumstances. The untyped axiomatic theories that have been proposed look far more sophisticated than the unrestricted Tarski-biconditionals. So one may ask whether there is a philosophical account analogous to that told by Boolos about set theory that can be used to motivate or perhaps even justify the truth-theoretic axioms, and that can explain, e.g., why not all the unrestricted Tarski-biconditionals are acceptable.

Many authors believe that a good axiomatic theory of reflexive truth should be embedded in a wider philosophical context and should be underpinned by a winning philosophical account that motivates the choice of the axiom. The philosophical story may comprise an account of how the content of the concept of truth is acquired, how new truths are in ordinary situations established on the basis of truths that have already been acquired (the revision theory of truth can be motivated such a story),

[7] The modality behind this counterfactual should probably an epistemic one. At any rate we do not claim that the inconsistency of comprehension is merely contingent. As with respect to a philosophical story about the rules for logical connectives, one might argue that some philosophical story is needed (that does not apply to 'tonk', for instance). But we would not classify this as a philosophical story that tells us something about the nature of conjunction.

how sentences of $\mathcal{L}_T$ that once were asserted are later withdrawn, how disagreements about propositions from $\mathcal{L}_T$ are resolved, or what purpose truth may serve. Classical "substantial" accounts of truth such as the utility view fall within this scope, but so does the story of the deflationists about truth as a device of generalisation. In this way, the axiomatic truth theory must somehow present a picture. It has to somehow expresses the main tenets of this philosophical account in a succinct and perspicuous way.

According to one such account, truth and falsehood are grounded in non-semantic facts: The truth of sentences ultimately supervenes on the truth of T-free sentences. This thesis is often seen as supported by Kripke's (1975) story about how the concept of truth is learned, and Kripke's minimal fixed point of the Strong Kleene scheme is regarded as one toy model of grounded truth. From iterated semantic ascent and compositionality it follows that many grounded truths of $\mathcal{L}_T$ ought to be included in the extension of the truth predicate. So including many grounded truths can be seen as a derived desideratum. This desideratum can be directly satisfied by a truth theory, by proving positive sentences that contain long iterations of the truth predicate T for instance. But it can also be indirectly met by containing natural interpretations of initial segments of the Tarskian compositional hierarchy, as described in Halbach 1995. Indeed, even from the point of reflexive truth it must be recognisable that the Tarskian hierarchy is fundamentally sound. Of course there is a limit to the length of the initial segment of the Tarski hierarchy that can be recovered by any axiomatic theory. The requirement that only grounded sentences should be classified as true or false cannot be regarded as derived from other desiderata. As mentioned earlier, it is implemented in a theory proposed by Burgess 2009.

Such a concomitant philosophical account or picture should not be confused with a mathematical model or class of mathematical models. In the axiomatic programme, models can at best have a heuristic use. Toy models can help us get a grip on an explanatory account. They can help us explore the structure of a philosophical account, which can then help us to formulate principles of truth. But that is all. In fact, as emphasised earlier, when we formulate our truth axioms for the strongest theories as base theory (set theory being an instance), then we will not even be able to establish the existence of models for the base theory to start from.

Here we do not a stance on the various philosophical accounts. We also do not exclude truth-theoretic pluralism that would admit incompatible truth theories as justified and underpinned by different philosophical accounts. For certain purposes we might be happy with one truth theory that may be better motivated by the actual use of truth in less theoretical contexts while another story is told about the purpose of the truth predicate in the philosophy of mathematics, and still another one about the purpose of the truth predicate in ethics.

The request for a philosophical account in the sense of this section may even be rejected altogether. One of the authors of this paper indeed denies the need for a philosophical account in this deeper sense. In some of the recent investigations into formal theories of truth (such as Friedman and Sheard 1987) this norm plays no role whatsoever. On the other hand, some research in the field has been motivated by desire to build axiomatic theories on the basis of a philosophical account.

12.4 Discussion

We claim that the desiderata that we have listed are more complete than rival lists that have hitherto been proposed. The hope is that all desiderata for theories of type-free truth derive from our list. But it can of course not be excluded that in the future desiderata for self-referential truth theories are discovered that are independent of the list that we propose here.

It is clear from our list of five desiderata that simple adding of marks on each dimension does not give a reliable judgement about the suitability of a theory of truth. For instance, if one is willing to accept a null score on coherence by giving up consistency altogether while classical logic is retained, then one can easily obtain maximal scores on most other dimensions and thus obtain a very high overall score. Yet most researchers would find such truth theories of little value.

Even though an individual researcher might find a very strong form of a desideratum (such as containing the unrestricted Tarski-biconditionals, or being completely classical) simply false, it seems unlikely that researchers who strongly support one of the criteria on the list—and there are many supporters for each of these criteria—are completely mistaken. So it seems not unreasonable to hold that a satisfactory theory of reflexive truth should do at least fairly well on each of the criteria on the list.

12.4.1 Comparison with Sheard

Sheard (2002) contains a list of maxims for truth theories. They are not intended specifically as desiderata for *axiomatic* truth theories. Nonetheless, it is worthwhile to compare them with our list.

Sheard's first maxim says that truth is an objective semantic concept. We agree on this point with Sheard, insofar as we understand it, and think that this point is captured by our background assumptions.

Sheard's second maxim says that provability preserves truth. What he means by this is that the axiomatic theory of truth has to be closed under the necessitation rule. This of course falls under our disquotation and ascent constraint. So we regard this as a specific gradation of one of our desiderata. A stronger version of Sheard's second maxim would hold that every truth theory S should be closed under the following rule of inference:

$$\frac{\begin{array}{c}\varphi\\ \vdots\\ \psi\end{array}}{T\ulcorner\varphi\urcorner \rightarrow \ulcorner\psi\urcorner}$$

In the presence of a deduction theorem, this rule is a consequence of our specific desiderata of compositionality and identity between internal and external logic.

Simplicity is Sheard's third maxim. We know from the literature in philosophy of science that simplicity is a theoretical virtue for scientific theories in general that is very difficult to spell out with any degree of precision. Simplicity should certainly not be equated with proof-theoretic weakness and, in particular, not with conservativity over the base theory. To some extent it is covered by coherence: the axioms of a theory should somehow hang together. Nonetheless, also in the case of truth theories we find it hard to determine fully in which way theories ought to be simple.

Sheard's fourth maxim is a difficult one: it says that often a "local truth analysis" is sufficient. The meaning of this is not completely clear to us, but it largely says that it is not necessarily required of a truth theory that it captures all aspects and uses of the truth predicate. In other words, this maxim is saying that it should not be a presupposition that all axiomatic truth theories can and should be compared to the same standard. It might be that one truth theory captures one use of the concept of truth very well, and another captures another very well, whilst no decent truth theory captures both at the same time. In our view, Sheard is onto something important here. This is why we have been explicit at the outset about what we expect our axiomatic truth theories to do: to give a decent account of reflexive uses of truth. This is what many (but not all!) contemporary axiomatic theories of truth are concerned with. And we do not want to claim that this is the sole cluster of uses of truth that one might be interested in capturing axiomatically. So Sheard's fourth maxim is one we have attempted to deal with in the preamble to our list.

As a fifth maxim, Sheard postulates the infinitary closure of the truth predicate. This seems a very specific requirement. It is captured by the formula

$$\forall x T^{\ulcorner}\varphi(\dot{x})^{\urcorner} \rightarrow T^{\ulcorner}\forall x \varphi(x)^{\urcorner}.$$

Of course this is a specific instance of our more general constraint of compositionality.

As a last and tentative maxim, Sheard contemplates requiring truth to be non-conservative. But he notes that this "almost always" follows from the previous maxim. He therefore in any case does not want to list it as a *fundamental* desideratum; we agree that it should be regarded as derived at most. But we take issue with Sheard's contention that it follows almost always from the previous maxim on Sheard's list (which is in our view also only a derived norm). It follows "almost always" from compositionality and the unrestricted presence of T in the induction scheme. But we have emphasized that there may be sound reasons for restricting the presence of truth in the induction scheme. So we cannot without qualification claim that non-conservativeness is even a derived norm for truth theories.

In general, our dissatisfaction with Sheard's list stems from the fact that his maxims are not fundamental enough.

12.4.2 Comparison with Leitgeb

Leitgeb compiles a list of norms that is significantly longer than ours and then goes on to discuss particular axiomatic truth theories as implementations of maximal

consistent subsets of his list of norms. The norms on his list are mostly much closer to ours than Sheard's maxims. Let us take them in turn.

Leitgeb's first norm says that truth should be treated as a predicate, and his third norm says that truth should be treated as a type-free notion. Both of these belong to our background assumptions.

We skip Leitgeb's second norm for the nonce, and move on to number four on his list. This imperative says that good axiomatic truth theories should derive the unrestricted Tarski-biconditionals. This is a very strong form of our disquotation/ascent desideratum, but not necessarily the strongest one. Unrestricted substitution of φ and $T(\ulcorner\varphi\urcorner)$ in formulas of $\mathcal{L}_T$ is a stronger version, at least if certain weak underlying logics are assumed. It is one of the virtues of Field's theory of truth that it even satisfies the unrestricted substitutivity requirement of Field 2008.

Leitgeb's fifth norm is compositionality. As we have seen, this is also one of our norms.

The sixth norm on Leitgeb's list requires the existence of standard interpretations. We have discussed this norm above under our coherence norm.

Leitgeb's eighth norm requires the outer logic to be classical. This is a strong form of sustaining ordinary reasoning.

Norm seven on Leitgeb's list surprises us. It requires internal and external logic to coincide. It follows from norm four and norm eight. As intimated earlier, we see norm seven as a not-so-strong version of the disquotation / ascent desideratum. We suspect that Leitgeb lists norm seven because norm four, from which it can be derived in the presence of norm eight, is in effect by many researchers rejected on the ground of being excessively strong.

Let us, to conclude, turn to Leitgeb's norm two: it says that the truth theory should prove the truth of the background theory. In a weak form this can be derived from Leitgeb's requirement that the inner logic should coincide with the outer logic. But the uniform statement that all theorems of PA are true is of course a reflection principle from which nonconservativeness follows. We believe that Sheard is right in saying that this should not be a fundamental norm. If it is a norm at all, then it should derive from more fundamental considerations (involving considerations about the presence of truth in the induction scheme), and it is not so clear whether it follows without qualification.

In sum, we like Leitgeb's list better than Sheard's list. We agree with most of what is on its list, although we would regard many of his norms as strong instantiations of more general desiderata. But one desideratum is missing from Leitgeb's list just as it is missing from Sheard's list: we (or better: one of us!) want an axiomatic truth theory to capture a philosophical picture. And just as with Sheard's list, we object to the presence of non-conservativeness as an (almost) fundamental desideratum.

12.5 Applications

Supposing now that our list is a correct and complete list of fundamental norms for theories of reflexive truth: How can they then be applied?

It would be unreasonable to expect that by testing axiomatic theories of truth on our five dimensions, the question about which is (or are) the most satisfactory theory (theories) of reflexive truth can eventually be settled. One reason is that it is not completely clear how it can be measured how high a given theory scores on a given dimension. But a second, and perhaps more fundamental reason, is that different researchers will attach different weights to a given dimension. It is a familiar fact of the present situation that some researchers attach much value to ability to sustain ordinary reasoning, whereas other researchers attach much less weight to it. This is of course only to be expected, and it is an exact analogue of the situation concerning theoretical virtues of scientific theories. Some will put much stock on empirical adequacy, whereas others will be content to sacrifice some empirical accuracy to fruitfulness. So there are fundamental limitations on the use as a methodological tool of the norms discussed in this article.

This is not to say that judgements about weights attached to individual dimensions are not amenable to rational discussion. In fact, as we have remarked earlier, many researchers may be unwilling to accept that satisfying some particular criterion on the list to a high degree is desirable at all. Many researchers will take many of the unrestricted Tarski-biconditionals to be simply false, for instance. The most we can say is that everyone should agree that a satisfactory axiomatic system of reflexive truth should satisfy the norms on our list to a "reasonable" degree (except possibly the meta-requirement of giving a philosophical account).

There is a sense in which the norms discussed in this paper can be taken to demarcate an arena within which interesting formal theories of truth can be developed without stifling new research. Recall that Leitgeb has the requirement that inner logic coincides with external logic on his list but also the unrestricted Tarski-biconditionals and a demand for classicality in the outer logic. As we have remarked earlier, just classicality and the unrestricted Tarski-biconditionals would have met the case just as well. Proceeding in the way that he does, plays a large role in his identification of the main existing theories of reflexive truth as instantiations of the maximal consistent sublists of his list in a way that exhausts the maximal consistent sublists. His article is in a sense working towards this result. We do not aim at covering the maximal consistent degrees of satisfying our list with existing truth theories. We want to leave open the possibility and indeed hope that novel ways of having a mix of all the five norms to a reasonable degree will give rise to new interesting axiomatic theories of reflexive truth.

There is a related sense in which our methodological investigation might be of use for research 'on the ground'. Suppose we have an established and influential theory of reflexive truth that has a decidedly low score on some of the dimensions in the list, whilst having a high score on other dimensions on the list. Then we can attempt to modify the theory in such a way that we improve the score on the 'weak'

dimension(s) whilst avoiding to push the scores on the 'strong' dimensions down significantly. Here is one brief example of how this can play out.

The Kripke–Feferman theory (KF), introduced by Feferman 1991 under the name Ref(PA), is one of the most popular axiomatic theories of reflexive truth. It does not score well on the disquotation / ascent dimension (since its inner logic does not coincide with its outer logic), whereas it does reasonably well on the other dimensions. The theory PKF proposed and investigated by Halbach and Horsten 2006 is close to KF, in that it is also an axiomatization of Kripke's truth theory. But PKF has unrestricted substitution of formulas φ by $T(\ulcorner\varphi\urcorner)$ and vice versa: thus it scores very well on the disquotation / ascent dimension. But of course there is a price to be payed. PKF is formulated in partial logic. So the question is whether the reduction of sustaining of ordinary reasoning is worth paying in exchange for improvement in on the disquotation dimension.

12.6 Conclusion

From the outside, it might look as if the field of axiomatic theories of reflexive truth is a methodologically strongly constrained enterprise. We hope that the present article has at least shown that this is far from the case. The methodological principles that are operative in this field are vague, variably adhered to, and pull in opposite directions. In this sense the field of axiomatic theories of reflexive truth does not differ from any other philosophical discipline.

References

Boolos, G. (1971). The iterative conception of set. *Journal of Philosophy,* 68, 215–231.

Burgess, A. G., & Burgess, J. P. (2011). *Truth*. Princeton: Princeton Foundations of Contemporary Philosophy. Princeton University Press.

Burgess, J. P. (2009). Friedman and the axiomatization of Kripke's theory of truth. Ohio State, 2009. paper delivered at the Ohio State University conference in honor of the 60th birthday of Harvey Friedman.

Cieśliński, C. (2007). Deflationism, conservativeness and maximality. *Journal of Philosophical Logic,* 36, 695–705.

Feferman, S. (1984). Towards useful type-free theories I. *Journal of Symbolic Logic,* 49, 75–111.

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

Field, H. (2008). *Saving truth from paradox*. Oxford: Oxford University Press.

Friedman, H. & Sheard, M. (1987). An axiomatic approach to self-referential truth. *Annals of Pure and Applied Logic,* 33, 1–21.

Fujimoto, K. (2012). Classes and truths in set theory. *Annals of Pure and Applied Logic* (to appear).

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. (1999). Disquotationalism and infinite conjunctions. *Mind,* 108, 1–22.

Halbach, V. (2006). How not to state the T-sentences. *Analysis,* 66, 276–280 (Correction of printing error in vol. 67, 268).

Halbach, V. (2011). *Axiomatic theories of truth*. Cambridge: Cambridge University Press.
Halbach, V., & Horsten, L. (2006). Axiomatizing Kripke's theory of truth. *Journal of Symbolic Logic,* 71, 677–712.
Horsten, L. (2011). *The Tarskian turn: Deflationism and axiomatic truth*. Cambridge: MIT Press.
Kripke, S. (1975). Outline of a theory of truth. *Journal of Philosophy,* 72, 690–712 (reprinted in 1984).
Leitgeb, H. (2007). What theories of truth should be like (but cannot be). In *Blackwell Philosophy Compass 2/2*, pages 276–290. Blackwell, 2007.
Martin, R. L. (ed.). (1984). *Recent essays on truth and the liar paradox*. Oxford, New York: Clarendon Press, Oxford University Press.
McGee, V. (1985). How truthlike can a predicate be? A negative result. *Journal of Philosophical Logic,* 14*,* 399–410.
McGee, V. (1992). Maximal consistent sets of instances of Tarski's schema (T). *Journal of Philosophical Logic,* 21, 235–241.
Sheard, M. (1994). A guide to truth predicates in the modern era. *Journal of Symbolic Logic,* 59*,* 1032–1054.
Sheard, M. (2002). Truth, probability, and naive criteria. In V. Halbach & L. Horsten (ed.), *Principles of truth*. Frankfurt a. M.: Dr. Hänsel-Hohenhausen.
Terzian, G. (2012). *Uncovering the norms of truth. A meta-theoretic inquiry*. PhD thesis, University of Bristol.

Chapter 13
Some Weak Theories of Truth

Graham E. Leigh

Abstract In this article we present a number of axiomatic theories of truth which are conservative extensions of arithmetic. We isolate a set of ten natural principles of truth and prove that every consistent permutation of them forms a theory conservative over Peano arithmetic.

It seems that each axiomatic theory of truth that occurs in the literature is equipped with its own unique proof-theoretic ordinal. The four most studied theories, the compositional theory of truth CT, Friedman and Sheard's theory of finitely iterated truth FS, the strongly compositional Kripke–Feferman theory KF, and Cantini's super-valuationist theory VF, occupy the four ordinals ϵ_{ϵ_0}, $\phi 20$, $\phi\epsilon_0 0$ and the Bachmann–Howard ordinal respectively (see Halbach 1994; Feferman 1991; Cantini 1990). Moreover, theories of truth have been proposed corresponding to many other ordinals of proof-theoretic interest, such as ϵ_0, $\phi\omega 0$ and Γ_0 (*cf.* Halbach 1999; Leigh and Rathjen 2010; Feferman 1991 respectively). Such a plethora of ordinals motivates the following question:

What assumptions about the nature of truth are responsible for deciding the proof-theoretic strength of a theory of truth?

Aside from mathematical interest, an answer to this question also has philosophical applications. Horsten (1995), Ketland (1999) and Shaprio (1998), for example, each suggest that a deflationary account of truth should be at least conservative over its base theory. To the supporter of this view, an answer to the question above will outline which collections of truth-theoretic principles are acceptable to the deflationist and which combinations are to be avoided.

The proof-theoretic analysis of the theories mentioned earlier already provide a partial answer to the question. It is known, for example, that the sub-theory of FS

This research was supported by an Arts and Humanities research council grant number AH/H039791/1. The author would like to thank Kentaro Fujimoto for his helpful comments.

G. E. Leigh
Institute of Discrete Mathematics and Geometry, Vienna University of Technology, Vienna, Austria
e-mail: graham.leigh@tuwien.ac.at

T. Achourioti et al. (Eds.), *Unifying the Philosophy of Truth,* Logic, Epistemology, and the Unity of Science 36, DOI 10.1007/978-94-017-9673-6_13

consisting of the principles Elim and U-Inf (respectively, the rule stating that from a proof that ϕ is true one can infer ϕ, and the predicate version of the Barcan formula) proves the same arithmetical statements as full FS, over a suitably weak base theory (Sheard 2001). Although proof-theoretically equivalent to FS, this sub-theory is weak from a truth-theoretic perspective as it does not contain the axioms expressing that truth commutes with the logical connectives and existential quantifier, which are present in FS. In fact it can be shown that it is *minimal* in this regard: Dropping either the principle Elim or U-Inf from the axiom list results in a theory conservatively extending Peano arithmetic. Similarly, while the theory VF has axioms expressing "it is true that truth is a consistent concept" and "a sentence is true only if it is true that it is true", it is only the principles Out (the axiom schema "if ϕ is true then ϕ") and U-Inf that are required to witness the arithmetic strength of VF (Friedman and Sheard 1987). Again this sub-theory is minimal with this property.

A systematic approach to the proof-theoretic analysis of theories of truth was proposed by (Friedman and Sheard 1987) and undertaken in (Leigh and Rathjen 2010). All maximal consistent sets of a collection of twelve principles of truth (see Table 13.1) were analysed for their proof-theoretic content over a weak base theory of truth. This base theory, denoted $\mathsf{Base}_{\mathrm{T}}$, is truth-theoretically innocent in the sense that it only stipulates that the set of sentences falling under the truth predicate is deductively closed and comprises primitive recursive arithmetic. Thus any desirable properties of truth such as compositionality or non-trivial instances of the T-schema must arise from additional principles.

The analysis of (Leigh and Rathjen 2010) determines the proof-theoretic strength of all axiomatisations of truth over $\mathsf{Base}_{\mathrm{T}}$ + U-Inf based on the principles listed in Table 13.1. In the present paper our attention is focused on the rôle of the truth quantifier axiom U-Inf in fixing the proof-theoretic strength of a theory of truth. We hope that by doing so we can shed further light on the subtle interactions between principles of truth and their effect on the truth-free consequences of a theory.

One future avenue of application is in the truth versus satisfaction debate. The language of arithmetic stands out in that it possesses a name for every element in the standard model. Therefore, a sentence $\forall x \phi(x)$ of arithmetic is "satisfiable" iff the sentence $\phi(t)$ is "true" for every closed term t, and truth is synonymous with satisfaction. On the contrary, in Zermelo–Fraenkel set theory, or even in natural language, this phenomenon is not present and it can be difficult to motivate (or, in some cases, even formulate) the principle "a sentence $\forall x \phi(x)$ is true iff $\phi(s)$ is true of all objects s." Although the results of this paper apply only to theories of truth over arithmetic it should not be difficult to lift them to other base theories, thereby giving a proof-theoretic account of the difference between truth and satisfaction in settings in which the conflation of the two concepts is questionable.

Before we proceed it is worth noting that $\mathsf{Base}_{\mathrm{T}}$ is not a sub-theory of all theories of truth. For instance the truth predicate of KF (and its variants such as Halbach and Horsten's PKF or Burgess' μKF) is necessarily non-classical. That said, $\mathsf{Base}_{\mathrm{T}}$ is

a sub-theory of a large number of theories found in the literature, including CT,[1] FS, VF and of course all the theories isolated by Friedman and Sheard (Friedman and Sheard 1987). Therefore we believe $\mathsf{Base}_{\mathrm{T}}$ is still an acceptable base theory for addressing the problem at hand. Moreover, in line with (Leigh and Rathjen 2012) the following can be read as a precursor to a more general analysis dealing with theories of truth formulated in non-classical logic.

Let $\mathcal{L}$ denote the language of arithmetic which we assume contains function symbols for all primitive recursive functions. Let T is a fresh unary predicate symbol not in $\mathcal{L}$ and set $\mathcal{L}_{\mathrm{T}} = \mathcal{L} \cup \{\mathrm{T}\}$. We assume that the axiomatisation of PA contains the defining equations for each function symbol in $\mathcal{L}$. PA_{T} denotes the theory PA formulated in the expanded language $\mathcal{L}_{\mathrm{T}}$ including induction for $\mathcal{L}_{\mathrm{T}}$ formulae. For the remainder of this paper we fix some primitive recursive coding $\ulcorner . \urcorner$ of $\mathcal{L}_{\mathrm{T}}$ formulae. Associated to this coding are primitive recursive functions $\dot{\rightarrow}$, $\dot{\vee}$, $\dot{\forall}$, etc. such that for $\phi, \psi \in \mathcal{L}_{\mathrm{T}}$, $\ulcorner \phi \urcorner \dot{\rightarrow} \ulcorner \psi \urcorner = \ulcorner \phi \rightarrow \psi \urcorner$, $\ulcorner \phi \urcorner \dot{\vee} \ulcorner \psi \urcorner = \ulcorner \phi \vee \psi \urcorner$, etc., as well as the usual substitution function $subn(m, n)$ defined so that $subn(m, n) = \ulcorner \phi(\bar{n}) \urcorner$ if m is the code of an $\mathcal{L}_{\mathrm{T}}$ formula $\phi(x)$ with at most x free and $\bar{n}$ represents the n-th numeral, and $subn(m, n) = \ulcorner \bar{0} = \bar{1} \urcorner$ otherwise. Let $\mathtt{Sent}_{\mathcal{L}}(x)$ (resp. $\mathtt{Sent}_{\mathcal{L}_{\mathrm{T}}}(x)$) be an arithmetical predicate expressing that x is the code of a formula of $\mathcal{L}$ (respectively $\mathcal{L}_{\mathrm{T}}$). Both predicates are primitive recursive, so can be chosen to be Δ^0_0. For a recursively axiomatisable theory S, $\mathtt{Bew}_{\mathsf{S}}(x)$ is one of the standard provability predicates for S of complexity Σ^0_1.

We now introduce some useful abbreviations. Lower case Greek symbols ϕ, ψ, etc. range over codes of $\mathcal{L}_{\mathrm{T}}$-formulae, lower case Roman symbols x, y, z, etc. are intended to range over natural numbers. Thus the quantifier $\forall \phi \ldots$ is an abbreviation for $\forall z(\mathtt{Sent}_{\mathcal{L}_{\mathrm{T}}}(z) \rightarrow \ldots)$, and $\forall \phi(x)$ (or $\forall \ulcorner \phi(x) \urcorner$ for the sake of readability) abbreviates $\forall z(\mathtt{Sent}_{\mathcal{L}_{\mathrm{T}}}(subn(z, \bar{0})) \rightarrow \ldots)$. The notation $\phi(\dot{x})$ will abbreviate the term $subn(\phi, x)$, or $subn(\ulcorner \phi \urcorner, x)$ where appropriate. Finally, a formula is called *arithmetical* if it is a formula of $\mathcal{L}$.

Definition 1 $\mathsf{Base}_{\mathrm{T}}$ is defined to be the theory of truth formulated in $\mathcal{L}_{\mathrm{T}}$ whose axioms are: (i) the axioms of Peano arithmetic; (ii) the schema of induction for all formulae in $\mathcal{L}_{\mathrm{T}}$; (iii) the axioms

$$\forall \phi \forall \psi ((\mathrm{T}(\phi \dot{\rightarrow} \psi) \wedge \mathrm{T}\phi) \rightarrow \mathrm{T}\psi),$$
$$\forall \phi (\mathtt{Ax}_{\mathsf{PRA}}\phi \rightarrow \mathrm{T}ucl\phi),$$
$$\forall \phi (\mathtt{val}\phi \rightarrow \mathrm{T}ucl\phi),$$

where $\mathtt{Ax}_{\mathsf{PRA}}(x)$ is a predicate expressing that x is (the code of) a true equation involving primitive recursive functions, $ucl(\phi)$ is (the code of) the universal closure of ϕ, and $\mathtt{val}(x)$ expresses that x is a valid formula of first-order logic.

$\mathsf{Base}_{\mathrm{T}}$ was utilised as a base theory of truth in (Friedman and Sheard 1987 ;Halbach 1994; Leigh and Rathjen 2010). It is sufficiently weak that $\mathsf{Base}_{\mathrm{T}} + \mathsf{Cons}$ is

[1] If formulated with a type-free truth predicate.

Table 13.1 Principles of truth

Name	Axiom schema
Out	$\mathrm{T}^{\ulcorner}\phi(\dot{x_1},\ldots,\dot{x_n})^{\urcorner} \to \phi(x_1,\ldots,x_n)$
In	$\phi(x_1,\ldots,x_n) \to \mathrm{T}^{\ulcorner}\phi(\dot{x_1},\ldots,\dot{x_n})^{\urcorner}$ (ϕ with at most $x_1,\ldots,x_n$ free)
Name	Axiom
U-Inf	$\forall\phi(x)(\forall z\mathrm{T}(\phi(\dot{z})) \to \mathrm{T}(\dot{\forall}x\phi))$
E-Inf	$\forall\phi(x)(\mathrm{T}(\dot{\exists}x\phi) \to \exists z\mathrm{T}(\phi(\dot{z})))$
Cons	$\forall\phi\neg(\mathrm{T}\phi \wedge \mathrm{T}(\dot{\neg}\phi))$
Comp	$\forall\phi(\mathrm{T}\phi \vee \mathrm{T}(\dot{\neg}\phi))$
Rep	$\forall\phi(\mathrm{T}\phi \to \mathrm{T}(^{\ulcorner}\mathrm{T}\dot{\phi}^{\urcorner})$
Del	$\forall\phi(\mathrm{T}(^{\ulcorner}\mathrm{T}\dot{\phi}^{\urcorner})) \to \mathrm{T}\phi)$
Name	Inference rule
Intro	from ϕ infer $\mathrm{T}^{\ulcorner}\phi^{\urcorner}$
Elim	from $\mathrm{T}^{\ulcorner}\phi^{\urcorner}$ infer ϕ
¬Intro	from $\neg\phi$ infer $\neg\mathrm{T}^{\ulcorner}\phi^{\urcorner}$
¬Elim	from $\neg\mathrm{T}^{\ulcorner}\phi^{\urcorner}$ infer $\neg\phi$

a conservative extension of arithmetic. A simple embedding of the theory into PA is obtained by interpreting the truth predicate as provability in $\mathsf{I}\Sigma_1$, the sub-theory of PA in which the induction schema applies to Σ_1^0 formulae only.

For two theories S and T, we say S *conservatively extends* T if all theorems of S in the language of T are derivable in T and this fact is provable in $\mathsf{I}\Sigma_1$. A theory S is *proof-theoretically* equivalent to a theory T if S conservatively extends the set of arithmetical theorems of T and vice versa.

For a set S of principles from Table 13.1, $\mathsf{Base}_{\mathrm{T}} + \mathsf{S}$ denotes the theory whose axioms and rules are those of $\mathsf{Base}_{\mathrm{T}}$ and elements of S. It is convenient to identify the set S with the theory $\mathsf{Base}_{\mathrm{T}} + \mathsf{S}$. Suppose S, T are subsets of the 12 principles laid out in Table 13.1. Then T is said to be *(proof-theoretically) minimal over S* if $\mathsf{Base}_{\mathrm{T}} + (\mathsf{T} \cup \mathsf{S})$ is consistent and for each $\mathsf{R} \subsetneq \mathsf{T}$, there is an arithmetical statement provable in $\mathsf{Base}_{\mathrm{T}} + (\mathsf{T} \cup \mathsf{S})$ but not in $\mathsf{Base}_{\mathrm{T}} + (\mathsf{R} \cup \mathsf{S})$. T is said to be *minimal* if it is minimal over $\emptyset$. A set T is *maximal* if for every principle P listed in Table 13.1, $\mathsf{Base}_{\mathrm{T}} + (\mathsf{T} \cup \{\mathsf{P}\})$ is not proof-theoretically equivalent to $\mathsf{Base}_{\mathrm{T}} + \mathsf{T}$.

The main theorem of this paper can be stated as follows.

Theorem 1 *Each of the sets* {*Cons*, *U-Inf*}, {*Elim*, *U-Inf*}, {*Elim*, *Intro*, *Del*, *U-Inf*} *and* {*Out*, *U-Inf*} *is proof-theoretically minimal. Moreover, the following nine sets of truth principles are the only maximal sets whose induced theory over* $\mathsf{Base}_{\mathrm{T}}$ *conservatively extends PA.*

A. In, Intro, ¬Elim, Del, Rep, Comp, U-Inf, E-Inf.
B. Rep, Cons, Comp.

C. Del, Cons, Comp.
D. Intro, Elim, ¬Intro, ¬Elim, Cons, Comp.
E. Intro, Elim, ¬Intro, Del, Cons.
F. Intro, Elim, ¬Elim, Del.
G. Intro, Elim, ¬Elim, Rep.
H. Out, Elim, ¬Intro, Del, Rep, Cons.
I. Elim, ¬Elim, Rep, Del.

The following theorems combine the known results in the proof-theoretic analysis of theories of truth. Theorem 2 is due to (Friedman and Sheard 1987), and theorems 3 and 4 are from (Leigh and Rathjen 2010).

Theorem 2 *Over* $\mathsf{Base}_{\mathrm{T}}$*, the axioms of* A *are inconsistent with all other principles in Table 13.1; the axioms of* B*,* C *and* D *are each inconsistent with all other principles apart from* U-Inf *and* E-Inf*; the axioms of* E–I *are each inconsistent with all other principles except* U-Inf*.*

Theorem 3 *The addition of the axiom* U-Inf *to any of the theories* B–I*, or the addition of* E-Inf *to the theories* B–D *results in a theory proof-theoretically stronger than* PA*.*

Proof We first note that over $\mathsf{Base}_{\mathrm{T}}$, U-Inf is a consequence of the conjunction of Comp, Cons and E-Inf. Secondly, the two theories $\mathsf{Base}_{\mathrm{T}}$ + U-Inf + Cons and $\mathsf{Base}_{\mathrm{T}}$ + U-Inf + Elim each prove the formalised consistency statement for PA (corollary 2.13 and theorem 2.41 of Leigh and Rathjen 2010 respectively). Since any of the theories listed in the statement of the theorem will extend one of these two theories, we are done. ■

Theorem 4 *The following four sets are proof-theoretically minimal over* {U-Inf}*:* {Cons}*;* {Elim}*;* {Elim, Intro, Del}*; and* {Out}*. Moreover, any other set of truth principles is either inconsistent or proof-theoretically equivalent over* $\mathsf{Base}_{\mathrm{T}}$+U-Inf *to one of these four minimal sets.*

Proof Friedman and Sheard (1987) establish that the set {Out} is minimal over {U-Inf}. That {Elim} is minimal over {U-Inf} is due to Sheard (2001). The remaining two minimality claims and the second part of the theorem are corollaries of results in (Leigh and Rathjen 2010). ■

Theorems 2 and 3 yield the following.

Corollary 1 *Suppose the theories* A *to* I *listed in theorem 1 are conservative extensions of* PA*. Then the defining sets are maximal and the sets* {Cons, U-Inf}*,* {Elim, U-Inf}*,* {Elim, Intro, Del, U-Inf} *and* {Out, U-Inf} *are all minimal.*

The remainder of the paper serves to prove that the theories A to I are conservative extensions of PA and hence establish theorem 1. That A is a conservative extension of PA was established in (Leigh and Rathjen 2010, cor. 2.8), while the conservativity of D over PA was proved in (Friedman and Sheard 1987, §7). The result for D also suffices to manage B and C as they are interpretable in D.

Lemma 1 *Every arithmetical theorem of either* B *or* C *is derivable in* D.

Proof We begin with B. Let f_{B} be a primitive recursive function such that for any sentence ϕ of the language $\mathcal{L}_{\mathrm{T}}$, $f_{\mathsf{B}}(\ulcorner\phi\urcorner)$ is the code for the sentence resulting from substituting in ϕ each sub-formula of the form $\mathrm{T}(s)$ by $s = s$. If x is not the code of a sentence we assume $f_{\mathsf{B}}(x) = x$. Let $\mathrm{T}^{\mathsf{B}}(x)$ denote $\mathrm{T}(f_{\mathsf{B}}(x))$.

As instances of Comp and Cons, both $\forall\phi(\mathrm{T}^{\mathsf{B}}\phi \vee \mathrm{T}^{\mathsf{B}}\dot{\neg}\phi)$ and $\forall\phi\neg(\mathrm{T}^{\mathsf{B}}\phi \wedge \mathrm{T}^{\mathsf{B}}\dot{\neg}\phi)$ are theorems of D. Moreover, $\mathsf{Base}_{\mathrm{T}} \vdash \forall\phi\, \mathrm{T}^{\mathsf{B}}(\ulcorner \mathrm{T}^{\mathsf{B}}(\dot{\phi})\urcorner)$, so

$$\mathsf{Base}_{\mathrm{T}} \vdash \forall\phi(\mathrm{T}^{\mathsf{B}}\phi \rightarrow \mathrm{T}^{\mathsf{B}}(\ulcorner \mathrm{T}^{\mathsf{B}}\dot{\phi}\urcorner)).$$

Finally, $\mathsf{Base}_{\mathrm{T}^{\mathsf{B}}}$ is a sub-theory of $\mathsf{Base}_{\mathrm{T}}$, so B can be embedded into D by interpreting $\mathrm{T}(s)$ as $\mathrm{T}^{\mathsf{B}}(s)$. Since this interpretation leaves truth-free formulae unchanged, the desired result is obtained.

C is the dual of B; the axioms of C are all derivable in D for the predicate T^{C} defined as $\mathrm{T}(f_{\mathsf{C}}(x))$ where $f_{\mathsf{C}}(\ulcorner\phi\urcorner)$ is the code of the sentence resulting from substituting in ϕ each sub-formula of the form $\mathrm{T}(s)$ by $s \neq s$. ■

Corollary 2 *D, B and C are each conservative extensions of* PA.

Proof Lemma 1 entails that every arithmetical theorem of B and C is derivable in D, and by (Friedman and Sheard 1987), D is a conservative extension of PA. Note that the proof of lemma 1 can be easily carried with $\mathsf{I}\Sigma_1$ as the background theory.■

We now turn our attention to the theory F. Define an hierarchy of theories:

$$\mathsf{F}_0 = \mathsf{PA}_{\mathrm{T}},$$
$$\mathsf{F}_{m+1} = \mathsf{Base}_{\mathrm{T}} + \mathsf{Del} + \{\mathrm{T}(\ulcorner\phi\urcorner) : \mathsf{F}_m \vdash \phi \text{ and } \phi \text{ is an } \mathcal{L}_{\mathrm{T}}\text{-sentence}\}.$$

We claim that for each m,

1. F_m is a conservative extension of PA, and
2. $\mathsf{F} \vdash \phi$ entails $\mathsf{F}_m \vdash \phi$ for some m.

Establishing (i) is straightforward: If $\mathsf{F}_m \vdash \phi$, then $\mathsf{PA} \vdash \phi^{\dagger}$, where† is translation that maps $\mathrm{T}(s)$ to $s = s$ and commutes with all connectives and quantifiers, leaving arithmetical formulae unchanged. In order to prove (ii), we are required to show that each theory F_m is closed under Elim and $\neg\mathsf{Elim}$.

Lemma 2 *Let* $\mathfrak{F}_0$ *be the* $\mathcal{L}_{\mathrm{T}}$*-structure* $\langle\mathbb{N}, \emptyset\rangle$*, and for each* k*, let* $\mathfrak{F}_{k+1}$ *be the structure* $\langle\mathbb{N}, \{\ulcorner\phi\urcorner : \mathsf{F}_k \vdash \phi \wedge \mathtt{Sent}_{\mathcal{L}_{\mathrm{T}}}\ulcorner\phi\urcorner\}\rangle$*. Then for each* $m < \omega$*,*

1. $\mathfrak{F}_m \models \mathsf{F}_m$,
2. F_m *is closed under* Elim,
3. F_m *is a sub-theory of* F_{m+1}.

Proof (i) is immediate if $m = 0$, as is (ii) and (iii). If $m = n + 1$, to show (i) it suffices to prove $\mathfrak{F}_m \models \mathsf{Del}$ which is an immediate consequence of the induction hypothesis for (ii). Given (i), if $\mathsf{F}_m \vdash \mathrm{T}(\ulcorner\phi\urcorner)$ then $\mathsf{F}_n \vdash \phi$, so $\mathsf{F}_m \vdash \phi$ by (iii), and thus (ii) holds. Finally, (iii) is a direct consequence of the induction hypothesis. ■

Let $\widetilde{\mathsf{F}}$ be the theory whose axioms are given by

$$\mathtt{Ax}(\widetilde{\mathsf{F}}) = \bigcup_{m<\omega} \{\phi : \mathsf{F}_m \vdash \phi\}.$$

Note that $\mathtt{Ax}(\widetilde{\mathsf{F}})$ is closed under logical deduction, whence $\widetilde{\mathsf{F}}$ forms a theory closed under Intro and extending $\mathsf{Base}_{\mathrm{T}} + \mathsf{Del}$. Moreover, lemma 2 implies $\widetilde{\mathsf{F}}$ is closed under Elim. Since the $\mathcal{L}_{\mathrm{T}}$-structure $\langle \mathbb{N}, \mathbb{N} \rangle$ naturally forms a model of $\widetilde{\mathsf{F}}$, we also determine that $\widetilde{\mathsf{F}}$ is closed under $\neg\mathsf{Elim}$, and so extends F.

Combining the previous lemma with the observations made earlier, it is clear that F proves no more arithmetical statements than PA. The only step left is to determine that the proof (and statement) of lemma 2 can be formalised within $\mathsf{I}\Sigma_1$. This is not a simple task however, as there is no obvious manner in which to define the structures $\mathfrak{F}_k$ within PA. Let $k \geq 0$. Our attempt sees us interpret the predicate "$\mathfrak{F}_{k+1} \models \phi$" as "$\mathsf{PA} \vdash \phi^*$" for a suitably chosen interpretation $^* : \mathcal{L}_{\mathrm{T}} \to \mathcal{L}$. The translation * is defined to mimic the definition of $\mathfrak{F}_{k+1}$, namely we ensure that $\mathtt{T}(s)^* \to \mathtt{Bew}_{\mathsf{F}_k}(s)$ is derivable for all terms s.

Following the description laid out above, the reflection principle implicit in the definition of $\mathfrak{F}_{k+1}$, that $\mathfrak{F}_{k+1} \models \mathtt{T}\ulcorner\phi\urcorner$ entails $\mathsf{F}_k \vdash \phi$, is interpreted as the conditional

$$\mathsf{PA} \vdash \mathtt{T}(\ulcorner\phi\urcorner)^* \Rightarrow \mathsf{F}_k \vdash \phi, \tag{13.1}$$

which again is required to be derivable within $\mathsf{I}\Sigma_1$. As such, it becomes clear that $\mathtt{T}(s)$ cannot be interpreted as simply $\mathtt{Bew}_{\mathsf{F}_k}(s)$.

In order to maintain (13.1) in the formalised setting, we introduce a further level of stratification in the proof of lemma 2, this time according to levels of induction. Let F^n_k be a representation of the theory F_k but with $\mathsf{I}\Sigma^{\mathrm{T}}_n$ as the underlying theory, not PA. Formally, let $\mathsf{Base}^n_{\mathrm{T}}$ be the sub-theory of $\mathsf{Base}_{\mathrm{T}}$ in which the schema of induction is restricted to only formulae in the class Σ^0_n (including the truth predicate).

$$\begin{aligned} \mathsf{F}^n_0 &= \mathsf{Base}^n_{\mathrm{T}}, \\ \mathsf{F}^n_{k+1} &= \mathsf{Base}^n_{\mathrm{T}} + \mathsf{Del} + \forall\phi(\mathtt{Bew}_{\mathsf{F}^n_k}\phi \wedge \mathtt{Sent}_{\mathcal{L}_{\mathrm{T}}}\phi \to \mathtt{T}\phi). \end{aligned}$$

The following proposition establishes that for each k and n, F^n_k is a conservative extension of $\mathsf{I}\Sigma_n$. What it does not establish, however, is that $\mathsf{F}^n_k \vdash A$ for some n and k whenever $\mathsf{F} \vdash A$. This will be proven in lemma 3. Let $\mathtt{Bew}^n_k(s)$ abbreviate the provability predicate for F^n_k.

Proposition 1 *Let f_{B} be the interpretation of $\mathcal{L}_{\mathrm{T}}$ into $\mathcal{L}$ defined in the proof of lemma 1. Then $\mathsf{I}\Sigma_1 \vdash \forall k \forall n \forall x (\mathtt{Bew}^n_k(x) \to \mathtt{Bew}_{\mathsf{I}\Sigma_n}(f_{\mathsf{B}}(x)))$.*

Proof Let p be the primitive recursive function such that if d encodes a deduction of ϕ in F^n_k, then $p(d)$ is a deduction of ϕ^* in $\mathsf{I}\Sigma_1$, where * is the translation represented by f_{B}. Formalising this argument is possible in $\mathsf{I}\Sigma_1$, yielding the proposition. ■

Lemma 3 *The following are derivable in $\mathsf{I}\Sigma_1$.*

i) $\forall n \forall k \forall \phi \forall \psi (\mathtt{Bew}^n_k(\phi \dot{\to} \psi) \to (\mathtt{Bew}^n_k\phi \to \mathtt{Bew}^n_k\psi))$
ii) $\forall n \forall k \forall \phi (\mathtt{Bew}^n_k\phi \to \mathtt{Bew}^n_{k+1}\phi)$.

Proof (i) is simply the first derivability condition that holds of standard provability predicates, while (ii), formalising that F^n_k is a sub-theory of F^n_{k+1}, is proved by induction on k. ■

For each k let g^n_{k+1} be a primitive recursive function representing the interpretation that maps $\mathtt{T}(s)$ to $\mathtt{Bew}^n_k(s)$, commutes with all connectives and quantifiers and leaves arithmetical formulae unchanged and set $g^n_0 = f_{\mathsf{B}}$. Let $\mathtt{ClTerm}(x)$ be a predicate that holds just in case x encodes a closed term of $\mathcal{L}$, and *val* a recursive function such that $val(s) = n$ if s is the code of a closed term s with value n.

Lemma 4 *The following holds for each $n > 0$ and k.*

1. $\mathsf{I}\Sigma_1 \vdash \forall\phi(\mathtt{Bew}^n_k\phi \to \mathtt{Bew}_{\mathsf{I}\Sigma_{n+k+1}}(g^n_k\phi)$.
2. $\mathsf{I}\Sigma_{n+k+2} \vdash \forall s(\mathtt{ClTerm}(s) \wedge \mathtt{Bew}^n_k(\ulcorner \mathtt{T}(s) \urcorner) \to \mathtt{Bew}^n_k(val(s)))$.

Proof Fix $n > 0$. The proof proceeds by induction on k. For the case $k = 0$, (i) amounts to proving $\mathsf{I}\Sigma_1 \vdash \mathtt{Bew}^n_0\phi \to \mathtt{Bew}_{\mathsf{I}\Sigma_{n+1}}(g^n_0\phi)$ which is a consequence of proposition 1. Using the interpretation f_{C} defined in lemma 1, one can also show $\forall s \neg \mathtt{Bew}^n_0(\ulcorner \mathtt{T}(s) \urcorner)$ within $\mathsf{I}\Sigma_1$, hence (ii) holds.

Now suppose $k = l + 1$. We begin with (i) and argue informally within $\mathsf{I}\Sigma_1$. Let * be the interpretation represented by g^n_k. It is required to show that $\mathsf{F}^n_k \vdash \phi$ entails $\mathsf{I}\Sigma_{n+k+1} \vdash \phi^*$, which is established via induction on the length of the former derivation.

We deal with the case ϕ is an axiom of F^n_k; the other cases are straightforward. If ϕ is an instance of induction in $\mathsf{Base}^n_{\mathrm{T}}$ then ϕ^* is an instance of induction in $\mathsf{I}\Sigma_{n+1}$. If ϕ is an axiom of $\mathsf{Base}_{\mathrm{T}}$ other than Imp, $\mathsf{I}\Sigma_{n+1} \vdash \phi^*$ is obvious, and lemma 3 provides for the case that ϕ is Imp. By the main induction hypothesis, $\mathsf{I}\Sigma_{n+k+1} \vdash \forall s(\mathtt{Bew}^n_l(\ulcorner \mathtt{T}(s) \urcorner) \to \mathtt{Bew}^n_l(val(s)))$, so $\mathsf{I}\Sigma_{n+k+1} \vdash (\mathsf{Del})^*$. Finally, if ϕ is the remaining axiom of F^n_k, ϕ^* is vacuous true.

By (i), $\mathsf{F}^n_k \vdash \mathtt{T}(s)$ entails $\mathsf{I}\Sigma_{n+k+1} \vdash \mathtt{Bew}^n_l(s)$. But the uniform Σ^0_1-reflection principle for $\mathsf{I}\Sigma_{n+k+1}$ is derivable in $\mathsf{I}\Sigma_{n+k+2}$, that is

$$\mathsf{I}\Sigma_{n+k+2} \vdash \forall x(\mathtt{Bew}_{\mathsf{I}\Sigma_{n+k+1}}(\ulcorner \phi(\dot{x}) \urcorner) \to \phi(x))$$

for every Σ^0_1 formula ϕ with at most x free, whence $\mathsf{I}\Sigma_{n+k+2} \vdash \forall s(\mathtt{ClTerm}(s) \wedge \mathtt{Bew}^n_k(\ulcorner \mathtt{T}(s) \urcorner) \to \mathtt{Bew}^n_k(val(s)))$, as required. ■

Theorem 5 *F is a conservative extension of PA.*

Proof We argue informally within $\mathsf{I}\Sigma_1$. Suppose $\mathsf{F} \vdash \phi$ and ϕ is arithmetical. By lemma 4 there exists k such that $\mathsf{PA} \vdash \mathtt{Bew}^{\tilde{k}}_{\tilde{k}}(\ulcorner \phi \urcorner)$ holds, whence lemma 4(i) implies $\mathsf{PA} \vdash \mathtt{Bew}_{\mathsf{I}\Sigma_{2\tilde{k}+1}}(\ulcorner \phi \urcorner)$ and so $\mathsf{PA} \vdash \phi$ by reflection. As this argument can be carried out within $\mathsf{I}\Sigma_1$, we are done. ■

A similar argument can also be used in the analysis of E, axiomatised by $\mathsf{Base}_{\mathrm{T}} + \mathsf{Intro} + \mathsf{Elim} + \neg\mathsf{Intro} + \mathsf{Del} + \mathsf{Cons}$. Define a hierarchy of theories,

$$\mathsf{E}_0 = \mathsf{PA}_{\mathrm{T}},$$
$$\mathsf{E}_{k+1} = \mathsf{Base}_{\mathrm{T}} + \mathsf{Del} + \mathsf{Cons} + \{\mathtt{T}(\ulcorner \phi \urcorner) : \mathsf{E}_k \vdash \phi\}.$$

By a proof analogous to lemma 2, it can be shown that $\bigcup_m \mathsf{E}_m$ forms a consistent theory closed under Intro, Elim and ¬Intro, and so extends E. The presence of Cons means that interpreting truth in E_{k+1} as provability in E_k requires first determining that E_k is consistent. We proceed directly with the formalised version of the consistency proof.

Theorem 6 *E is a conservative extension of PA.*

Proof Let ${}^*\mathtt{Bew}_k^n$ be a provability predicate for E_k^n. For each $n > 0$ and every k, define

$$\mathsf{E}_0^n = \mathsf{Base}_{\mathrm{T}}^n,$$
$$\mathsf{E}_{k+1}^n = \mathsf{Base}_{\mathrm{T}}^n + \mathsf{Del} + \mathsf{Cons} + \forall\phi({}^*\mathtt{Bew}_k^n\phi \wedge \mathtt{Sent}_{\mathcal{L}_{\mathrm{T}}}\phi \rightarrow \mathtt{T}\phi).$$

Pick primitive recursive functions f_k^n representing the interpretation that maps $\mathtt{T}s$ to ${}^*\mathtt{Bew}_{k-1}^n s$ if $k > 0$ and that maps $\mathtt{T}s$ to $\mathtt{Bew}_{\mathsf{I}\Sigma_1}(s)$ otherwise. By expanding lemma 4, we can show

i) $\mathsf{I}\Sigma_1 \vdash \forall\phi({}^*\mathtt{Bew}_k^n\phi \rightarrow {}^*\mathtt{Bew}_{\mathsf{I}\Sigma_{n+k+1}}(f_k^n\phi)$,
ii) $\mathsf{I}\Sigma_{n+k+2} \vdash \forall s(\mathtt{ClTerm}(s) \wedge {}^*\mathtt{Bew}_k^n(\ulcorner \mathtt{T}s \urcorner) \rightarrow {}^*\mathtt{Bew}_k^n s)$,
iii) $\mathsf{I}\Sigma_{n+k+2} \vdash \forall\phi\neg({}^*\mathtt{Bew}_k^n\phi \wedge {}^*\mathtt{Bew}_k^n(\dot{\neg}\phi))$.

(iii) is a consequence of the formalised consistency statement for $\mathsf{I}\Sigma_{n+k+1}$ and suffices to determine that the interpretation of Cons is derivable for each n and k. A consequence of (ii) is that E_k^n is closed under Elim, whence we deduce that $\mathsf{E} \vdash A$ iff $\mathsf{E}_k^n \vdash A$ for some n and k (notice ¬Intro is derivable from Intro and Cons). Therefore E is a conservative extension of PA. ■

To analyse G, we define an hierarchy of theories,

$$\mathsf{G}_0 = \mathsf{Base}_{\mathrm{T}} + \mathsf{Rep} + \{\mathtt{T}\ulcorner\phi\urcorner : \phi \text{ is an } \mathcal{L}_{\mathrm{T}}\text{-sentence}\},$$
$$\mathsf{G}_{m+1} = \mathsf{Base}_{\mathrm{T}} + \mathsf{Rep} + \{\mathtt{T}\ulcorner\phi\urcorner : \mathsf{G}_m \vdash \phi \wedge \mathtt{Sent}_{\mathcal{L}_{\mathrm{T}}}\phi\},$$

along with $\mathcal{L}_{\mathrm{T}}$-structures $\mathfrak{G}_0 = \langle\mathbb{N}, \mathbb{N}\rangle$ and $\mathfrak{G}_{m+1} = \langle\mathbb{N}, \{\ulcorner\phi\urcorner : \mathsf{G}_m \vdash \phi\}\rangle$.

Lemma 5 *For each $m < \omega$,*

1. $\mathfrak{G}_m \models \mathsf{G}_m$,
2. G_m *is closed under* Intro,
3. G_{m+1} *is a sub-theory of* G_m.

Proof Proceed by induction on m. This is immediate if $m = 0$, so suppose $m = n + 1$. By the induction hypothesis for (ii), we know G_n is closed under Intro, so $\mathfrak{G}_m \models \mathsf{Rep}$; therefore $\mathfrak{G}_m \models \mathsf{G}_m$. If $\mathsf{G}_m \vdash \phi$ then $\mathtt{T}\ulcorner\phi\urcorner$ is an axiom of G_{m+1}, whence $\mathsf{G}_m \vdash \mathtt{T}\ulcorner\phi\urcorner$ by the induction hypothesis, and (ii) holds. Finally, (iii) results directly from the induction hypothesis which implies every axiom of G_{m+1} is an axiom of G_m. ■

Theorem 7 *G is a conservative extension of PA.*

Proof Define a theory $\widetilde{\mathsf{G}}$ whose axioms are given by

$$\mathrm{Ax}(\widetilde{\mathsf{G}}) = \{\phi : \mathsf{G}_n \vdash \phi \wedge \mathtt{Sent}_{\mathcal{L}_\mathrm{T}}\ulcorner\phi\urcorner \text{ for every } n\}.$$

$\mathrm{Ax}(\widetilde{\mathsf{G}})$ is closed under deduction, so if ϕ is a sentence then $\widetilde{\mathsf{G}} \vdash \phi$ implies $\phi \in \mathrm{Ax}(\widetilde{\mathsf{G}})$. As such $\widetilde{\mathsf{G}}$ forms a theory closed under Intro. Moreover, if $\widetilde{\mathsf{G}} \vdash \mathrm{T}\ulcorner\phi\urcorner$, we may deduce $\mathsf{G}_n \vdash \mathrm{T}\ulcorner\phi\urcorner$ for every n and so $\mathsf{G}_n \vdash \phi$ for every n by lemma 5(i), while $\widetilde{\mathsf{G}}$ is closed under ¬Elim since $\mathfrak{G}_0 \models \widetilde{\mathsf{G}}$.

Thus $\mathsf{G} \vdash \phi$ implies $\widetilde{\mathsf{G}} \vdash \phi$ for every $\mathcal{L}_\mathrm{T}$-sentence ϕ. Since G_0 is naturally a sub-theory of A and $\widetilde{\mathsf{G}}$ is a sub-theory of G_0, one easily obtains that G proves no more arithmetical statements than PA. Finally, we note that this argument can be formalised within PA in much the same way as for F. ■

The theory H can be axiomatised by Base_T, Out and Rep; all other axioms and rules are derivable from Out. The addition of U-Inf to H results in a theory proof-theoretically equivalent to the theory of one inductive definition, ID_1 (Cantini 1990). In analogy with E and F, we will interpret truth in H as provability, so Out corresponds to a reflection principle.

Theorem 8 *H is a conservative extension of PA.*

Proof Let $\mathsf{S} = \mathsf{I}\Sigma_1 + \mathsf{Intro}$ and define an interpretation $^* : \mathcal{L}_\mathrm{T} \to \mathcal{L}$ that commutes with all quantifiers and connectives, leaves arithmetical formulae unchanged and interprests $\mathrm{T}(s)$ as $\mathtt{Bew}_\mathsf{S}(s)$. Then $\mathsf{PA} \vdash (\mathsf{Imp})^* \wedge (\mathsf{Rep})^*$ is immediate. By proving

$$\mathsf{PA} \vdash \forall x[\mathtt{Bew}_\mathsf{S}(\ulcorner\phi(\dot{x})\urcorner) \to \phi^*(x)] \tag{13.2}$$

for each formula ϕ of $\mathcal{L}_\mathrm{T}$ with at most x free, we may easily conclude $\mathsf{PA} \vdash \phi^*$ whenever $\mathsf{H} \vdash \phi$.

In order to establish (13.2) we must first analyse the structure of proofs within S. For this we argue informally within PA. S can be formulated in a Tait-style sequent calculus in which one has the usual rules $(\wedge)$, $(\vee)$, $(\exists)$, $(\forall)$ and (Cut), together with a rule (Intro) and the induction schema as an axiom,

(Intro) $\vdash \phi$ entails $\vdash \Gamma, s \neq \ulcorner\phi\urcorner, \mathrm{T}s$
(Ind) $\vdash \Gamma, \neg\phi(\bar{0}), \neg\forall x(\phi(x) \to \phi(x+1)), \phi(x)$

whenever Γ is a finite set of $\mathcal{L}_\mathrm{T}$-sentences and ϕ is a Σ^0_1 formula of $\mathcal{L}_\mathrm{T}$. Due to the fact that induction in S is restricted to only Σ^0_1 formulae, this system supports partial cut-elimination: Explicitly, if $\mathsf{S} \vdash \Gamma$, there is a derivation of Γ in S in which all cut formulae are contained in $\Sigma^0_2 \cup \Pi^0_2$. As usual this cut-elimination argument is formalisable within PA.

Let $\mathtt{Bew}^2_\mathsf{S}(x)$ be the predicate formalising provability in this sequent calculus for S with cuts only on Σ^0_2 or Π^0_2 formulae. Formally, the cut elimination argument above yields

$$\forall\phi(\mathtt{Bew}_\mathsf{S}\phi \to \mathtt{Bew}^2_\mathsf{S}\phi).$$

For each k, let $\mathtt{Tr}_k$ denote a formal truth predicate for arithmetical Σ^0_k formulae. Moreover, let $\mathtt{Sent}_k(x)$ express that x is the code of a Σ^0_k sentence. We then derive

$$\forall\phi(\mathtt{Bew}^2_{\mathsf{S}}\phi \wedge \mathtt{Sent}_k\phi \rightarrow \mathtt{Tr}_{k+4}\phi^*),$$

for each k. This can be seen by the following argument. Assume $S \vdash \Gamma$, where Γ is a sequent containing only Σ^0_k sentences. Denote by Γ^* the sequent $\{\phi^* : \phi \in \Gamma\}$ and let $\bigvee\Gamma$ be the disjunction of all elements of Γ.

If Γ is an axiom of S then either $\mathtt{Tr}_1(\ulcorner\bigvee\Gamma^*\urcorner)$ holds, or $\Gamma = \Gamma', \neg\phi(\bar{0}), \neg\forall x(\phi(x) \rightarrow \phi(x+1)), \phi(x)$ is an instance of the induction axiom with ϕ a Σ^0_1 formula. In the latter case ϕ^* is at most Σ^0_2 and there is a sequent $\Delta \subseteq \Gamma$ such that $\bigvee\Delta^*$ is Σ^0_4 and arithmetical, and $\mathtt{Tr}_4(\ulcorner\bigvee\Delta^*\urcorner)$ holds, whence also $\mathtt{Tr}_4(\ulcorner\bigvee\Gamma^*\urcorner)$.

If Γ is derived by a rule other than (Cut) or (Intro), one easily obtains $\mathtt{Tr}_{k+4}(\ulcorner\bigvee\Gamma^*\urcorner)$ by the induction hypothesis. In the case the last rule was (Cut), the cut formula is Σ^0_3, and so at most Σ^0_4 under the interpretation ϕ^*, whence the induction hypothesis may be applied and one obtains $\mathtt{Tr}_{k+4}(\ulcorner\bigvee\Gamma^*\urcorner)$.

This leaves only applications of (Intro) to be concerned with, so suppose Γ contains $\{\mathtt{T}s, s \neq \ulcorner\phi\urcorner\}$ for some sentence ϕ, and $\mathsf{S} \vdash \phi$. Let $\psi = \mathtt{T}s$. The complexity of ϕ may, of course, be larger than that of Γ, but $\mathtt{Tr}_{k+4}(\ulcorner s \neq \ulcorner\phi\urcorner \vee \psi^*\urcorner)$ holds by definition, so $\mathtt{Tr}_{k+4}(\ulcorner\bigvee\Gamma^*\urcorner)$ holds as required. Thus

$$\mathsf{PA} \vdash \mathtt{Bew}_{\mathsf{S}}\phi \wedge \mathtt{Sent}_k\phi \rightarrow \mathtt{Tr}_{k+4}\phi^*$$

holds for each k. Since $\mathtt{Tr}_k$ is a truth predicate for Σ^0_k formulae, we obtain $\mathsf{PA} \vdash \forall x(\mathtt{Bew}_{\mathsf{S}}(\phi(\dot{x})) \rightarrow \phi^*(x))$ for every formula ϕ of $\mathcal{L}_{\mathrm{T}}$ (with at most x free), and $\mathsf{PA} \vdash (\mathsf{Out})^*$ is deduced. Finally, the above argument can be readily formalised within $\mathsf{I}\Sigma_1$, so H is a conservative extension of PA. ■

As a corollary of theorem 8, we also obtain the desired result for I, completing the proof of theorem 1.

Theorem 9 *I is a conservative extension of* PA.

Proof Since the $\mathcal{L}_{\mathrm{T}}$-structure $\langle\mathbb{N}, \mathbb{N}\rangle$ is a model of I, I does not derive any formulae of the form $\neg\mathtt{T}s$, whence I is vacuously closed under the rule $\neg\mathsf{Elim}$. It then follows that I is a sub-theory of H and so is conservative over PA. ■

References

Cantini, A. (1990). A theory of formal truth arithmetically equivalent to ID1. *Journal of Symbolic Logic, 55*(1), 244–259.

Feferman, S. (1991). Reflecting on Incompleteness. *Journal of Symbolic Logic, 56*(1), 1–49.

Friedman, H., & Sheard, M. (1987). An axiomatic approach to self-referential truth. *Annals of Pure and Applied Logic, 33,* 1–21.

Halbach, V. (1994). A system of complete and consistent truth. *Notre Dame Journal of Formal Logic, 35*(3), 311–327.

Halbach, V. (1999). Conservative theories of classical truth. *Studia Logica*, 62(3), 353–370.

Horsten, L. (1995). The semantical paradoxes, the neutrality of truth and the neutrality of the minimalist theory of truth. In: Cortois, P. (ed.), *The many problems of realism (Studies in the general philosophy of science: Volume 3)* (pp. 173–187). Tilburg: Tilburg University Press.

Ketland, K. (1999). Deflationism and Tarski's paradise. *Mind, 109*(429), 69–94.

Leigh, G. E., & Rathjen, M. (2010). An ordinal analysis for theories of self-referential truth. *Archive for Mathematical Logic, 49*(2), 213–247.

Leigh, G. E., & Rathjen, M. (2012). The Friedman-Sheard programme in intuitionistic logic. *Journal of Symbolic Logic, 77*(3), 777–806.

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

Sheard, M. (2001). Weak and strong theories. *Studia Logica, 68,* 89–101.

Chapter 14
Deflationism and Instrumentalism

Martin Fischer

Abstract Deflationist theories of truth have become increasingly popular. In this paper we are going to pursue an instrumentalist reading of deflationism. This reading is particularly interesting considering the role a truth predicate can have in facilitating proofs. Based on an instrumentalist conception we will provide a new answer to one of the arguments challenging deflationism.

14.1 Introduction

In recent decades interest in deflationist accounts of truth has developed rapidly. Field and Horwich are two influential deflationist authors of the last part of the twentieth century. Although several theories have been suggested it is still not settled what exactly deflationism amounts to. There is a variety of different theories and conceptions that are covered by this terminology.

In this paper we are interested in conceptions of deflationism that are in line with the following two claims: First, the truth predicate has an irreducible expressive function. Second, this expressive function is the only function the truth predicate has.

The first claim is usually explained by referring to the truth predicate as an expressive device that allows generalizations which could not be expressed otherwise, e.g. 'All tautologies are true'. The second claim is supposed to capture the

Thanks to the organizers of the conference 'Axiomatic Theories of Truth' at Oxford for the invitation and to the members of the MCMP and participants of the Plurals, Predicates and Paradox Seminar for valuable comments, especially Johannes Stern, Sean Walsh and Florian Steinberger. The research was carried out in the project 'Syntaktische Ansätze für interagierende Modalitäten' financed by the DFG. For their financial support I wanted to thank the DFG and the Alexander von Humboldt foundation.

M. Fischer
MCMP, LMU-Munich, Germany
e-mail: martin_fischer@hotmail.com

T. Achourioti et al. (Eds.), *Unifying the Philosophy of Truth,* Logic, Epistemology, and the Unity of Science 36, DOI 10.1007/978-94-017-9673-6_14

non-substantiality of a deflationist truth predicate. It has also been suggested by several authors that this non-substantiality claim commits a deflationist to conservativeness.[1]

In the paper we want to sketch a version of deflationism which we shall label instrumentalist deflationism. The idea is to conceive of the truth predicate as an instrumental device. The introduction of a truth predicate is similar to the introduction of ideal elements in mathematics, as for example Hilbert understood it. The instrumental function of the truth predicate is then comparable to the instrumental function of ideal elements. This function consists mainly in facilitating proofs. Moreover the instrumentalist view on truth gives independent reasons for conservativeness. The similarities of the two philosophical approaches are discussed in the second part of this paper.

The axiomatic theory of truth introduced in the third part of the paper exemplifies an instrumental deflationist theory of truth. For this purpose we will use a positive theory of truth called PT^{-}. The instrumental function of the truth predicate in this theory is justified by a speed-up result.

In the last part of the paper we want to reconsider an argument against deflationism given by Shapiro and Ketland.[2] The argument aims at showing that a conservative theory of truth cannot be adequate because it cannot show the soundness of the base theory. We will argue that there is a version of soundness that can be expressed in a conservative theory of truth.

14.2 Deflationism and Instrumentalism

The deflationist conception of truth consists of a positive and a negative doctrine. According to the positive doctrine the truth predicate is an expressive device that allows us to express generalizations in a very natural way that we couldn't express without introducing more cumbersome devices.

The question what it exactly means for a theory of truth to enable the expression of generalizations has not been resolved, although there have been some interesting suggestions.[3] In this paper we will only focus on the simplest form of generalizations. Let's assume that we have codes $\ulcorner\varphi\urcorner$ for sentences φ and that there is a set of sentences, which we can refer to via a formula $\Theta(x)$. If we try to assert all sentences in Θ we could use a schematic form like $\Theta(\ulcorner\varphi\urcorner) \rightarrow \varphi$ for all sentences φ. But this is only an approximation and not an actual assertion of all the sentences in Θ in one utterance. If we have a truth predicate $T(x)$ at hand we can quite naturally express a generalization $\forall x(\Theta(x) \rightarrow T(x))$ that can be asserted and that implies all instances.

[1] See for instance Horsten (1995), Shapiro (1998) and Ketland (1999).
[2] Shapiro (1998) and Ketland (1999).
[3] See for example Halbach (1999).

The negative claim of deflationism characterizes the truth predicate as having only an expressive function. The literature contains claims that deflationist truth is non-substantial, or the truth predicate does not express a property. Shapiro, Ketland and others argued that this commits deflationists to a form of conservativeness, conservativeness of a theory of truth $T(B)$ over a base theory B, in our case PA. The conservativeness is understood either as proof-theoretic conservativeness, saying that for all formulas φ in the language of arithmetic $\mathcal{L}_A$ we have, if $T(\text{PA}) \vdash \varphi$, then $\text{PA} \vdash \varphi$. Or, we can understand it as model-theoretic conservativeness saying that every model of PA can be expanded to a model of $T(\text{PA})$. The conservativeness commitment of deflationism is not generally accepted and there are several authors who resist it such as Halbach (2011), Horsten (2011) and others.

Although it might be questionable whether deflationism in general is committed to conservativeness, there is no doubt that the specific version of deflationism we are interested in here, namely instrumental deflationism, is bound to conservativeness. Instrumental deflationism is a form of deflationism that takes truth as an instrumental predicate. It is built on the philosophical tradition of mathematical instrumentalism, as exemplified by Hilbert's program.

David Hilbert proposed his program as a new way for the foundations of mathematics. Central to the program is the idea to prove the consistency of mathematics. The need for consistency proofs is already apparent in his formulations of the open problems in mathematics in his talk from 1900, where he states the need for a new kind of proof of the consistency of arithmetic as the second problem. Hilbert's program is not a unique and clear-cut position. It developed over time and the most important formulations are in the 1920's where he introduced the distinction between 'real' and 'ideal' mathematics.[4]

The 'real' mathematics part is the safe part. The real statements are meaningful and have content. They are based on the conception of concrete signs, such as strokes and concern the finitary part. The real mathematics does not need a justification, because it is intuitively acceptable. This part also covers formalized proofs and is therefore called metamathematics or proof theory and should be used according to Hilbert to prove the consistency of the ideal part.

The 'ideal' part on the other hand contains transfinite elements that have to be justified. The real statements contain already 'problematic' statements. Examples are generalizations that have no negation in the realm of the real statements. The ideal statements are introduced in order regain the usual logical laws. These ideal statements have no meanings as long as they do not express finite statements. So in order to judge them they have to be formalized and their consistency has to be established. In other words the introduction of the ideal elements has to be proven innocent.

So in order to carry out Hilbert's program to show that all of classical mathematics is consistent, it would suffice to show that ideal mathematics is conservative over the real part and the conservativeness proof has to be carried out in the real part. This

[4] For example Hilbert (1923, 1926).

shift is also apparent in the second half of the 1920's in Hilbert's writings.[5] This idea led Feferman to the more general concept of proof-theoretic reducibility.

Whereas for Hilbert mathematics has to be formalized, it is not necessary to formalize metamathematics. Gödel developed methods to formalize metamathematics and others have tried to pin down what Hilbert's finite mathematics amounts to. Tait suggested PRA as a suitable candidate and Kreisel suggested the realm of PA-provably total functions.

Usually Gödel's theorems especially the second incompleteness theorems are taken to show that Hilbert's program is not realizable. The rough argument is this: The real mathematics is a part of mathematics. So to prove the consistency of all of mathematics in real mathematics one would have to prove the consistency of real mathematics in real mathematics, which is impossible according to Gödel's theorems. Moreover, if we take the conservativeness line of the program, then already Gödel's first incompleteness theorem is devastating as Smoryński (1977) argues convincingly.[6]

Although Gödel's theorems appear to be disastrous for Hilbert's program it is still alive. On the one hand there are modifications of the original program that seem more feasible. Gentzen's proof of the consistency of PA with transfinite induction and the development of modern proof theory of the Schütte school and ordinal analysis can be understood as a form of Hilbert's program with an extended finitism. But also the proof-theoretic reduction by Feferman, as well as the area of reverse mathematics are valuable successors. On the other hand there have been attempts to show the compatibility of Hilbert's program and Gödel's theorems. To this effect Detlefsen uses some kind of omega rule. This approach is based on ideas by Hilbert himself.[7] Those are different forms of Hilbert's program that are still influential in the philosophy of mathematics.

In the remainder we will focus on Feferman's notion of proof theoretic reducibility because we think that it is particularly suited for a deflationist theory of truth. As we already remarked Hilbert's program has two aspects which are connected and which can be separated as a consistency program and a conservation program.[8]

Let R be the system of real mathematics and I be the system of ideal mathematics. R contains means to codify derivations from I, like a proof predicate $Proof_I(x, y)$ and a provability predicate $Pr_I(x)$. With this we can also formulate consistency of I as $\forall x \neg Proof_I(x, \ulcorner 0 = 1 \urcorner)$, which is a Π^0_1-sentence. And we can formulate conservativeness of I over R, by $\forall x(Pr_I(x) \rightarrow Pr_R(x))$, which is not Π^0_1 anymore.

In this reading the consistency program consists in proving the consistency of I in R, which is $R \vdash Con_I$. So if I contains R, and R is axiomatizable, then we cannot hope to accomplish the consistency program in this form.

[5] For example Hilbert (1926, p. 179).
[6] For a different interpretation see Detlefsen (1990).
[7] See also Ignjatović (1994). Another line of reasoning is given in Niebergall and Schirn (2002).
[8] Compare Smoryński (1977).

The conservation program on the other hand may be realizable to a certain degree. Hilbert wanted to show by finitistic means that the statements of real mathematics which are derivable in ideal mathematics could already be derived by real mathematics. So to carry out the program one would have to show that $R \vdash \forall x(\Pi^0_1(x) \wedge Pr_I(x) \rightarrow Pr_R(x))$. But at this point we already encounter a problem. As was noted earlier the formalized conservativeness claim is not a Π^0_1-sentence and therefore possibly also not part of real mathematics.

According to Smoryński we could nevertheless fulfill the conservation program if we could fulfill the consistency program.[9] The idea is this: Assume that the consistency program is satisfied with respect to I, i.e. $R \vdash Con_I$. To show conservation we assume that $I \vdash \varphi$ for some Π^0_1 sentence φ. By representability we get $R \vdash Pr_I(\ulcorner \varphi \urcorner)$ and by consistency $R \vdash \neg Pr_I(\ulcorner \neg\varphi \urcorner)$. If we assume that R proves the Σ^0_1 completeness for I we also get $R \vdash \neg\varphi \rightarrow Pr_I(\ulcorner \neg\varphi \urcorner)$ and therefore $R \vdash \varphi$.

There is an alternative if we drop the restriction to real statements. There are two versions of PRA in use. One formulated as an equational theory and one formulated in the language of arithmetic with induction restricted to formulas containing only bounded quantifiers. The second version would allow us to carry out the conservation program for a small part of mathematics, such as for example $I\Sigma_1$ or WKL_0. This is the case because we have $\mathrm{PRA} \vdash Con_{\mathrm{PRA}} \leftrightarrow Con_{I\Sigma_1}$ and $\mathrm{PRA} \vdash Con_{\mathrm{PRA}} \leftrightarrow Con_{\mathrm{WKL}_0}$.[10]

On a more liberal version of instrumentalism, not bound to finitism, it is possible to allow for a different foundational theory. An alternative option is a local perspective rather than a global perspective by comparing various theories and their relations. This last option is Feferman's point of view. The notion of proof-theoretic reducibility is introduced by Feferman as follows:[11]

Definition 2.1 Let S, T, U be theories and Φ a set of formulas. S is proof-theoretically reducible to T, provable in U, with respect to formulas of Φ, iff $U \vdash \forall x(\Phi(x) \wedge Pr_s(x) \rightarrow Pr_t(x))$.

This notion of proof-theoretic reducibility seems a particularly suited version of mathematical instrumentalism for an interpretation of deflationism.

14.3 An Instrumentalist Theory of Truth

In this section we will sketch what an instrumentalist theory of truth could amount to and how it relates to deflationism.[12] Whereas we drop Hilbert's finitism we can still preserve a separation of a 'real' and an 'ideal' part in the investigation of truth. The

[9] Smoryński (1977, p. 824).
[10] Compare Caldon and Ignjatović (2005, p. 7).
[11] See for example Feferman (2000).
[12] The line taken in this paper is only one of many possible connections. As was pointed out by one of the referees Reinhardt (1986) had an interesting proposal of connecting theories of truth and Hilbert's program. Reinhardt compared Hilbert's ideal part with meaningful, but nonsignificant sentences, i.e. sentences such as the liar that are wellformed formulas but are neither true nor false. Reinhardt proposed to use the theory IKF, which is the inner logic of KF as the real part. Reinhardt's

underlying theory of syntax-in our case PA-has a special independent justification and is therefore a possible candidate for the real part. The extension of the arithmetical language by a truth predicate introduces an ideal element. In this case we do not introduce a new significant predicate that commits one to a new sort of entities, rather it allows for the formulation of new formulas, such as generalizations, analogous to the case that Hilbert describes, when going from a schematic version to a formula using free variables.[13]

The real part contains the substantial object theory. Statements in the language of the real part have meaning. For the object theory we can accept a full-blown truth-conditional semantics in which terms denote objects and predicates express properties.

In contrast to the real part the truth predicate does not express a property. Still the truth predicate has its use and the use is governed by the rules of inference or the axioms. So one could accept that the truth predicate has meaning in a less substantial way. Here the inferentialist picture may be helpful. The meaning of the truth predicate is given purely by the rules of inference or the axioms.[14]

This reading of the truth predicate fits the vague deflationist claims very well. The truth predicate is non-substantial in this sense and it does not express a property. It also goes well with Shapiro's demand of semantical conservativity. One could also read this as some kind of invariance result and try to establish truth as a logical device. What is added to this picture from the instrumentalist side is Feferman's proof-theoretic reducibility. So instrumentalist theories of truth are considered to be proof-theoretically reducible to their base theory. This guarantees that the truth predicate introduced is unproblematic, relative to the base theory.

Moreover the instrumental theory of truth should be useful. This usefulness has two aspects. The first is expressiveness. Analogous to the case of Hilbert's program we allow new forms of expressions. In Hilbert's case the restriction to real statements was severe and caused problems as the real statements were not closed under negation. The introduction of ideal statements allowed for the full apparatus of unrestricted quantification. And the move from a schematic version to a general statement seems to be structurally similar to the deflationist move from a schematic version to a generalization.

The second aspect is facilitating proofs. Whereas we do not want our instrumentalist theory of truth to prove new substantial theorems it should help in proving old theorems. This help is to be understood as a form of efficiency in establishing old theorems. Speed-up results exactly show that proofs can be significantly shortened. Moreover there are speed-up results even in the case of conservative extensions. A

proposal has some problems that were pointed out by Halbach and Horsten (2006). They show that many derivations of significant statements in IKF necessarily contain nonsignificant statements.

[13] Compare Hilbert (1926, p. 175).

[14] Horsten (2011) already suggested connections of truth and inferentialism in the context of deflationism.

speed-up result for a conservative theory of truth over PA precisely witnesses the kind of instrumental function we are looking for.[15]

There are interesting axiomatic theories of truth that are acceptable as instrumentalist theories. One of them is a theory of positive truth called PT^-. In the discussion that follows this specific theory is only one example out of a class of truth theories that seem to be acceptable as an instrumentalist theory of truth. The main criteria are proof-theoretic reducibility and speed-up.[16]

The language $\mathcal{L}_T$ is the language of arithmetic $\mathcal{L}_A$ expanded by a one place predicate $T(x)$. T is intended to be a truth predicate for the language $\mathcal{L}_A$. The axioms of PT^- are the following

$$(PT1) \qquad \forall s \forall t(T(s \doteq t) \leftrightarrow val(s) = val(t))$$

$$(PT2) \qquad \forall s \forall t(T(\dot{\neg}(s \doteq t)) \leftrightarrow \neg val(s) = val(t))$$

$$(PT3) \qquad \forall x \forall y(Snt(x \dot{\wedge} y) \rightarrow (T(x \dot{\wedge} y) \leftrightarrow T(x) \wedge T(y)))$$

$$(PT4) \qquad \forall x \forall y(Snt(\dot{\neg}(x \dot{\wedge} y)) \rightarrow (T(\dot{\neg}(x \dot{\wedge} y)) \leftrightarrow T(\dot{\neg} x) \vee T(\dot{\neg} y)))$$

$$(PT5) \qquad \forall x(Snt(\dot{\forall} y\, x) \rightarrow (T(\dot{\forall} y\, x) \leftrightarrow \forall z T(x(\dot{z}))))$$

$$(PT6) \qquad \forall x(Snt(\dot{\neg}\dot{\forall} y\, x) \rightarrow (T(\dot{\neg}\dot{\forall} y\, x) \leftrightarrow \exists z T(\dot{\neg} x(\dot{z}))))$$

$$(PT7) \qquad \forall x(Snt(\dot{\neg}\dot{\neg} x) \rightarrow (T(\dot{\neg}\dot{\neg} x) \leftrightarrow T(x)))$$

$$(PT8) \qquad \forall x(T(x) \rightarrow Snt(x))$$

$$(PT9) \qquad \forall s \forall t \forall x(val(s) = val(t) \rightarrow (T(x(s)) \leftrightarrow T(x(t))))$$

$$tot(x) :\Leftrightarrow Fml^1(x) \wedge \forall y(T(x(\dot{y})) \vee T(\dot{\neg} x(\dot{y}))).$$

The induction axiom is a form of internal induction restricted to total formulas:

$$(I_t I) \qquad \forall x(tot(x) \wedge T(x(0)) \wedge \forall y(T(x(\dot{y})) \rightarrow T(x(y \dot{+} 1))) \rightarrow \forall y T(x(\dot{y})))$$

$$\mathrm{PT}^- := \mathrm{PA} \cup (PT1) - (PT9) \cup (I_t I)$$

PT^- is a typed theory of truth because the truth predicate is only intended for sentences not containing the truth predicate. An alternative typefree theory introduced by Cantini (1989) is KF_t. PT^- is contained in KF_t. The proof-theoretic reducibility of PT^- to PA can be established via a theorem by Cantini (1989) that shows that KF_t is proof-theoretically reducible to PA.

One of the special features of the theories PT^- and KF_t is the form of induction they share. In contrast to other theories induction is formulated as an axiom and not as a schema which allows for a finite axiomatization. So in this respect they are similar to subsystems of second-order arithmetic with an induction axiom. This form of restriction seems justified for an instrumentalist notion of truth, whereas

[15] Caldon and Ignjatović (2005) combine mathematical instrumentalism and speed-up.
[16] Also KF_t and $\mathrm{CT}\restriction + $ 'all axioms of PA are true', are viable candidates.

for a substantial notion a restriction of induction appears unmotivated. From an instrumentalist point of view the truth predicate has no meaning in itself and does not express a property. But we can make use of it in a context in which we can show it to be unproblematic. And those are cases in which we can prove the formulas in question to be total, such that we can use the usual logical laws.

In the case of type-free theories of truth the instrumentalist picture is naturally associated with the notion of groundedness. Only those sentences containing the truth predicate that are grounded in non-semantical facts have a meaning, and only a derivative meaning.

The usefulness of the truth predicate in the case of PT^- is exemplified by fact that it has super-exponential speed-up over PA.

Theorem 1 PT^- *has super-exponential speed-up over* PA.

A proof of this statement would go beyond the scope of the philosophical discussion in this paper.[17] We will only highlight the most important facts. The speed-up result says that there is a sequence of formulas φ_n that are provable in PT^- and PA for all $n \in \omega$. Moreover in the case of PT^- we can give a bound on the length of proofs polynomial in n. In contrast we can only give a super-exponential bound on the length of proofs in PA.

A first thing to notice is that the formulas φ_n that are used to establish the speed-up result are partial consistency statements of the form $Con_{pa}(\overline{n})$ defined as $\neg\exists x(x \leq \overline{n} \wedge Proof_{pa}(x, \ulcorner 0 = 1 \urcorner))$.

In mathematics those consistency statements are often considered to be not proper mathematical statements but rather artificial. But in the case of truth they are particularly interesting. This seems to suggest that the speed-up is quite significant in the case of truth.

Another interesting aspect worth explaining is that the derivations of partial consistency statements in PA are rather involved. There is no way to establish the partial consistency statements in PA as instances of a provable generalization. But in PT^- we do have a more general way. We can prove the consistency of PA on a cut. Cuts play an important role in the proof. A cut $C(x)$ in a theory U is a formula with one free variable, such that $U \vdash$

$$
\begin{array}{ll}
(i) & C(\overline{0}) \\
(ii) & \forall x(C(x) \rightarrow C(x+1)) \\
(iii) & \forall x \forall y(x < y \wedge C(y) \rightarrow C(x))
\end{array}
$$

A cut $C(x)$ is said to be proper if we cannot prove $\forall x C(x)$. The truth predicate allows us to define a proper cut $C(x)$ in PT^-, such that PT^- proves the consistency of PA on this cut, i.e. $\mathrm{PT}^- \vdash \forall x(C(x) \rightarrow Con_{pa}(x))$.

Cuts are initial segments containing the standard numbers. Whereas in the case of the standard model the extension of the cut formula is just the natural numbers,

[17] For details see Fischer (2014).

in the case of nonstandard models it might well be a proper subset of the domain. So consistency on a cut is a general version from which all the partial consistency statements for all $n \in \omega$ can be derived easily.

The speed-up result suggests a close connection between deflationism and instrumentalism. Caldon and Ignjatović (2005) understand their speed-up results for $I\Sigma_1$ over PRA as a partial realization of Hilbert's program. Moreover a lemma connected to the speed-up also opens up the possibility for a new answer to a challenge for conservative theories, which we will consider in the next section.

14.4 Truth and Reflection

The restriction to conservative extensions of PA has well-known limitations for a theory of truth. We want to address one of the main concerns in this respect which was raised for example by Ketland (1999) and Shapiro (1998). The deflationist truth predicate is a natural candidate for expressing the soundness of the base theory that we accept. The problem is that by Gödel's second incompleteness theorem a conservative extension of PA is not able to prove a standard consistency statement for PA. This also implies that a conservative theory of truth cannot prove a soundness statement in form of the global reflection principle $\forall x(Snt(x) \wedge Pr_{pa}(x) \rightarrow T(x))$.[18]

We will argue that the reasoning only commits one to a weaker version of reflection which a conservative theory is able to prove. In the following we will take a closer look at the argument that a deflationist theory of truth should prove the global reflection principle for its base theory. The first exposition follows one of the more convincing presentations of the argument given by Cieśliński (2010).

There are two steps in the argument:

(I) By accepting PA as a base theory one has the epistemic obligation to accept the soundness of PA.
(II) The global reflection principle is 'the' way to express this soundness claim.

The first step is a move from the acceptance of a set of sentences to the acceptance of all the consequences. Sometimes this is taken to be expressed by the local reflection principle Rfn_{pa}:

$$Pr_{pa}(\ulcorner\varphi\urcorner) \rightarrow \varphi, \text{ for } \varphi \text{ closed.}$$

And the epistemic obligation is then interpreted as the extension of PA by Rfn_{pa}. In the second step the argument goes on. Rfn_{pa} is a schematic principle and exactly of a form, such that a deflatonist truth predicate should enable a generalization. The global reflection principle is the generalization of Rfn_{pa}.

This version of the argument is convincing in so far as it tries to establish that it is reasonable for a theory of truth to prove the global reflection principle for its base

[18] Under the condition that the theory of truth satisfies some minimal criteria, such as containing all T-sentences for the arithmetical language.

theory. What it does not establish is that a theory of truth is obliged to prove it. There are various ways for a deflationist to resist the argument. We will sketch one option that has not been considered so far (to the best of the authors knowledge).

The first part of the solution consists in a reinterpretation of the epistemic obligation. It is reasonable that if one accepts a set of sentences, then one should accept also its theorems. This seems to suggest that if we accept Σ and $\Sigma \vdash \varphi$, then we should also accept φ. One way to formalize this is Rfn_{pa}. But it is not the only way to read it and maybe not even the most faithful way. Accepting all the theorems can equally be understood as saying that if we accept Σ and we have a proof $d : \Sigma \vdash \varphi$, then we should accept φ. Formalizing this amounts to the following principle, which we will call explicit reflection principle, following Artemov[19]:

$$Proof(\overline{d}, \ulcorner \varphi \urcorner) \rightarrow \varphi \text{ for all proofs } d \text{ and sentences } \varphi.$$

This schematic form allows only for standard proofs in contrast to the local reflection principle, in which we existentially quantify over the first position, $\exists x\, Proof(x, \ulcorner \varphi \urcorner)$. This formulation seems also to better fit Hilbert's conception of proof as a 'concrete and surveyable object'.[20]

This interpretation of the epistemic obligation allows for more options concerning generalizations which will be the second part of our solution. The global reflection principle is only one of the possibilities to capture all the instances of $Proof(\overline{d}, \ulcorner \varphi \urcorner) \rightarrow \varphi$ for all proofs d and sentences φ. Alternatives can be build on a notion of proof with a restriction on the length of the proofs, as long as all standard proofs are captured. This is the case for restrictions to a cut $C(x)$.

A restriction of the proof length is not artificial considering the intuition that the standard proofs are the intended proofs. The non-standard proofs are the artificial proofs that are to be excluded. We know that we cannot have a predicate that only allows for standard proofs, but provability on a cut seems to be a good approximation.

Viewed in this light, it is not at all clear that the global reflection principle is 'the' correct way to state soundness or the only one. PT^- cannot prove the global reflection principle in its general form. But if we restrict the proofs in a specific way as is done in the proof of consistency on a cut, then PT^- is able to prove a restricted version of global reflection, namely global reflection on a cut.

Theorem 2 *There is a cut $C(x)$ in PT^-, such that*

$$\mathrm{PT}^- \vdash \forall x \forall y (Snt(x) \wedge C(y) \wedge Proof_{pa}(y, x) \rightarrow T(x)).$$

Also for this theorem we do not provide a complete proof here.[21] The idea is that with the help of the truth predicate we can define a cut $C(x)$ in such a way that all the formulas in the cut are total. This allows us to use the truth predicate for those

[19] See Artemov (2001, p. 6).
[20] Compare 'Ein formalisierter Beweis ist ... ein konkreter und überblickbarer Gegenstand' Hilbert (1926, p. 179).
[21] For a proof see Fischer (2014).

formulas in a standard way. Moreover all the formulas in the cut have a restricted complexity, so that we can prove all axioms on this cut to be true. Since we can also prove that the rules are truth preserving on this cut, we can prove the global reflection on a cut in a rather standard way.

With this result and the fact that for all standard proofs d we can prove $C(\overline{d})$ it is easy to see that PT^- proves all instances of the explicit reflection principle. This observation allows us to understand global reflection on a cut as a generalization that is sufficient to express the soundness of the base theory. PT^- is an example of a conservative extension of PA that is still adequate in expressing the soundness of its base theory. By giving the soundness requirement a slightly weaker reading than proving global reflection the deflationist can answer the challenge by Shapiro and Ketland.

In the remaining paragraphs we will discuss a possible objection. Global reflection on a cut is built on a notion of provability that is non standard. Rosser provability and other non standard provability predicates show that extensional adequacy is not sufficient to guarantee for an intended behavior.

Since Gödel it is well known how to canonically define provability predicates for theories in an arithmetical theory, such as PA. Let pa be a Δ_1 formula strongly representing the set of axioms of PA. Then the canonical way to define provability and consistency is:

$$\begin{aligned} Proof_{pa}(x,y) :\leftrightarrow & Seq(x) \wedge (x)_{lh(x)\dot{-}1} = y \wedge \forall i < lh(x) \\ & (lAx((x)_i) \vee pa((x)_i) \vee \exists j,k < i((x)_j = (x)_k \dot{\rightarrow} (x)_i)) \\ Pr_{pa}(y) :\leftrightarrow & \exists x\, Proof_{pa}(x,y) \\ Con_{pa} :\leftrightarrow & \neg Pr_{pa}(\ulcorner 0 = 1 \urcorner) \end{aligned}$$

We also know that this is not the only way to represent the theorems of PA. Any formula $Pr^*(x)$ numerating the theorems of PA is called a provability predicate. But provability predicates can have very different properties and it would be nice to have a clear cut-criterion for provability predicates to be adequate or not. There are some suggestions for 'standard' provability predicates for PA.

A minimal requirement for the provability predicate to be 'standard' is that it is extensionally correct, i.e. that it is a representation of theoremhood. Usually the provability predicate should be Σ_1.

But as Feferman (1960) and others pointed out there are extensionally correct Σ_1 provability predicates which are not intended, such as the Rosser provability predicate. So moreover the provability predicates should be intensionally correct and those are often referred to as 'standard' provability predicates.

For an intensionally correct arithmetization we usually have the three Löb derivability conditions:

(Löb1) if $\mathrm{PA} \vdash \varphi$, then $\mathrm{PA} \vdash Pr_{pa}(\ulcorner \varphi \urcorner)$

(Löb2) $\mathrm{PA} \vdash Pr_{pa}(\ulcorner \varphi \urcorner) \wedge Pr_{pa}(\ulcorner \varphi \rightarrow \psi \urcorner) \rightarrow Pr_{pa}(\ulcorner \psi \urcorner)$

(Löb3) $\mathrm{PA} \vdash Pr_{pa}(\ulcorner \varphi \urcorner) \rightarrow Pr_{pa}(\ulcorner Pr_{pa}(\ulcorner \varphi \urcorner) \urcorner)$

These derivability conditions are important. One of the reasons for the importance of the Löb conditions is that they are sufficient conditions to establish Löb's theorem according to which the following holds: if PA $\vdash Pr_{pa}(\ulcorner\varphi\urcorner) \rightarrow \varphi$, then PA $\vdash \varphi$. And with Löb's theorem it is easy to derive Gödel's second incompleteness theorem.

It is not totally clear that the Löb derivability conditions are exactly the necessary and sufficient conditions a standard proof predicate should satisfy. (Löb 2) seems to be directly justified. We want our notion of provability to be closed under modus ponens. (Löb3) on the other hand is harder to intuitively justify but there is an indirect justification via formalized Σ_1 completeness, PA $\vdash \varphi \rightarrow Pr_{pa}(\ulcorner\varphi\urcorner)$ for all φ in Σ^0_1. Formalized Σ_1 completeness is a further criterion for standardness.

Guaspari and Solovay suggest the following plausible criterion:[22] A provability predicate Pr is a standard provability predicate iff it is a Σ^0_1 representation of the theorems of PA satisfying (Löb 2) and formalized Σ_1 completeness.

Remark: Every provability predicate that is provably equivalent to a standard provability predicate will satisfy the Löb derivability conditions.

There are at least two ways in which a provability predicate can be non-standard. Either the representation pa of the axioms of PA is more complex than Σ_1. In this case (Löb3) will no longer hold as we cannot use formalized Σ_1 completeness. Or, some different definition is used, such as in the Rosser provability predicate:

$$Proof^R_{pa}(x, y) :\leftrightarrow Proof_{pa}(x, y) \wedge \forall z < x(\neg Proof_{pa}(z, \dot{\neg} y)).$$

The Rosser provability predicate is non-standard as PA $\vdash Con^R_{pa}$. So we have that the Rosser provability predicate cannot satisfy the three Löb conditions. Moreover it can be shown that Rosser provability does neither satisfy (Löb 2) nor (Löb3) and therefore also not formalized Σ_1 completeness.[23] The Rosser provability predicate is also different from the usual provability predicate in that its fixed points are not necessarily unique. For example we have $0 = 0$ and $0 = 1$ both as fixed points of $Pr^R_{pa}(x)$.

For a theory S and a Σ^{exp}_0 representation s of the axioms of S we can introduce a proof predicate on a cut in the following way:[24]

$$Proof^C_s(x, y) :\leftrightarrow Proof_s(x, y) \wedge C(lh(x));$$
$$Pr^C_s(y) :\leftrightarrow \exists x\, Proof^C_s(x, y);$$
$$Con^C_s :\leftrightarrow \neg Pr^C_s(\ulcorner 0 = 1\urcorner).$$

For such provability predicates it is not clear whether they are adequate or not.

In our case the cuts are formulated with the help of the truth predicate and are therefore no longer strictly Σ_1. But the underlying proof predicate is standard as well

[22] See Guaspari and Solovay (1979, p. 83).
[23] See Guaspari and Solovay (1979).
[24] Σ^{exp}_0 allows only bounded quantifiers but the the exponentiation function. By Craig's theorem we can find for a deductively closed set of sentences T, which is Σ_1, a Σ^{exp}_0 set of axioms t, such that the deductive closure of t is T.

as the representation *pa* of the axioms of PA. Moreover the restriction of the proof length is not as artificial as the built-in consistency in case of Rosser provability.

The Löb conditions in their original form are also not satisfied in the case of provability on a cut but a slightly weaker version can be established:[25]

Lemma 4.1 *For each cut C in S, there is a cut C' in S, such that*

$$(i) \qquad S \vdash \varphi \Rightarrow S \vdash Pr_s^C(\ulcorner \varphi \urcorner)$$

$$(ii) \qquad S \vdash Pr_s^{C'}(\ulcorner \varphi \urcorner) \wedge Pr_s^{C'}(\ulcorner \varphi \rightarrow \psi \urcorner) \rightarrow Pr_s^C(\ulcorner \psi \urcorner)$$

$$(iii) \qquad S \vdash Pr_s^{C'}(\ulcorner \varphi \urcorner) \rightarrow Pr_s^C(\ulcorner Pr_s^{C'}(\ulcorner \varphi \urcorner) \urcorner)$$

The reason that one cut is not sufficient is that cuts are not closed under exponentiation and for the concatenation function we need faster growing functions than addition and multiplication. But those conditions are sufficient to get a strengthened version of Gödel's second incompleteness theorem:[26]

Theorem 3 (Pudlák). *Let S be a consistent theory containing $I\Sigma_1$, $C(x)$ a cut in S and s a Σ_0^{exp} definition of the axioms of S. Then S does not prove Con_s^C.*

The theorem says that for a sufficiently strong consistent theory S, S does not prove it's own consistency on a cut. So there will also be no theory proving it's own reflection on a cut. So provability on a cut is in this respect different from Rosser provability and closer to standard provability. This is also witnessed by the fact that we have a corresponding result for formalized Σ_1-completeness, namely:[27]

Lemma 4.2 *Let S be a consistent theory containing $I\Sigma_1$, $C(x)$ a cut in S and s a Σ_0^{exp} definition of the axioms of S. Let φ be Σ_0^{exp}, then there is a k such that*

$$S \vdash \varphi \rightarrow \exists z < 2_k Proof_s(z, \ulcorner \varphi \urcorner).$$

With this background information we can reconsider the objection that provability on a cut is not adequate because it is not intensionally correct. If we consider only provability predicates to be adequate that are standard in the sense of Guaspari and Solovay, then this objection would be applicable. But for the objection to be convincing it would be nice to have an argument why we should have such a strict criterion for provability predicates. If the reason for the strict standardness criterion is to exclude Rosser-like provability predicates that allow for a consistency proof, then it is too strict as provability on a cut does not allow for such a proof. Even if there are good reasons for using the strict notion of standard provability for a general investigation of provability, it is not at all clear that the same notion is relevant for deflationist purposes. And by satisfying the weaker Löb conditions provability on a cut does not only satisfy extensional correctness but captures also the intension at least partially. So for the objection to work one would have to provide good reasons why

[25] See Hájek and Pudlák (1993, Chap. III. 3)
[26] See Hájek and Pudlák (1993, Chap. III. 3).
[27] See Hájek and Pudlák (1993, Chap. III. 3).

only in the case of the original Löb conditions we can adequately express soundness and why the weaker conditions are not sufficient for this purpose.

To conclude, we think that an instrumentalist perspective allows for a new reading of deflationism that is worth exploring. This instrumentalist reading of truth is justified by speed-up results. Furthermore instrumentalism provides a philosophical story underpinning deflationist claims and also helps in answering challenges posed to deflationism.

References

Artemov, S. N. (2001). Explicit provability and constructive semantics. *Bulletin of Symbolic Logic, 7*(1), 1–36.

Caldon, P., & Ignjatović, A. (2005). On mathematical instrumentalism. *Journal of Symbolic Logic, 70*(3), 778–794.

Cantini, A. (1989). Notes on formal theories of truth. *Zeitschrift für mathematische Logik und Grundlagen der Mathematik, 35,* 97–130.

Cieśliński, C. (2010). Truth, conservativeness, and provability. *Mind, 119,* 409–422.

Detlefsen, M. (1990). On an alleged refutation of Hilbert's program using Gödel's first incompleteness theorem. *Journal of Philosophical Logic, 19,* 343–377.

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

Feferman, S. (2000). Does reductive proof theory have a viable rationale? *Erkenntnis, 53*(1/2), 63–96.

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

Guaspari, D., & Solovay, R. M. (1979). Rosser sentences. *Annals of Mathematical Logic, 16,* 81–99.

Hájek, P. & Pudlák, P. (1993) *Metamathematics of first-order arithmetic*. Berlin: Springer.

Halbach V. (1999). Disquotationalism and infinite conjunctions. *Mind, 108,* 1–22.

Halbach, V. (2011). *Axiomatic theories of truth*. Cambridge: Cambridge University Press.

Halbach, V.,& Horsten, L. (2006). Axiomatizing Kripke's theory of truth. *The Journal of Symbolic Logic, 71,* 677–712.

Hilbert, D. (1923). Die logischen Grundlagen der Mathematik. *Mathematische Annalen, 88,* 151–165.

Hilbert, D. (1926). Über das Unendliche. *Mathematische Annalen, 95,* 161–190.

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* (pp. 173–187). Tilburg: Tilburg University Press.

Horsten, L. (2011). *The Tarskian turn. Deflationism and axiomatic truth*. Cambridge: MIT Press.

Ignjatović, A. (1994). Hilbert's program and the omega-rule. *The Journal of Symbolic Logic, 59*(1), 322–343.

Ketland, J. (1999). Deflationism and Tarski's paradise. *Mind,* 108*,* 69–94.

Niebergall, K.-G., & Schirn, M. (2002). Hilbert's programme and Gödel's theorems. *Dialectica, 56*(4), 347–370.

Reinhardt, W. N. (1986). Some remarks on extending and interpreting theories with a partial predicate for truth. *The Journal of Philosophical Logic, 15,* 219–251.

Shapiro, S. (1998). Proof and truth: Through thick and thin. *The Journal of Philosophy, 95,* 493–521.

Smoryński, C. (1977). The incompleteness theorems. In J. Barwise (Ed.), *Handbook of mathematical logic* (pp. 821–865). Amsterdam: Fischer.

Chapter 15
Typed and Untyped Disquotational Truth

Cezary Cieśliński

Abstract We present an overview of typed and untyped disquotational truth theories with the emphasis on their (non)conservativity over the base theory of syntax. Two types of conservativity are discussed: syntactic and semantic. We observe in particular that TB—one of the most basic disquotational theories—is not semantically conservative over its base; we show also that an untyped disquotational theory PTB is a syntactically conservative extension of Peano Arithmetic.

15.1 Disquotational Truth Theories

Disquotationalists believe that the whole content of the notion of truth is captured by the so called schema T:

$$\text{(T)} \qquad \text{`}p\text{' is true iff } p.$$

The general intuition is that the result of adding "is true" to a name of a sentence (or an utterance, or a proposition) is equivalent—in some weaker or stronger sense—to the sentence (the utterance, the proposition) itself. Apart from that, no further explanation of the meaning of "is true" is needed. Thus e.g. Field in (1994) states that for a given person and an utterance u, "the claim that u is true (true-as-he-understands-it) is cognitively equivalent (for the person) to u itself (as he understands it)" (p. 250); and he proceeds to assert that "the cognitive equivalence of the claim that u is disquotationally true to u itself provides a way to understand disquotational truth independent of any nondisquotational concept of truth or truth conditions" (p. 251). Another example of a philosophical position arising from disquotational intuitions is Horwich's minimalism (see his (Horwich 1990)). According to Horwich, all the facts about truth can be explained on the basis of the so called "minimal theory", whose axioms have a disquotational form.[1] Horwich claims that the minimal theory fully

[1] It's worth mentioning that Horwich attributes truth to propositions, not sentences or utterances.

C. Cieśliński
University of Warsaw, Warsaw, Poland
e-mail: c.cieslinski@uw.edu.pl

T. Achourioti et al. (Eds.), *Unifying the Philosophy of Truth,* Logic, Epistemology, and the Unity of Science 36, DOI 10.1007/978-94-017-9673-6_15

characterizes the content of the notion of truth. In particular, to understand the truth predicate it is simply enough to be ready to accept the substitutions of the relevant T-schema. Accordingly, truth has no hidden nature which could and should be revealed by scientific enquiry (Horwich 1990, p. 2). In effect all the traditional, substantive conceptions of truth (like correspondence, coherence and warranted assertability theory) turn out to be useless at best and probably misleading.

In this context disquotationalists quite often cite Alfred Tarski, recalling his famous Convention T. Why are we ready to call Tarski's definition a definition of truth, and not of some other property? The reason is that it permits us to derive all the instances of the T-schema for sentences of the object language. Tarski gives us a clear hint: his predicates (for various languages) are *truth* predicates, because they satisfy the same condition of material adequacy, i.e. they conform to Convention T. After we recognize this, there is only one last step to be taken: one can declare that all the apparatus of classical semantics (compositional approach, involving an inductive characterization of reference and satisfaction) is really unnecessary in the context of our general project of explaining the intuitive notion of truth. "Truth has a certain purity" (see Horwich 1990, p. 12)—we can explain it directly in terms of T-schemata without any appeal to other semantic notions.

What's the form of the theory of truth most adequate to disquotational intuitions? The most direct approach consists in stipulating that the set of axioms of our theory of truth will take the form of a collection of chosen instantiations of a T-schema.[2] Of course due to the contradictions of the liar type, we can't have all substitutions on our list—some restrictions are necessary. Nevertheless, the disquotationalist will claim that no other sort of axioms is really needed.

There are two basic variants of disquotational theories of truth. One possibility consists in adopting as axioms the substitutions of the local T-schema:

(L) $\quad T(\ulcorner\varphi\urcorner) \equiv \varphi$

The second option is to adopt a schema of uniform disquotation:

(U) $\quad \forall a_1 \ldots a_n [T(\ulcorner\varphi(a_1 \ldots a_n)\urcorner) \equiv \varphi(a_1 \ldots a_n)]$

In the second case our axioms are formulas obtained from (U) by substituting concrete variables for $a_1 \ldots a_n$ and concrete formulas for a schematic letter φ. In fact (L) can be viewed as a special case of (U), with φ being a sentence and the sequence of quantified variables being empty.

Comment. Using (U) instead of (L) retains a lot of the disquotationalist spirit, although one could complain that it is more a satisfaction than a truth schema. The intended meaning of "$T(\ulcorner\varphi(a_1 \ldots a_n)\urcorner)$" is after all that a formula $\varphi(x_1 \ldots x_n)$ is satisfied by objects $a_1 \ldots a_n$. Admittedly, in an arithmetical context, where every object has a

[2] Admittedly, it is not the only possible option. Cf. Beall (2009), where disquotationalism is understood as a view that truth is a "fully transparent device" (p. 3); more exactly: it's "a device introduced via rules of intersubstitution: that $Tr(\ulcorner\alpha\urcorner)$ and α are intersubstitutable in all (nonopaque) contexts" (p. 1). On this approach, adopting T-biconditionals as axioms might be just one of the possible ways to give justice to the disquotationalist's intuitions.

standard numeral denoting it, we can express this thought employing just a one place truth predicate: we say in effect that the result of substituting in φ numerals for $a_1 \ldots a_n$ is true. However, in other contexts, where nameability assumption can't be employed, we would have to use a satisfaction predicate instead.

In what follows both types of disquotational axioms will be discussed, in two variants: typed and untyped one. I will concentrate on arithmetical context, taking PA as the base theory of syntax and stressing each time the arithmetical strength of the resulting theory.

15.2 Conservativeness

The main emphasis will be on (non)conservativity results. Conservativeness has been one of the major issues in recent debates about truth theories. Should we expect from a theory of truth that it be conservative over its base? The opinions have been divided. On the one hand, some philosophers of deflationary bent claimed that truth is an innocent and metaphysically thin notion. An explication of this claim has been proposed by Horsten in (1995) and elaborated by Shapiro (1998) and Ketland (1999). On this view, an adequate theory of truth for a given language should conservatively extend a base theory of syntax for this language. The motivation for accepting conservativeness demand is succinctly formulated in the following fragment of Shapiro's paper:

> How thin can the notion of arithmetic truth be, if by invoking it we can learn more about the natural numbers?

(see Shapiro 1998, p. 499.) As we see, the underlying intuition is that if by invoking the notion of truth we can learn more about natural numbers, then the notion of truth is not thin. In the next step the notion of conservativeness is used to analyze the situation in more detail. A representative fragment from Shapiro's paper runs as follows:

> I submit that in one form or another, conservativeness is essential to deflationism. Suppose, for example, that Karl correctly holds a theory B in a language that cannot express truth. He adds a truth predicate to the language and extends B to a theory B' using only axioms essential to truth. Assume that B' is not conservative over B. Then there is a sentence Φ in the original language (so that Φ does not contain the truth predicate) such that Φ is a consequence of B' but not a consequence of B. That is, it is logically possible for the axioms of B to be true and yet Φ false, but it is not logically possible for the axioms of B' to be true and Φ false. This undermines the central deflationist theme that truth is in-substantial. (Shapiro1998, p. 497)

Observe that although in the quoted passage the claim of insubstantiality of truth is explicated in terms of conservativeness, Shapiro remains noncommittal about a particular form of a conservativeness demand. (In fact various versions of the demand can be considered; see below, Definition 1.)

Others have argued that the deflationists have no reason to embrace conservativeness as a condition on truth theories; in addition, a theory of truth with this property

would be too weak.[3] I am not going to engage into this debate here; I stress only that results about arithmetical strength of truth theories are philosophically important no matter what one's standpoint in the debate is.[4]

Let us introduce now the definition of two notions of conservativeness.

Definition 1 Let T_1 and T_2 be theories in languages L_1 and L_2 (with $L_1 \subseteq L_2$). Then:

(a) T_2 is syntactically conservative over T_1 iff $T_1 \subseteq T_2$ and $\forall \psi \in L_1[T_2 \vdash \psi \rightarrow T_1 \vdash \psi]$.
(a) T_2 is semantically conservative over T_1 iff every model M of T_1 can be expanded to a model of T_2 (interpretations for new expressions of L_2 can be provided in M in such a way as to make T_2 true).

The two notions of conservativeness do not coincide. Semantic conservativeness is a more general notion: it gives via completeness theorem the syntactic version, but the opposite implication does not hold. Examples of truth theories being syntactically, but not semantically conservative over their base theories will be given below. Both notions are invoked by Shapiro in (1998). Later however most of the authors writing on the subject concentrated almost exclusively on the syntactic notion, ascribing to the deflationist a commitment to syntactic conservativeness. One of the few pleas for semantic conservativeness can be found in McGee (2006).

15.3 Typed Disquotation

I will discuss typed disquotational theories in two variants: local and uniform one.

15.3.1 Typed Uniform Disquotation

Adopting the typed approach, we obtain a Tarskian hierarchy of truth predicates and a family of theories characterizing the notion of truth for languages with truth predicates of all lower levels. Let L_0 be the language of Peano arithmetic; let L_{n+1} be the extension of L_n with a new one place predicate "T_n". Denote by $Ind(L_n)$ the set of all induction axioms for formulas of the language L_n. Then we define ("UTB" reads "uniform Tarski biconditionals"):

[3] See e.g. Halbach (2001), p. 188: "As far as I can see, neither have deflationists subscribed to conservativeness explicitly nor does it follow from one of their other doctrines. (...) But if the deflationist understands his claim that truth is not a substantial notion as implying that his truth theory has no substantial consequences, he commits a mistake."

[4] For a philosophical discussion of conservativeness as a demand for deflationary truth theories, see also (Cieśliński 2010), (Ketland 2010) and (Tennant 2010).

Definition 2

- $UTB_0 = PA$
- $UTB_{n+1} = UTB_n \cup \{\forall a_1 ... a_n [T_n(\ulcorner \varphi(a_1 ... a_n) \urcorner) \equiv \varphi(a_1 ... a_n)] : \varphi \in L_n\} \cup Ind(L_{n+1})$

(Observe that UTB_n is always in the language L_n.) Then the following result can be obtained:

Theorem 3 *For every n, UTB_{n+1} is syntactically conservative over UTB_n.*[5]

Proof Assume that $UTB_{n+1} \vdash \varphi$, $\varphi \in L_n$. Consider all disquotational axioms employing T_n in a (fixed) proof of φ. Let $\psi_0 ... \psi_i$ be all formulas mentioned in these axioms in the scope of T_n (i.e. every such an axiom has a form "$\forall a_1 ... a_n [T_n(\ulcorner \psi_k(a_1 ... a_n) \urcorner) \equiv \psi_k(a_1 ... a_n)]$" for some $k \leq i$). Taking m as a maximal quantifier rank of $\psi_0 ... \psi_i$, we observe that that there is a predicate "$Tr_m(x)$" of the language L_n, which is a truth predicate for formulas of L_n with a quantifier rank smaller or equal m.[6] Since UTB_n proves all biconditionals of the form "$\forall a_1 ... a_n [Tr_m(\ulcorner \psi_k(a_1 ... a_n) \urcorner) \equiv \psi_k(a_1 ... a_n)]$" for $k \leq i$, the proof of φ can be reconstructed in UTB_n by substituting "Tr_m" for "T_n" and by supplying proofs for the resulting biconditionals when necessary. □

Before analyzing the semantic conservativeness property, I want to remind the reader an important notion of a recursively saturated model.

Definition 4

- Let Z be a set of formulas with one free variable x and parameters $a_1 ... a_n$ from a model M. Z is realized in M iff there is an $a \in M$ such that every formula in Z is satisfied in M under a valuation assigning a to x.
- Z is a type of M iff every finite subset of Z is realized in M.
- M is recursively saturated iff every recursive type of M is realized in M.

It is a well known fact that every infinite model is elementarily equivalent with a recursively saturated structure (see e.g. Kaye (1991), Proposition 11.4, p. 14).

Theorem 5 *UTB_{n+1} is not semantically conservative over UTB_n.*[7]

Proof The proof consists in observing that only recursively saturated models of UTB_n can be expanded to models of UTB_{n+1}. Given a model M_1 of UTB_n, assume that it's possible to expand it to a model $M_2 = (M, T_0 ... T_n)$ in such a way that $M_2 \models UTB_{n+1}$. Let $p(x, a_1 ... a_n)$ be a recursive type over M_1. Let "$s \in p$" be an arithmetical formula representing in PA the recursive set of formulas (without parameters) used in forming the type $p(x, a_1 ... a_n)$. Then we have:

$$\forall k \in N M_2 \models \exists z \forall \varphi(x, y_1 ... y_n) < k [\varphi(x, y_1 ... y_n) \in p \rightarrow T_n(\varphi(z, a_1 ... a_n))]$$

[5] Cf. Halbach (2011), p. 55, where the proof is given that UTB_1 is conservative over PA.
[6] On partial truth predicates, see Kaye (1991), p. 119 ff.
[7] Cf. (Kaye 1991), p. 228, Proposition 15.4.

So by overspill, there is a nonstandard $b \in M_2$ such that:

$$M_2 \models \exists z \forall \varphi(x, y_1 ... y_n) < b[\varphi(x, y_1 ... y_n) \in p \rightarrow T_n(\varphi(z, a_1 ... a_n))]$$

Then such a z realizes our type $p(x, a_1 ... a_n)$ in M_1.[8] □

15.3.2 Typed Local Disquotation

In an analogous manner, we define now the hierarchy of typed theories based on the local disquotational schema.

Definition 6

- $TB_0 = PA$
- $TB_{n+1} = TB_n \cup \{T_n(\ulcorner \varphi \urcorner) \equiv \varphi : \varphi \in L_n\} \cup Ind(L_{n+1})$

Since local disquotation is a special variant of uniform disquotation, some results from the last subsection carry over to the present case. In particular, Theorem 3 applies without any changes—TB_n is syntactically conservative over TB_k for $k < n$. As for Theorem 5, although its proof doesn't carry over to our present case, the result still holds.

Theorem 7 *TB_{n+1} is not semantically conservative over TB_n.*[9]

Before giving the proof, I would like to remind the reader some basic concepts, which will be used later on.

Explanation 1 (the notion of coding). A set Z of natural numbers is coded in a model M by an element a of this model iff $Z = \{n : M \models n \in a\}$. Expression of the form "$x \in y$" is taken to be an arithmetical formula used for the purposes of coding; it can be e.g. "$p_x \mid y$" ("the x^{th} prime divides y"). In the standard model it is exactly finite sets which are coded. The situation is different in nonstandard models, where some infinite sets will be coded as well.[10]

Explanation 2 (the notion of a prime model). Let S be a consistent extension of PA in the language L with new predicates $\tilde{A}_1 ... \tilde{A}_n$, with full induction for L. Let $M^* = (M, A_1 ... A_n)$ be a model for S. A *prime model* K of S is obtained from M^* in the following manner:

- The universe of K is defined as $\{a \in M$: a is definable in M^* by some formula of $L\}$

[8] Although we worked in M_2, the transition to M_1 is made possible by the fact that all formulas in our type belong to the language L_n, i.e. they do not contain "T_n", so if they are satisfied in M_2, they are also satisfied in M_1.

[9] After obtaining Theorem 7, I found out that the result was proved earlier by Fredrik Engström. Engström's work is unpublished.

[10] For more about coded sets, see e.g. Kaye (1991) p. 141 ff.

- The operations of K are the operations of M^* restricted to the universe of K
- for $i \leq n$, $A_i^K = A_i \cap K$.

It is possible to show that K is an elementary submodel of M^*, with all elements of K being definable in K.[11]
We now start with the following lemma:

Lemma 8 *For every $n \in N$, the following conditions are equivalent for an arbitrary nonstandard model $M^* = (M, T_0 ... T_{n-1})$ of TB_n:*

(a) M^* can be expanded to a model of TB_{n+1}
(b) M codes $Th(M^*)$.

Proof For the direction from (b) to (a), assume that a is a code of $Th(M^*)$ in M. Define: $T_n = \{x \in M : M \models \text{“}x \in a\text{”}\}$. Then $(M, T_0 ... T_n) \models TB_{n+1}$ as required (observe in particular, that T_n is inductive, since it's definable with parameters in M). For the opposite direction, assume that M^{**} is an expansion of M^* satisfying TB_{n+1}. Then we have:

$$\forall k \in N M^{**} \models \exists z \forall s[s \in z \equiv (s < k \wedge T_n(s))]$$

Therefore by overspill there is a nonstandard $a \in M$ such that:

$$M^{**} \models \exists z \forall s[s \in z \equiv s < a \wedge T_n(s)]$$

(Observe that overspill can be used, because we assumed that T_n is inductive.) Picking such a z, we obtain a code for $Th(M^*)$ in M, as required. □

With Lemma 8 at hand, the proof of Theorem 7 is immediate.

Proof of theorem 7. Let $M^* = (M, T_0 ... T_{n-1})$ be a prime nonstandard model of TB_n. We show that it can't be expanded to a model of TB_{n+1}. For an indirect proof, assume that it can. Then by Lemma 8, M codes $Th(M^*)$, and since it's prime, a code c of $Th(M^*)$ is definable in M^*. Take a formula $\alpha(x)$ defined as:

$$\alpha(x) := \exists z[\psi(z) \wedge x \in z]$$

with $\psi(x)$ being a formula of L_n which defines c in M^*. It's easy to observe that $\alpha(x)$ is a truth predicate of the language L_n for L_n sentences in M^*, which contradicts Tarski's indefinability theorem. □

15.4 Untyped Disquotation

If we decide to drop the typing restrictions, the situation may change drastically, depending on our choice of the substitution class for the T-schemata. Even a seemingly weaker schema (L) can generate quite powerful theories once a suitable set

[11] More information about prime models can be found in Kaye (1991), p. 91 ff.

of instances is selected. The key observation was made by Vann McGee in (1992). Consider an arbitrary sentence φ of the arithmetical language extended with the truth predicate (it will be denoted as L_T). Let PAT be a theory in the language L_T whose all extralogical axioms are just those of PA. Then there is a substitution of (L) which is provably (in PAT) equivalent with φ. The method of finding an appropriate substitution of (L) is effective; it is also possible to "decode" effectively the sentence φ given a corresponding substitution of (L). In effect we obtain the following:

Theorem 9 *Let H be an arbitrary recursive set of sentences of the language L_T. Then there is a recursive set G of substitutions of (L) such that H and G are (over PAT) recursive axiomatizations of the same theory.*

Superficially, Theorem 9 might look like a great news for the disquotationalist. Whatever your favourite theory of truth is, you can always axiomatize it by substitutions of (L). Nothing else is needed! However, in fact McGee's result leads the disquotationalist nowhere. The main problem is that the disquotationalist wants to treat the substitutions of the T-schemata as epistemologically basic. Whatever more substantial principles of truth we accept, he plans to justify them by recourse to the T-schemata, and not the other way round. (In particular, it won't do to justify the acceptance of a given set of substitutions by saying that they are equivalent to the axioms of our favourite (substantial) theory of truth.) Unfortunately, McGee's result shows, that in general there is nothing basic about the schema (L). *False* sentences are provably equivalent to substitutions of (L); arithmetical truths unknown to us are also provably equivalent to such substitutions. It seems that (in many cases) accepting a given substitution of (L) requires a special argument, which goes beyond a mere saying that it is a substitution of a disquotational schema.

In short: the disquotationalist needs to characterize a set S satisfying the following conditions: (1) S is a recursive set of substitutions of a T-schema (the local or the uniform one) (2) we have good reasons to treat elements of S as epistemologically basic. In particular, we do not accept S because of its equivalence (over PAT) with some substantial truth theory of our choice.

The disquotationalist's predicament is that it seems quite difficult to find a comprehensive set S satisfying (1) and (2).[12] The difficulty will be illustrated below, by considering a concrete candidate for the role of such an S: a set of *positive* substitutions of a T-schema.

15.4.1 Untyped Uniform Disquotation

The proposal is described by Volker Halbach in (2009). It arises from an analysis of the way paradoxes are produced. The initial insight is that in paradoxical reasonings

[12] One path could consist in considering maximal conservative sets of substitutions of a T-schema. It has been shown however, that there are uncountably many such sets and none of them is axiomatizable. See Cieśliński (2007).

we apply the truth predicate to sentences containing a negated occurrence of this predicate (see (Halbach 2009), p. 788). This is plainly the case with the liar sentence: the standard, diagonal construction of the liar produces a sentence with "T" within a scope of one negation. In effect one could try to avoid the paradoxes by restricting the set of substitutions of (U): from now on we admit *positive* substitutions only, i.e. our axiom is whatever can be obtained from (U) by substituting a positive formula for a schematic letter φ.

The notion of a positive formula is defined for a language containing $\neg$, $\wedge$ and $\vee$ as the only connectives. (Implication is not a primitive symbol. A reflection on Curry's paradox forces us to treat apparently positive occurrences of "T" in an antecedent of an implication as negative.[13]) From now on we stipulate that L_T (the extension of the language of PA with the truth predicate) contains just those connectives. Then we define:

Definition 10

(a) A formula φ of L_T is positive iff every occurrence of "T" in φ appears in the scope of an even number of negations.
(b) PUTB ("positive uniform Tarski biconditionals") is a theory axiomatized by all axioms of PAT with extended induction and all substitutions of the uniform truth schema (U) by positive formulas.

Halbach's main theorem characterizes the arithmetical strength of PUTB. Far from being conservative over PA, PUTB is arithmetically very strong—it is in fact arithmetically equivalent to the Kripke-Feferman theory KF, one of the strongest truth theories discussed in contemporary literature.[14]

Theorem 11 $\forall \psi \in L_{PA}[PUTB \vdash \psi \equiv KF \vdash \psi]$

The proof consists in showing that PUTB defines the truth predicate of KF, i.e. there is a formula $\alpha(x)$ such that PUTB proves all sentences obtained from axioms of KF by replacing the truth predicate $T(x)$ with $\alpha(x)$. Together with the information that $PUTB \subseteq KF$, this implies Theorem 11. For details, see (Halbach 2009) (lemma 4.3 and theorem 5.1). It's also worth mentioning, that nevertheless PUTB is truth-theoretically weaker than KF—it doesn't prove compositional truth axioms (Halbach (2009), lemma 6.1 and below).

Halbach ended his paper with an open question about the arithmetical strength of the theory taking as axioms all positive substitutions of the *local* truth schema (L). I will sketch the answer in the next subsection.

---

[13] In Curry's paradox we consider a sentence ψ satisfying the condition: $\psi \equiv [T(\ulcorner \psi \urcorner) \rightarrow 0 = 1]$. It turns out then that adopting a T-biconditional for ψ results in a contradiction. However, a Curry sentence ψ constructed by diagonalization contains an occurrence of the truth predicate which is not negated.

[14] A presentation of KF can be found in Halbach (2011), starting on p. 195.

15.4.2 Untyped Local Disquotation

Let's consider now a case of a positive local disquotation. The basic definition is as follows.

Definition 12 PTB ("positive Tarski biconditionals") is a theory axiomatized by all axioms of PAT with extended induction and all substitutions of the local truth schema (L) by positive sentences.

We formulate now the main theorem about PTB.

Theorem 13 *PTB is syntactically conservative over Peano Arithmetic.*[15]

The proof consists in showing that:

(*) For every finite set S of axioms of PTB, for every recursively saturated model M of Peano arithmetic, M can be expanded to a model of S (i.e. an interpretation of the truth predicate can be found in M in such a way as to make all sentences in S true).

After (*) is obtained, Theorem 13 follows immediately.

Proof of Theorem 13 from (*). Assume that $PTB \vdash \varphi$ for an arithmetical sentence φ. Then there is a finite set S of axioms of PTB such that $S \vdash \varphi$. By (*), every recursively saturated model of PA can be expanded to a model of S; therefore φ is true in every recursively saturated model of PA. But every model of PA is elementarily equivalent with a recursively saturated model, therefore φ is true in every model of PA, which by completeness implies that $PA \vdash \varphi$. □

Sketch of the proof of (*). A handy tool in this proof is a notion of a translation function $t(a, \psi)$, which takes as arguments a number a (possibly nonstandard) from a given model and a formula ψ belonging to the language with the truth predicate. The value of this function is an arithmetical formula (no "T" inside) with a parameter a—a $translation$ of ψ. The translation is obtained by substituting all occurrences of "$T(t)$" in ψ by "$t \in a$"—an arithmetical formula used for the purposes of coding sets (see Explanation 1). In effect the translation interprets the truth predicate in ψ as referring to the set coded by a.

With the notion of a translation at hand, we can define, for a recursively saturated model M, a family of recursive types over M, a family of elements realizing these types and of models expanding M with an interpretation of the truth predicate. In what follows the predicates $Sent_{PA}$ and $Sent_T^+$ denote (respectively) the set of all sentences of the language of PA and the set of all positive sentences of the language L_T.

[15] The proof of Theorem 13 was presented on the "Truth be told" conference in Amsterdam (2011), and also in (Cieśliński 2011).

Definition 14

1. • $p_0(x) = \{\varphi \in x \equiv \varphi : \varphi \in Sent_{PA}\} \cup \{\forall w(w \in x \Rightarrow w \in Sent_{PA}\}$
   • d_0 realizes $p_0(x)$
   • $T_0 = \{a : M \models a \in d_0\}$
   • $M_0 = (M, T_0)$
2. • $p_{n+1}(x, d_n) = \{\varphi \in x \equiv t(d_n, \varphi) : \varphi \in Sent_T^+\} \cup \{\forall z(z \in d_n \Rightarrow z \in x)\} \cup \{\forall z(z \in x \Rightarrow z \in Sent_T^+)\}$
   • d_{n+1} *realizes* $p_{n+1}(x, d_n)$
   • $T_{n+1} = \{a : M \models a \in d_{n+1}\}$
   • $M_{n+1} = (M, T_{n+1})$

The idea behind Definition 14 is as follows. A number d_0 obtained at the start codes the set of all arithmetical sentences true in M—we denote it as T_0. Building a model M_0, we interpret the truth predicate with just this set.[16] In the next stage we obtain a number d_1 coding the set of all positive sentences of the language L_T, which become true in M once "T" is interpreted by a set T_0 (i.e. once "$T(t)$" is replaced by "$t \in d_0$"). And then we iterate the construction for all natural numbers.

At this moment two observations become useful. The first is that $T_0 \subseteq T_1 \subseteq T_2 \ldots$. The second is a general fact about positive formulas: if A and B are subsets of the universe of a model M with $A \subseteq B$, then every positive formula satisfied under some valuation in (M, A) will be also satisfied in (M, B). From these two observations it follows that given a finite set $S = \{T(\ulcorner\varphi_0\urcorner) \equiv \varphi_0 \ldots T(\ulcorner\varphi_k\urcorner) \equiv \varphi_k\}$ of axioms of PTB, there will be a natural number n such that $M_{n+1} \models S$. We simply find an n such that:

$$\forall i \leq k[M_n \models \varphi_i \vee \neg\exists l \in N\, M_l \models \varphi_i]$$

and then observe that all the equivalences from S are true in M_{n+1}. Since T_{n+1} is definable with parameters in M (by a formula "$x \in d_{n+1}$), M_{n+1} satisfies also all the induction axioms for the extended language. □

Although PTB is syntactically conservative over PA, it doesn't have the semantic conservativeness property. This follows easily from Theorem 7.

Corollary 15 *PTB is not semantically conservative over PA.*

Proof Otherwise, since TB_1 can be treated as a subtheory of PTB, every model of PA could be expanded to a model of TB_1, contrary to Theorem 7. □

[16] Strictly speaking, T_0 will contain not only (codes of) arithmetical sentences true in M, but also these nonstandard numbers a, for which the formula "$a \in d_0$" is satisfied in M.

15.5 Justification of Disquotational Axioms

Is disquotationalism a philosophically attractive position? An answer to this question depends on the one hand, on the assessment of the strength of disquotational theories, and on the second, on the justification of disquotational axioms. In these final comments I want to concentrate on the second issue. As we saw, the key move consists in choosing a substitution class S for a T-schema (local or uniform one). In view of Theorem 9, the choice of S must be well motivated—it won't do in general to accept S as a set of "mere substitutions of a T-schema" (see remarks below Theorem 9). What motivations can be offered?

Restoration of the consistency of disquotational theory is a natural aim. Naive, unrestricted T-schema generates a contradiction—that's a fact to which all truth theorists must react and the disquotationalist is no exception. Restoring the consistency of a theory of truth should be treated as a permissible motivation for the disquotationalist to proceed. The question is only how far it can take us.

I will discuss two worries concerning this type of justification. Eventually I will dismiss the first one; in contrast, the second seems to me a real issue.

Objection 1 (cf. Halbach 2011, p. 311). The disquotationalist should not only guarantee that his theory is safe from a contradiction, but he must do it by using safe proof methods. In particular, since his aim is to characterize a satisfactory notion of truth, he shouldn't be allowed to use model theoretic arguments. A model theoretic notion of truth goes beyond the disquotational notion, therefore he can't employ it. Perhaps he will achieve his aim for typed theories: using the means available in PA, he can prove e.g. relative consistency of UTB_1 (i.e. he can prove in PA that if PA is consistent, then UTB_1 is consistent), so he is entitled to claim that his trust in the consistency of UTB_1 is no less warranted than his trust in PA itself. But the problem is that this approach fails as a general strategy. We can't do the same for strong disquotational theories, whose relative consistency is not provable in PA.

Answer. I can see no reason why model theoretic means shouldn't be available to the disquotationalist, discussing the notion of arithmetical truth. Observe that the notion of truth in a model can be expressed by set theoretical means; the completeness theorem can also be viewed as a set theoretical result. Questioning set theory is not a part of the disquotationalist's baggage; he questions rather a substantial notion of truth. As I take it, he is free to use "truth in a model" as a technical notion, useful for obtaining consistency results. He can just add: "this is not the same as the concept of (unrelativized) truth-as-we-understand-it, which I try to characterize by means of a T-schema. These are just two different concepts".

Although I take the above answer as basically correct, one qualification is needed. The real issue is not whether the disquotationalist can use set theory with its notion of model theoretic truth (he can!); the issue is rather *how* he uses it. In particular, I would find his employment of the notion of truth in the standard (or the intended) model quite problematic. If a given philosophical argument hinged on identifying truth-as-we-understand-it with truth in the intended model, it would seem indeed

that a stronger (non-disquotational) notion of truth is needed to make the argument work. As we will see, this provides a basis for the second and more serious objection.

Objection 2. It is not enough that a disquotational theory be consistent. If a T-schema is to be treated as epistemologically basic, some argument is needed to show that the obtained theory is arithmetically sound. By Theorem 9, false arithmetical sentences are also provably equivalent to substitutions of (L). Why should we trust that the disquotational theory of our choice doesn't produce false results?

Comment. An advocate of typed disquotation—say, of UTB_1—could retort that PA proves not only relative consistency, but also conservativeness of UTB_1 over PA. In effect we (as users of PA) are entitled to trust UTB_1 just as we are entitled to trust PA. No strong notion of truth is needed to establish that.[17]

However, the situation of an advocate of an untyped theory like PUTB is more problematic. A natural move could consist in arguing that PUTB admits a standard interpretation—that it's possible to interpret the truth predicate of PUTB in the intended model of arithmetic. But the problem with this rejoinder is that in its employment of the notion of truth in the standard model, it goes beyond the legitimate disquotational means. The disquotationalist can't argue "my axioms are trustworthy, since they produce true arithmetical results, which I know because they are true under the intended interpretation". In saying this he commits himself to a stronger notion of arithmetical truth than the disquotational one. And that's his predicament.

15.6 Problems

I end with listing what I take to be the main problems in this area of research. The problems are:

(1) Is there a natural substitution class for (U), which could be used to obtain not only the arithmetical content of KF, but also its compositional principles?
(2) Are there any plausible candidates for the role of a natural substitution class for (L), producing an arithmetically strong theory?
(3) Is there a disquotationally acceptable answer to the question "why should we trust positive disquotational axioms"?
(4) Is there an argument for conservativity of PTB over PA, which can be formalized in PA?

Questions (1)–(3) are philosophical; question (4) is formal. (1) relates to the fact that PUTB, although arithmetically strong, is quite weak in proving compositional principles (see Halbach 2009, pp. 793 ff.). Admittedly, it is not very clear what classes should count as "natural". The intuition is that principled, non ad-hoc reasons should stand behind selecting such a class. Question (2) is motivated by our observation, that

[17] In this respect the situation of an adherent of a typed disquotational theory is quite comfortable; his problems lie elsewhere: in the deductive weakness of his theory.

PTB is conservative over PA, so positive substitutions of (L) do not take us very far (are there other good candidates worth considering?) Question (3) is in effect whether a good answer to Objection 2 can be given. For question (4), observe that the proof of conservativity of PTB over PA given here is semantic and doesn't translate easily to a syntactic argument.

Acknowledgements This work was supported by a grant number N N101 170438 by the Polish Ministry of Science and Education (MNiSW).

References

Beall, Jc. (2009). *Spandrels of truth*. Oxford: Oxford University Press.
Cieśliński, C. (2007). Deflationism, conservativeness and maximality. *Journal of Philosophical Logic, 36,* 695–705.
Cieśliński, C. (2010). Truth, conservativeness, and provability. *Mind, 119,* 409–422.
Cieśliński, C. (2011). T-equivalences for positive sentences. *The Review of Symbolic Logic, 4,* 319–325.
Field, H. (1994). Deflationist views of meaning and content. *Mind, 103,* 249–284.
Halbach, V. (2001). How innocent is deflationism? *Synthese, 126,* 167–194.
Halbach, V. (2009). Reducing compositional to disquotational truth. *The Review of Symbolic Logic, 2,* 786–798.
Halbach, V. (2011). *Axiomatic theories of truth*. Cambridge: Cambridge University Press.
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 of studies in the general philosophy of science* (Vol. 3, pp. 173–187). Tilburg: Tilburg University Press.
Horwich, P. (1990). *Truth*. Oxford: Basil Blackwell.
Kaye, R. (1991). *Models of Peano arithmetic*. Oxford: Clarendon.
Ketland, J. (1999). Deflationism and Tarski's paradise. *Mind, 108,* 64–94.
Ketland, J. (2010). Truth, conservativeness, and provability: Reply to Cieslinski. *Mind, 119,* 423–436.
McGee, V. (1992). Maximal consistent sets of instances of Tarski's schema (T). *Journal of Philosophical Logic, 21,* 235–241.
McGee, V. (2006). In praise of the free lunch: Why disquotationalists should embrace compositional semantics. In V. F. Hendricks, S. A. Pedersen, & T. Bollander (Eds.), *Self reference* (pp. 95–120) Stanford: CSLI.
Shapiro, S. (1998). Proof and truth: Through thick and thin. *Journal of Philosophy, 95,* 493–521.
Tennant, N. (2010). Deflationism and the Gödel-Phenomena: Reply to Cieslinski. *Mind, 119,* 437–450.

Chapter 16
New Constructions of Satisfaction Classes

Ali Enayat and Albert Visser

Abstract We use model-theoretic ideas to present a perspicuous and versatile method of constructing full satisfaction classes on models of Peano arithmetic. We also comment on the ramifications of our work on issues related to conservativity and interpretability.

16.1 Introduction

In our forthcoming paper (Enayat and Visser 2012) we explore satisfaction classes over a wide variety of 'base theories' ranging from weak fragments of arithmetic to systems of ZF set theory and beyond. This note provides a synopsis of some of this work in the context of the most popular base theory adopted in investigations of axiomatic theories of truth, namely PA (Peano Arithmetic).

The notion of a satisfaction class was first introduced and investigated by Krajewski in his 1976 paper (Krajewski 1976). Two noteworthy accomplishments of (Krajewski 1976) are the following results:

(1) If a countable model of a 'base theory' (such as PA) carries at least one full satisfaction class, then it has a countable elementary extension that carries continuum-many full satisfaction classes.
(2) Every model of ZF has an elementary extension that carries a full satisfaction class.

This research was partially supported by a grant from the Descartes Center of Utrecht University, which supported the first author's visit to Utrecht to work closely with the second author.

A. Enayat
Department of Philosophy, Linguistics, and Theory of Science,
University of Gothenburg, Box 200, 405 30, Gothenburg, Sweden
e-mail: ali.enayat@gu.se

A. Visser
Department of Philosophy, Bestuursgebouw, Heidelberglaan 6,
584 CS Utrecht, Utrecht, The Netherlands
e-mail: albert.visser@phil.uu.nl

T. Achourioti et al. (Eds.), *Unifying the Philosophy of Truth,* Logic, Epistemology,
and the Unity of Science 36, DOI 10.1007/978-94-017-9673-6_16

The question whether the analogue of (2) holds for PA remained open until the appearance of the joint work (Kotlarski et al. 1981) of Kotlarski, Krajewski, and Lachlan in 1981, in which the rather exotic proof-theoretic technology of '$\mathcal{M}$-logic' (an infinitary logical system based on a nonstandard model $\mathcal{M}$), was invented to construct 'truth classes' over countable recursively saturated models of PA.[1] This model-theoretic result can be used to show that the analogue of (2) does indeed hold for PA, which in turn can be used to show that $\mathsf{PA}^{\mathsf{FT}}$ is conservative over PA, where $\mathsf{PA}^{\mathsf{FT}} = \mathsf{PA} + $ "T is a full truth class". The conservativity of $\mathsf{PA}^{\mathsf{FT}}$ over PA has attracted considerable philosophical attention, especially in relation to the grand debate concerning deflationism.[2]

In this paper we present a perspicuous method for the construction of full satisfaction classes that is dominantly based on *model-theoretic techniques* (e.g., expanding the language, compactness, and elementary chains). As we shall see, our construction method is quite versatile and can be used to construct many (if not all) of the results that have hitherto been only possible to establish with the use of $\mathcal{M}$-logic machinery. Furthermore, the method can also be employed to build new types of full satisfaction classes (see Sect. 16.6).

We present the necessary preliminaries in Sect. 16.2, and then in Sect. 16.3 we concentrate on the basic form of our new construction of full satisfaction classes, where it is used to show that every model of PA has an elementary extension that carries a full satisfaction class. The versatility of the methodology of Sect. 16.3 is illustrated in Sect. 16.4, in which an appropriate modification of the method is used to construct truth classes for models of PA. As explained in Sect. 16.5, certain arithmetizations of our construction can also be employed to establish that (1) $\mathsf{PA}^{\mathsf{FT}}$ is interpretable in PA; and (2) the conservativity of $\mathsf{PA}^{\mathsf{FT}}$ over PA can be verified in PRA (Primitive Recursive Arithmetic). Finally, in Sect. 16.6 we briefly describe further applications of the methods introduced in this paper.

Acknowledgments We are grateful to the editors of this volume for their interest in our work. Thanks also to Volker Halbach, Fredrik Engström, and James Schmerl for helpful feedback on preliminary drafts of this paper. We are particularly indebted to Schmerl for catching an inaccuracy in an earlier formulation of Lemma 3.1, and for his suggestion to distill the results of our paper (Enayat and Visser 2012) in this form for wider dissemination.

16.2 Preliminaries

Definition 2.1 Throughout this paper PA refers to Peano arithmetic formulated in a *relational language* $\mathcal{L}_{\mathsf{PA}}$ using the logical constants $\{\neg, \vee, \exists, =\}$. Note that in this

[1] As explained in Sect. 16.4, a truth class is essentially a well-behaved kind of satisfaction class. The $\mathcal{M}$-logic methodology was further elaborated to establish refined constructions of full truth classes by Smith (1984, 1987, 1989), Kaye (1991), and Engström (2002).

[2] A recent noteworthy paper in this connection is McGee's (2003).

formulation PA has no constant symbols; the arithmetical operations of addition and multiplication are construed as ternary relations; and conjunction, universal quantification, and other logical constants are taken as defined notions in the usual way.

It is well known that PA has more than sufficient expressive machinery to handle syntactic notions. The following list of $\mathcal{L}_{\mathsf{PA}}$-formulae will be useful here.[3]

- $\mathsf{Form}(x)$ is the formula expressing "x is the code of an $\mathcal{L}_{\mathsf{PA}}$-formula using variables $\{v_i : i \in \mathbb{N}\}$, and the non-logical symbols available in $\mathcal{L}_{\mathsf{PA}}$".
- $\mathsf{Asn}(x)$ is the formula expressing "x is the code of an assignment", where an assignment here simply refers to a function whose domain consists of a (finite) set of variables. We use α and its variants (α', α_0, etc.) to range over assignments.
- $y \in \mathsf{FV}(x)$ is the formula expressing "$\mathsf{Form}(x)$ and y is a free variable of x".
- $y \in \mathsf{Dom}(\alpha)$ is the formula expressing "the domain of α includes y".
- $\mathsf{Asn}(\alpha, x)$ is the following formula expressing "α is an assignment for x":

$$(\mathsf{Form}(x) \wedge \mathsf{Asn}(\alpha) \wedge \forall y(y \in \mathsf{Dom}(\alpha) \leftrightarrow y \in \mathsf{FV}(x)).$$

- $x \lhd y$ is the formula expressing "x is the code of an immediate subformula of the $\mathcal{L}_{\mathsf{PA}}$-formula coded by y", i.e., $x \lhd y$ abbreviates the conjunction of $\mathsf{Form}(y)$ and the following disjunction:

$$(y = \neg x) \vee \exists z\,((y = x \vee z) \vee (y = z \vee x)) \vee \exists i\,(y = \exists v_i\ x).^4$$

The theory $\mathsf{PA}^{\mathsf{FS}}$ (read as "PA with full satisfaction") is formulated in an *expansion* of the language $\mathcal{L}_{\mathsf{PA}}$ by adding a new *binary* predicate $\mathsf{S}(x, y)$. The binary/unary distinction is of course not an essential one since PA has access to a definable pairing function. However, the binary/unary distinction *at the conceptual level* marks the key difference between satisfaction classes and truth classes (the latter are discussed in Sect. 16.4). $\mathsf{PA}^{\mathsf{FS}}$ is defined below with the help of a collection of sentences $\mathsf{Tarski}(\mathsf{S},\mathsf{F})$.

When reading the definition below it is helpful to bear in mind that $\mathsf{Tarski}(\mathsf{S},\mathsf{F})$ expresses:

---

[3] All of the formulae in the list can be arranged to be Σ_0-formulae in the sense of Definition 2.4.

[4] Technically speaking, this formula should be written so as to distinguish the logical operations of the meta-langauge with those of the object-language. For example, using Feferman' s commonly used 'dot-convention', one would write:

$$\big(y = \dot{\neg} x\big) \vee \exists z\big(\big(y = x \mathbin{\dot{\vee}} z\big) \vee \big(y = z \mathbin{\dot{\vee}} x\big)\big) \vee \exists i\big(y = \dot{\exists} v_i\ x\big).$$

However, since the difference between the two kinds of operation will be always clear from the context, we have opted for the lighter notation.

F is a subset of Form that is closed under immediate subformulae; each member of S is an ordered pair of the form (x, α), where $x \in \mathsf{F}$ and α is an assignment for x; and S satisfies Tarski's compositional clauses.

Definition 2.2 $\mathsf{PA}^{\mathsf{FS}} := \mathsf{PA} \cup \mathsf{Tarski}(\mathsf{S}, \mathsf{Form})$, where $\mathsf{Tarski}(\mathsf{S},\mathsf{F})$ is the conjunction of the universal generalizations of the formulae $\mathsf{tarski}_0(\mathsf{S}, \mathsf{F})$ through $\mathsf{tarski}_4(\mathsf{S}, \mathsf{F})$ described below, all of which are formulated in $\mathcal{L}_{\mathsf{PA}} \cup \{\mathsf{F}(\cdot), \mathsf{S}(\cdot, \cdot)\}$, where S and F do not appear in $\mathcal{L}_{\mathsf{PA}}$.

In the following formulae R ranges over the *relations in* $\mathcal{L}_{\mathsf{PA}}$; $t, t_0, t_1, \cdots$ are *metavariables*, e.g, we write $R\,(t_0, \cdots, t_{n-1})$ instead of $R\left(v_{i_0}, \cdots, v_{i_{n-1}}\right)$; and $\alpha' \supseteq \alpha$ abbreviates

$$\big(\mathsf{Dom}(\alpha') \supseteq \mathsf{Dom}(\alpha)\big) \wedge \forall t \in \mathsf{Dom}(\alpha)\, \alpha(t) = \alpha'(t).$$

- $\mathsf{tarski}_0(\mathsf{S}, \mathsf{F}) := (\mathsf{F}(x) \to \mathsf{Form}(x)) \wedge (\mathsf{S}(x, \alpha) \to (\mathsf{F}(x) \wedge \mathsf{Asn}(\alpha, x))) \wedge$

$$(y \lhd x \wedge \mathsf{F}(x) \to \mathsf{F}(y)).$$

- $\mathsf{tarski}_{1,R}(\mathsf{S},\mathsf{F}) :=$

$$\left(\mathsf{F}(x) \wedge (x = \ulcorner R\,(t_0, \cdots, t_{n-1})\urcorner) \wedge \mathsf{Asn}(\alpha, x) \wedge \bigwedge_{i<n} \alpha(t_i) = a_i\right) \to$$

$$(\mathsf{S}(x, \alpha) \leftrightarrow R\,(a_0, \cdots, a_{n-1})).$$

- $\mathsf{tarski}_2(\mathsf{S}, \mathsf{F}) := (\mathsf{F}(x) \wedge (x = \neg y) \wedge \mathsf{Asn}(\alpha, x)) \to$

$$(\mathsf{S}(x, \alpha) \leftrightarrow \neg \mathsf{S}(y, \alpha)).$$

- $\mathsf{tarski}_3(\mathsf{S}, \mathsf{F}) := (\mathsf{F}(x) \wedge (x = y_1 \vee y_2) \wedge \mathsf{Asn}(\alpha, x)) \to$

$$(\mathsf{S}(x, \alpha) \leftrightarrow (\mathsf{S}\,(y_1, \alpha \restriction \mathsf{FV}(y_1)) \vee \mathsf{S}\,(y_2, \alpha \restriction \mathsf{FV}(y_2)))).$$

- $\mathsf{tarski}_4(\mathsf{S}, \mathsf{F}) := (\mathsf{F}(x) \wedge (x = \exists t\ y) \wedge \mathsf{Asn}(\alpha, x)) \to$

$$\big(\mathsf{S}(x, \alpha) \leftrightarrow \exists \alpha' \supseteq \alpha\ \mathsf{S}(y, \alpha')\big).$$

Definition 2.3 Suppose $\mathcal{M} \models \mathcal{L}_{\mathsf{PA}}$, $F \subseteq M$, and S is a binary relation on M.[5]

(a) S is an *F-satisfaction class* if $(\mathcal{M}, S, F) \models \mathsf{Tarski}(\mathsf{S},\mathsf{F})$.[6]

(b) Let $\omega_{\mathcal{M}}$ be the well-founded initial segment of $\mathcal{M}$ that is isomorphic to the ordinal ω. We say that F is the set of *standard* $\mathcal{L}_{\mathsf{PA}}$-formulae of $\mathcal{M}$ if

$$F = \mathsf{Form}^{\mathcal{M}} \cap\ \omega_{\mathcal{M}}.$$

[5] Throughout the paper we use the convention of using M, M_0, N, etc. to denote the universes of discourse of structures $\mathcal{M}$, $\mathcal{M}_0$, $\mathcal{N}$, etc.

[6] Note that the closure of F under direct subformulae does not guarantee that F should also contain 'infinitely deep' subformulae of a nonstandard formula in F.

In this case there is a unique F-satisfaction class on $\mathcal{M}$, known as the *Tarskian satisfaction class on* $\mathcal{M}$.

(c) S is a *full satisfaction class on* $\mathcal{M}$ if S is an *F-satisfaction class* for $F := \mathsf{Form}^{\mathcal{M}}$. This is equivalent to $(\mathcal{M}, S) \models \mathsf{PA}^{\mathsf{FS}}$.

Definition 2.4 $\Sigma_0 = \Pi_0 =$ the collection of $\mathcal{L}_{\mathsf{PA}}$-formulae all of whose quantifiers are of the form $\exists x < y\ \varphi$ or $\forall x < y\ \varphi$; Σ_{n+1} consists of formulae of the form $\exists x_0 \cdots \exists x_{k-1}\ \varphi$, where $\varphi \in \Pi_n$; and Π_{n+1} consists of formulae of the form $\forall x_0 \cdots \forall x_{k-1}\ \varphi$, where $\varphi \in \Sigma_n$. Here k ranges over ω, with the understanding that $k = 0$ corresponds to an empty block of quantifiers; this convention leads to the pleasant consequence that $\Sigma_n \subseteq \Sigma_{n+1}$ and $\Pi_n \subseteq \Pi_{n+1}$ for all n.

Theorem 2.5 *(Mostowski (Kaye 1991), (Hájek and Pudlák 1993)) For each nonzero $n < \omega$ there is a binary Σ_n-formula $\mathsf{Sat}_n(x, y)$ such that*

$$\mathit{PA} \vdash \mathit{Tarski}(\mathsf{Sat}_n, \Sigma_n),$$

where Σ_n is (the arithmetization of) the set of codes of formulae in Σ_n.

16.3 The Basic Construction

In this section we explain the basic methodology of building satisfaction classes using tools from model theory. The following lemma lies at the heart of the main result of this section.

Lemma 3.1 *Let $\mathcal{N}_0 \models \mathit{PA}$, $F_1 := \mathit{Form}^{\mathcal{N}_0}$, $F_0 \subseteq F_1$, and suppose S_0 is an F_0-satisfaction class. Then there is an elementary extension $\mathcal{N}_1$ of $\mathcal{N}_0$ that carries an F_1-satisfaction class $S_1 \supseteq S_0$ and $(c, \alpha) \in S_0$ whenever $c \in F_0$, $\alpha \in N_0$, and $(c, \alpha) \in S_1$.*

Proof Let $\mathcal{L}^+_{\mathsf{PA}}(\mathcal{N}_0)$ be the language obtained by enriching $\mathcal{L}_{\mathsf{PA}}$ with constant symbols for each member of N_0, and new *unary* predicates U_c for each $c \in \mathsf{Form}^{\mathcal{N}_0}$. It helps to have in mind that the intended interpretation of U_c is $\{\alpha \in A_c : S_1(c, \alpha)\}$, where $A_c := \{\alpha : \mathcal{N}_1 \models \mathsf{Asn}(\alpha, c)\}$.

We first wish to describe a new set of axioms

$$\Theta := \{\theta_c : c \in F_1\}$$

formulated in $\mathcal{L}^+_{\mathsf{PA}}(\mathcal{N}_0)$, where θ_c stipulates 'local Tarskian behavior' for U_c.
If $R \in \mathcal{L}_{\mathsf{PA}}$ and $\mathcal{N}_0 \models c = \ulcorner R(t_0, \cdots, t_{n-1})\urcorner$, then

$$\theta_c := \forall \alpha\ (\mathsf{U}_c(\alpha) \leftrightarrow \mathsf{Asn}(\alpha, c) \wedge R\ (\alpha(t_0), \cdots, \alpha(t_{n-1}))).$$

If $\mathcal{N}_0 \models c = \neg d$, then

$$\theta_c := \forall \alpha\ (\mathsf{U}_c(\alpha) \leftrightarrow \mathsf{Asn}(\alpha, c) \wedge \neg \mathsf{U}_d(\alpha)).$$

If $\mathcal{N}_0 \models c = d_1 \vee d_2$, then

$$\theta_c := \forall \alpha \left(\mathsf{U}_c(\alpha) \leftrightarrow \mathsf{Asn}(\alpha, c) \wedge \left(\mathsf{U}_{d_1}(\alpha \restriction \mathsf{FV}(d_1)) \vee \mathsf{U}_{d_2}(\alpha \restriction \mathsf{FV}(d_2)) \right) \right).$$

If $\mathcal{N}_0 \models c = \exists v_a\, b$, then

$$\theta_c := \forall \alpha \left(\mathsf{U}_c(\alpha) \leftrightarrow \mathsf{Asn}(\alpha, c) \wedge \exists \alpha' \supseteq \alpha\ \mathsf{U}_b(\alpha') \wedge \mathsf{Asn}(\alpha', b) \right).$$

Let

$$\Gamma := \{\mathsf{U}_c(\alpha) : c \in F_0 \text{ and } (c, \alpha) \in S_0\} \cup \{\neg \mathsf{U}_c(\alpha) : c \in F_0 \text{ and } (c, \alpha) \notin S_0\},$$

and let

$$\mathsf{Th}^+(\mathcal{N}_0) := \mathsf{Th}(\mathcal{N}_0, a)_{a \in N_0} \cup \Theta \cup \Gamma.$$

We now proceed to show that $\mathsf{Th}^+(\mathcal{N}_0)$ is consistent by demonstrating that each finite subset of $\mathsf{Th}^+(\mathcal{N}_0)$ is interpretable in $(\mathcal{N}_0, S_0)$. To this end, suppose T_0 is a finite subset of $\mathsf{Th}^+(\mathcal{N}_0)$ and let C consist of the collection of $c \in F_0$ such that U_c appears in T_0. If $C = \emptyset$, T_0 is readily seen to be consistent, so we shall assume that $C \neq \emptyset$ for the rest of the argument.

Our goal is to construct subsets $\{U_c : c \in C\}$ of N_0 such that the following two conditions hold when U_c is interpreted by U_c:

(1) $(\mathcal{N}_0, U_c)_{c \in C} \models \{\theta_c : c \in C\}$, and
(2) For $c \in C \cap F_0$, $U_c = \{\alpha \in N_0 : (c, \alpha) \in S_0\}$.

We shall construct $\{U_c : c \in C\}$ *in stages*, beginning with the simplest formulae in C, and working our way up using Tarski rules for more complex ones. Recall that $c \lhd d$ expresses "c is a direct subformula of d". Define $\lhd^*$ on C by:

$$c \lhd^* d \text{ iff } (c \lhd d)^{\mathcal{N}_0} \text{ and } \theta_d \in T_0 \cap \Theta.$$

Note that whenever $c \lhd^* d$, then for all $c' \lhd d$ we have $c' \in C$ and $c' \lhd^* d$. The finiteness of C implies that $(C, \lhd^*)$ is well-founded, which in turn helps us define a useful measure of complexity for $c \in C$ using the following recursive definition:

$$\mathsf{rank}_C(c) := \sup\{\mathsf{rank}_C(d) + 1 : d \in C \text{ and } d \lhd^* c\}.$$

Note that for $c \in C$, $\mathsf{rank}_C(c) = 0$ precisely when $\theta_c \notin T_0 \cap \Theta$. Next, let

$$C_i := \{c \in C : \mathsf{rank}_C(c) \leq i\}.$$

Observe that $C_0 \neq \emptyset$ (since C is finite and nonempty), and that if $c \in C_{i+1}$, then the codes of all immediate subformulae of the formula coded by c are in C_i. This observation ensures that the following recursive clauses yield a well-defined U_c for each $c \in C$.

- If $c \in C_0$ then $U_c := \begin{cases} \{\alpha : (c, \alpha) \in S_0\}, \text{if } c \in F_0; \\ U_c := \emptyset, \text{if } c \notin F_0. \end{cases}$
- If $c \in C_{i+1} \backslash C_i$ and $= \neg d$, then
$$U_c := \{\alpha \in A_c : \alpha \notin U_d\}.$$
- If $c \in C_{i+1} \backslash C_i$ and $c = a \vee b$, then
$U_c := \{\alpha \in A_c : \alpha \restriction \mathsf{FV}(a) \in U_a \text{ or } \alpha \restriction \mathsf{FV}(b) \in U_b\}$.
- If $c \in C_{i+1} \backslash C_i$ and $c = \exists v_a\, b$, then
$U_c := \left\{\alpha \in A_c : \exists \alpha' \in N\ (\alpha \subseteq \alpha' \text{ and } \alpha' \in U_b)\right\}$.

Note that in the first item above, the choice of $U_c := \emptyset$ when $c \in C_0$ and $c \notin F_0$ is completely arbitrary.[7] Also, in the third item above where $c = a \vee b$, both a and b will be in C_i, thanks to the properties of $\lhd^*$.

It is routine to verify, using induction on $\mathsf{rank}_C(c)$, that (1) and (2) hold for $(\mathcal{N}_0, U_c)_{c \in C}$. More specifically, if $\mathsf{rank}_C(c) = 0$, then $\theta_c \notin T_0 \cap \Theta$, so (1) is vacuously satisfied, and (2) is satisfied by design. On the other hand, when $\mathsf{rank}_C(c) > 0$ then (1) is satisfied since U_c is defined so as to comply with Tarski conditions; and (2) is satisfied since S_0 is an F_0-satisfaction class. This concludes the proof of the consistency of arbitrary finite subsets T_0 of $\mathsf{Th}^+(\mathcal{N}_0)$, which in turn shows that $\mathsf{Th}^+(\mathcal{N}_0)$ has a model, i.e., some elementary extension $\mathcal{N}_1$ of $\mathcal{N}_0$ has an expansion $\mathcal{N}_1^+$ of the form

$$\mathcal{N}_1^+ := (\mathcal{N}_1, U_c)_{c \in F_1}$$

with the property that $\mathcal{N}_1^+ \models \mathsf{Th}^+(\mathcal{N}_0)$. Let S_1 be the binary relation on N_1 defined via

$$S_1(c, \alpha) \Leftrightarrow \alpha \in U_c.$$

It is evident that S_1 is an F_1-satisfaction class, $S_1 \supseteq S_0$ and $(c, \alpha) \in S_0$ whenever $c \in F_0$, $\alpha \in N_0$, and $(c, \alpha) \in S_1$. □

Theorem 3.2 *Let $\mathcal{M}_0$ be a model of PA of any cardinality.*

(a) *If S_0 is an F_0-satisfaction class on $\mathcal{M}_0$, then there is an elementary extension $\mathcal{M}$ of $\mathcal{M}_0$ that carries a full satisfaction class that extends S_0.*
(b) *There is an elementary extension $\mathcal{M}$ of $\mathcal{M}_0$ that carries a full satisfaction class.*

Proof Note that (b) is an immediate consequence of (a) since we may choose F_0 to be the set of *atomic* $\mathcal{M}_0$-formulae and S_0 to be the obvious satisfaction predicate for F_0. To establish (a), we note that by Lemma 3.1 there is an elementary extension $\mathcal{M}_1$ of $\mathcal{M}_0$ that carries an F_1-satisfaction class, where $F_1 := \mathsf{Form}^{\mathcal{M}_0}$. Lemma 3.1 allows this argument to be carried out ω-times to yield two sequences $\langle \mathcal{M}_i : i \in \omega \rangle$ and $\langle S_i : i \in \omega \rangle$ that satisfy the following properties for each $i \in \omega$:

[7] As shown in (Enayat and Visser 2012) this feature can be exploited to construct 'pathological' satisfaction classes, such as the one mentioned at the end of Sect. 6 of this paper.

(1) $\mathcal{M}_i \prec \mathcal{M}_{i+1}$;
(2) S_{i+1} is an F_{i+1}-satisfaction class on $\mathcal{M}_{i+1}$ with $F_{i+1} := \mathsf{Form}^{\mathcal{M}_i}$; and
(3) $S_i = S_{i+1} \cap \{(c, \alpha) : c \in F_i, \mathcal{M}_i \models \mathsf{Asn}(\alpha, c)\}$.

Let $\mathcal{M} := \bigcup_{i \in \omega} \mathcal{M}_i$, and $S := \bigcup_{i \in \omega} S_i$. Tarski's elementary chain theorem and (1) together imply that $\mathcal{M}$ elementarily extends $\mathcal{M}_0$. It is easy to see, using (2) and (3), that S is a full satisfaction class on $\mathcal{M}$. □

Theorem 3.2, when coupled with the completeness theorem of first order logic, immediately yields the following conservativity result.

Corollary 3.3 *PA^{FS} is a conservative extension of PA.*

Proof Suppose not. Then for some arithmetical sentence φ we have:

(1) $\mathsf{PA}^{\mathsf{FS}} \vdash \varphi$, and
(2) $\mathsf{PA} \nvdash \varphi$.

Since (2) implies that $\mathsf{PA} \cup \{\neg\varphi\}$ is consistent, by the completeness theorem for first order logic, there is a model $\mathcal{M}_0 \models \mathsf{PA} \cup \{\neg\varphi\}$. On the other hand, by part (b) of Theorem 3.2 there is an elementary extension $\mathcal{M}_1$ of $\mathcal{M}_0$ that carries a full satisfaction class, and therefore by (1) $\mathcal{M}_1 \models \varphi$. This contradicts the fact that $\mathcal{M}_1$ elementarily extends $\mathcal{M}_0$. □

Corollary 3.4 *Every resplendent model of PA carries a full satisfaction class. In particular, every countable recursively saturated model of PA carries a full satisfaction class.*

Proof The first claim directly follows from the definition of a resplendent model. The second claim follows from the first claim, when coupled with the key result that countable recursively saturated models are resplendent (see (Kaye 1991, Sect. 15.2) for more detail). □

16.4 Truth Classes

With the exception of Krajewski's original paper (Krajewski 1976), what we refer to as a 'truth class' here has been dubbed 'satisfaction class' in the model-theoretic literature. More specifically, Krajewski (1976) employed the framework of satisfaction classes over base theories formulated in relational languages as in this paper, however, the later series of papers (Kotlarski et al. 1981; Smith 1987, 1989) all used the framework of truth classes over Peano arithmetic formulated in a relational language, augmented with 'domain constants'. Later, Kaye (1991) developed the theory of satisfaction classes over models of PA in languages incorporating function symbols; his work was extended by Engström (2002) to truth classes over models of PA in functional languages.

As explained in this section, there is a simple canonical correspondence between truth classes over models of PA (in a relational language) and certain types of satisfaction classes, here referred to as 'extensional'. The main aim of this section is to demonstrate that the method of building satisfaction classes in the previous section can be conveniently modified so as to yield full *extensional* satisfaction classes (and thereby: full truth classes) over appropriate models of PA.

Within PA one can easily define an injective function c that yields the code for a constant symbol $\overline{x}$ for each member x of the domain. This enables PA to internally represent the language $\mathcal{L}_{\mathsf{PA}}^{+} = \mathcal{L}_{\mathsf{PA}}+$ 'domain constants'. We can then add a unary predicate $\mathsf{T}(x)$ denoting a *truth class* (instead of a binary predicate $\mathsf{S}(x, y)$ for a satisfaction class) to $\mathcal{L}_{\mathsf{PA}}$, whose intended interpretation is "x is the code of a true *sentence* σ", where σ is an arithmetical sentence formulated in a language $\mathcal{L}_{\mathsf{PA}}^{+}$. We will make this more precise in the following definition.

Definition 4.1 $\mathsf{PA}^{\mathsf{FT}} := \mathsf{PA} \cup \mathsf{Tarski}(\mathsf{T})$, where $\mathsf{Tarski}(\mathsf{T})$ is the conjunction of the universal generalizations of $\mathsf{tarski}_0(\mathsf{T})$ through $\mathsf{tarski}_4(\mathsf{T})$, all formulated in the language $\mathcal{L}_{\mathsf{PA}} \cup \{\mathsf{T}(\cdot)\}$, as described below.[8] In what follows $\mathsf{Sent}(x)$ is the $\mathcal{L}_{\mathsf{PA}}$-formula that expresses "x is a formula of $\mathcal{L}_{\mathsf{PA}}^{+}$ with no free variables", and R ranges over relations symbols in $\mathcal{L}_{\mathsf{PA}}$.

- $\mathsf{tarski}_0(\mathsf{T}) := (\mathsf{T}(x) \rightarrow \mathsf{Sent}(x))$.
- $\mathsf{tarski}_{1,R}(\mathsf{T}) := \left(\ulcorner R(\overline{t}_0, \cdots, \overline{t}_{n-1})\urcorner = x\right) \rightarrow (R(t_0, \cdots, t_{n-1}) \leftrightarrow \mathsf{T}(x))$.
- $\mathsf{tarski}_2(\mathsf{T}) := (x = \neg y) \rightarrow (\mathsf{T}(x) \leftrightarrow \neg\mathsf{T}(y_1))$.
- $\mathsf{tarski}_3(\mathsf{T}) := (x = y_1 \vee y_2) \rightarrow (\mathsf{T}(x) \leftrightarrow (\mathsf{T}(y_1) \vee \mathsf{T}(y_2)))$.
- $\mathsf{tarski}_4(\mathsf{T}) := (x = \exists v_i \varphi) \rightarrow (\mathsf{T}(x) \leftrightarrow \exists z\, \mathsf{T}(\varphi(\overline{z})))$.

T is a *full truth class* on $\mathcal{M}$ if $(\mathcal{M}, T) \models \mathsf{PA}^{\mathsf{FT}}$.

Definition 4.2 A *substitution* for a formula ψ of $\mathcal{L}_{\mathsf{PA}}$ is a function

$$\sigma : \mathsf{FV}(\psi) \rightarrow \mathsf{Var}$$

such that σ respects substitutability in the 'usual way', i.e., if x is a free variable of ψ, then x is not in the scope of any quantifier that binds $\sigma(x)$. Given ψ and σ as above, let $\psi * \sigma$ be the formula obtained from ψ by applying the substitution σ, and A be the set of pairs (φ, α) such that α is an assignment for the formula φ. This allows us to define a key equivalence relation $\sim$ on A by decreeing that $(\varphi_0, \alpha_0) \sim (\varphi_1, \alpha_1)$ iff there is some $(\psi, \beta) \in A$, and there are substitutions σ_0 and σ_1 for ψ, with

$$\varphi_i = \psi * \sigma_i \text{ and } \beta = \alpha_i \circ \sigma_i, \text{ for } i = 0, 1.$$

In the above, $\alpha_i \circ \sigma_i$ is the composition of α_i and σ_i. It is important to bear in mind that, intuitively speaking, $(\varphi_0, \alpha_0) \sim (\varphi_1, \alpha_1)$ means that φ_0 and φ_1 are the same

[8] $\mathsf{PA}^{\mathsf{FT}}$ is the relational analogue of the theory of $\mathsf{CT}\restriction$ in Halbach's monograph (Halbach 2011). The base theory of $\mathsf{CT}\restriction$ is PA formulated in a *functional* language. The conservativity of $\mathsf{CT}\restriction$ over the functional language version of PA can also be established using the techniques of this paper (see Sect. 16.6).

except for their free variables, and for all variables x and y, if x occurs freely in the same position in φ_0 as y does in φ_1, then $\alpha_0(x) = \alpha_1(y)$.

- An F-satisfaction class S is *extensional* if for all φ_0 and φ_1 in F, $\mathcal{M} \models (\varphi_0, \alpha_0) \sim (\varphi_1, \alpha_1)$ implies $(\varphi_0, \alpha_0) \in S$ iff $(\varphi_1, \alpha_1) \in S$.[9]

The following proposition describes a canonical correspondence between extensional satisfaction classes and truth classes. The routine but laborious proof is left to the reader.

- In what follows c is the $\mathcal{M}$-definable injection $m \mapsto_{\mathsf{c}} \overline{m}$ that designates a constant symbol $\overline{m}$ for each $m \in M$, and $\varphi(\mathsf{c} \circ \alpha)$ is the sentence in the language $\mathcal{L}^+_{\mathsf{PA}}$ obtained by replacing each occurrence of a free variable x of φ with the constant symbol $\overline{m}$, where $\alpha(x) = m$.

Proposition 4.3 *Suppose $\mathcal{M} \models \mathsf{PA}$, T is a full truth class on $\mathcal{M}$, and S is an extensional full satisfaction class on $\mathcal{M}$.*

(a) *$\mathcal{S}(T)$ is an extensional satisfaction class on $\mathcal{M}$, where $\mathcal{S}(T)$ is defined as the collection of ordered pairs (φ, α) such that $\varphi(\mathsf{c} \circ \alpha) \in T$.*
(b) *$\mathcal{T}(S)$ is a truth class on $\mathcal{M}$, where $\mathcal{T}(S)$ is defined as the collection of $\varphi \in \mathcal{L}^+_{\mathsf{PA}}$ such that for some $\psi \in \mathcal{L}^+_{\mathsf{PA}}$ and some assignment α for ψ, $\varphi = \psi(\mathsf{c} \circ \alpha)$ and $(\psi, \alpha) \in S$.*
(c) *$\mathcal{S}(\mathcal{T}(S)) = S$, and $\mathcal{T}(\mathcal{S}(T)) = T$.*

Before describing the construction of extensional satisfaction classes we need the preliminaries presented in Definition 4.4 and Lemma 4.5.

Definition 4.4

(a) Given formulae φ_0 and φ_1 of $\mathcal{L}_{\mathsf{PA}}$, we write $\varphi_0 \approx \varphi_1$ if there is a formula ψ, and substitutions σ_0 and σ_1 for ψ such that $\varphi_i = \psi * \sigma_i$ for $i = 0, 1$.
(b) Given $c \in \mathsf{Form}^{\mathcal{M}}$, let $\mathbf{TC}_{\mathcal{M}}(c)$ be the *externally defined* transitive closure of c with respect to the direct subformula relation, i.e.,

$$\mathbf{TC}_{\mathcal{M}}(c) := \bigcup_{n<\omega} \mathbf{TC}_{\mathcal{M}}(c, n),$$

where $\mathbf{TC}_{\mathcal{M}}(c, 0) := \{c\}$ and

$$\mathbf{TC}_{\mathcal{M}}(c, n+1) := \{x \in M : x \lhd^{\mathcal{M}} d \text{ for some } d \in \mathbf{TC}_{\mathcal{M}}(c, n)\}.$$

The following lemma presents salient features of the two equivalence relations $\sim$ and $\approx$.

Lemma 4.5 *Let $\sim$ be as in Definition 4.2; and $\mathbf{TC}_{\mathcal{M}}(c)$ and $\approx$ be as in Definition 4.4.*

[9] Note that an extensional satisfaction predicate need not be closed under re-naming of bound variables.

(i) *If* $d \in \boldsymbol{TC}_{\mathcal{M}}(c)$ *and* $d \neq c$, *then* $\neg\,(c \approx\!\!\!\approx d)$.

(ii) $\approx\!\!\!\approx$ *preserves the principal connectives, i.e., it relates negations to negations, disjunctions to disjunctions, and existential formulae to existential formulae with the same bound variable. Moreover, if* $\neg c \approx\!\!\!\approx \neg d$, *then* $c \approx\!\!\!\approx d$; *if* $c \vee d \approx\!\!\!\approx c' \vee d'$, *then* $c \approx\!\!\!\approx c'$ *and* $d \approx\!\!\!\approx d'$; *and if* $\exists t\ c \approx\!\!\!\approx \exists t'\ c'$, *then* $t = t'$ *and* $c \approx\!\!\!\approx c'$.

(iii) *If* $(\varphi_0, \alpha_0) \sim (\varphi_1, \alpha_1)$, *then* $\varphi_0 \approx\!\!\!\approx \varphi_1$.

(iv) *If* $(\neg\varphi_0, \alpha_0) \sim (\neg\varphi_1, \alpha_0)$, *then* $(\varphi_0, \alpha_0) \sim (\varphi_1, \alpha_1)$.

(v) *If* $(\varphi_0 \vee \varphi_1, \alpha) \sim (\varphi_0' \vee \varphi_1', \alpha')$, *then* $(\varphi_0, \alpha \restriction \mathit{FV}(\varphi_0)) \sim \left(\varphi_0', \alpha' \restriction \mathit{FV}(\varphi_0')\right)$ *and* $(\varphi_1, \alpha \restriction \mathit{FV}(\varphi_1)) \sim \left(\varphi_1', \alpha' \restriction \mathit{FV}(\varphi_1')\right)$.

(vi) *If* $\varphi = \exists t\ \psi$, *and* $\varphi' = \exists t'\ \psi'$, *and* $(\varphi, \alpha) \sim (\varphi', \alpha')$, *then* $t = t'$ *and for some* e

$$(\varphi, \alpha[t : e]) \sim (\varphi', \alpha'[t' : e]).^{10}$$

The next Lemma presents a variant of Lemma 3.1 that is our main tool for constructing extensional satisfaction classes.

Lemma 4.6 *Let* $\mathcal{N}_0 \models \mathit{PA}$, $F_1 := \mathit{Form}^{\mathcal{N}_0}$, $F_0 \subseteq F_1$, *and suppose* S_0 *is an extensional* F_0*-satisfaction class. Then there is an elementary extension* $\mathcal{N}_1$ *of* $\mathcal{N}_0$ *that carries an extensional* F_1*-satisfaction class* $S_1 \supseteq S_0$ *and* $(c, \alpha) \in S_0$ *whenever* $c \in F_0$, $\alpha \in N_0$, *and* $(c, \alpha) \in S_1$.

Proof Let Θ and Γ be as in the proof of Lemma 3.1, and let

$$\mathsf{Th}^{+}(\mathcal{N}_0) := \mathsf{Th}(\mathcal{N}_0, a)_{a \in N_0} \cup \Theta \cup \Gamma \cup \Delta,$$

where $\Delta := \left\{\delta_{cc'} : c,\ c' \in F_1\right\}$, and

$$\delta_{cc'} := \forall\alpha\ \forall\alpha' \left((c, \alpha) \sim (c', \alpha') \rightarrow \left(\mathsf{U}_c(\alpha) \leftrightarrow \mathsf{U}_{c'}(\alpha')\right)\right).$$

The proof of the lemma would be complete once we verify that $\mathsf{Th}^{+}(\mathcal{N}_0)$ has a model. To this end, we shall demonstrate that every finite subset T_0 of $\mathsf{Th}^{+}(\mathcal{N}_0)$ is interpretable in $\mathcal{N}$. Let C be the collection of $c \in F_1$ such that c appears in T_0. Also, let $\lhd^{*}$ and $\mathsf{rank}_C(c)$ be precisely as in the proof of Lemma 3.1.

We can extend C to another finite set $\overline{C}$ so that it satisfies a certain closure property, namely: whenever we have $c \approx\!\!\!\approx c'$ and $d \lhd^{*} c$, where c, c' and d are all in $\overline{C}$, then there is some $d' \in \overline{C}$ such that $d' \lhd^{*} c'$ with $d \approx\!\!\!\approx d'$. This can be done simply by adding any missing direct subformulae d' by an appropriate recursion.[11] By replacing C by $\overline{C}$ we may therefore additionally assume:

---

[10] Here $\alpha[t : e]$ is the assignment obtained by redefining the value of α at the variable t to be e if $t \in \mathrm{Dom}(\alpha)$; note that $\alpha[t : e] = \alpha$ if $t \notin \mathrm{Dom}(\alpha)$.

[11] More specifically, first define $\lhd^{\circ}$ on C by $d' \lhd^{\circ} c$ iff $d' \lhd^{*} c' \approx\!\!\!\approx c$, for some $c' \in C$. Since $\lhd^{\circ}$ is cycle-free, C is well-founded, and therefore lends itself to a ranking function $\mathsf{rank}^{\circ}_C(c)$. Let $n = \max\left\{\mathsf{rank}^{\circ}_C(c) : c \in C\right\}$, and for $0 \leq i \leq n$ define $D_i := \{c \in C : \mathsf{rank}^{\circ}_C(c) = i\}$. Next use a 'backward' recursion to define $E_n, E_{n-1}, \cdots, E_0$ via:

- $E_n := D_n$;
- $E_{n-(i+1)} := D_{n-(i+1)} \cup \{d : d \lhd^{\mathcal{N}_0} c \text{ for some } c \in E_{n-i}\}$.

Finally, let $\overline{C} := E_n \cup \cdots \cup E_0$. It is easy to see that $\overline{C}$ is finite, extends C, and has the desired closure property.

(#) If c and c' are both in C with $c \approx\!\!\!\approx c'$, then $\mathsf{rank}_C(c) = \mathsf{rank}_C(c')$.
As in the proof of Lemma 3.1, we then recursively construct $\{U_c : c \in C\}$ such that:

(1) $(\mathcal{N}_0, U_c)_{c\in C} \models \{\theta_c : c \in C\}$ and
(2) For $c \in C \cap F_0$, $U_c = \{\alpha \in N_0 : (c, \alpha) \in S_0\}$.

It remains to show:

(3) $(\mathcal{N}_0, U_c)_{c\in C} \models \{\Delta_{cc'} : c,\ c' \in C\}$.

We establish (3) by using induction on $\mathsf{rank}_C(c)$ to show that $\forall c \in C\ P(c)$, where $P(c)$ abbreviates:

$$\forall c' \in C\ \ (\mathcal{N}_0, U_c)_{c\in C} \models \forall\alpha\ \forall\alpha'\ \big((c, \alpha) \sim (c', \alpha') \rightarrow \big(\mathsf{U}_c(\alpha) \leftrightarrow \mathsf{U}_{c'}(\alpha')\big)\big).$$

If $\mathsf{rank}_C(c) = 0$ and $(c, \alpha) \sim (c', \alpha')$, then by part (iii) of Lemma 4.5 we have $c \approx\!\!\!\approx c'$, which in turn by (#) assures us that $\mathsf{rank}_C(c') = 0$. This makes it clear that $P(c)$ holds when $\mathsf{rank}_C(c) = 0$ since S_0 is assumed to be an *extensional* F_0-satisfaction class.

To verify the inductive step, suppose:

(4) $P(x)$ holds for all $x \in C$ with $\mathsf{rank}_C(x) = i$.

Let $c \in C$ with $\mathsf{rank}_C(c) = i + 1$, and suppose $(c, \alpha) \sim (c', \alpha')$, where $c = \exists t\ d$. Then $c' = \exists t'\ d'$, and $d \approx\!\!\!\approx d'$ by part (ii) of Lemma 4.5. Observe that thanks to (#) we have:

(5) $\mathsf{rank}_C(c') = i + 1$ and $\mathsf{rank}_C(d) = \mathsf{rank}_C(d') = i$.

Now if $\alpha \in U_c$, then $\alpha[t : e] \in U_d$ for some e by (1), and therefore by part (vi) of Lemma 4.5, we obtain:

(6) $(d, \alpha[t : e]) \sim (d', \alpha'[t' : e])$.

Using (4), (5), and (6) we may now conclude that $\alpha'[t : e] \in U_{d'}$, which by (1) yields $\alpha' \in U_d$, thus completing the verification of the quantificational case (by symmetry). A similar reasoning can be carried out for propositional cases. This concludes the proof of consistency of T_0.

The rest is precisely as before: the consistency of $\mathsf{Th}^+(\mathcal{N}_0)$ implies that there is an elementary extension $\mathcal{N}_1$ of $\mathcal{N}_0$ that has an expansion $\mathcal{N}_1^+ := (\mathcal{N}_1, S) \models \mathsf{Th}^+(\mathcal{N}_0)$, and the binary relation S_1 on $\mathcal{N}_1$ defined via

$$S_1(c, \alpha) \Leftrightarrow \alpha \in U_c$$

has the property that S_1 is an extensional F_1 -satisfaction, $S_1 \supseteq S_0$, and $(c, \alpha) \in S_0$ whenever $c \in F_0$, $\alpha \in N_0$, and $(c, \alpha) \in S_1$. □

Theorem 4.7 *Let $\mathcal{M}_0 \models \mathsf{PA}$. There is an elementary extension $\mathcal{M}$ of $\mathcal{M}_0$ that carries a full extensional satisfaction class.*

Proof Since the satisfaction class S_0 on the collection F_0 of atomic formulae of $\mathcal{M}_0$ is extensional, we may use Lemma 4.6 instead of Lemma 3.1 in order to carry out the elementary chain argument of Theorem 3.2. □

By coupling Theorem 4.7 with part (b) of Proposition 4.3 we obtain:

Corollary 4.8 *Every model of PA has an elementary extension that carries a full truth class.*

Finally, the line of reasoning employed in the proof of Corollary 3.3 shows, using Corollary 4.8, that:

Corollary 4.9 PA^{FT} *is a conservative extension of PA.*

16.5 Arithmetization, Interpretability, and Conservativity

Here we briefly discuss the *arithmetization* of the constructions of the previous two sections, with an eye towards issues connected with interpretability and conservativity. As explained in (Enayat and Visser, Sect. 4) the compactness and elementary chain argument employed in the proofs of Theorems 3.2 and 4.7 can be implemented in the fragment $\mathsf{I}\Sigma_2$ of PA with the help of the 'Low Basis Theorem' of Recursion Theory. Coupled with Orey's Compactness Theorem, this can used to establish the following:

Theorem 5.1 *(Enayat and Visser)* PA^{FT} *is interpretable in PA.*[12]

On the other hand, the technology of LL_1-sets[13] of (Hájek and Pudlák 1993, Theorem 4.2.7.1, p. 104) can be used to show that the proofs of both theorems 3.2 and 4.7 can even be implemented in the fragment $\mathsf{I}\Sigma_1$ of PA. In light of the fact that the statement "$\mathsf{PA}^{\mathsf{FT}}$ is conservative over PA" is a Π_2-statement, and $\mathsf{I}\Sigma_1$ is well known[14] to be Π_2-conservative over PRA, we obtain the following:

Theorem 5.2 *(Enayat and Visser) The conservativity of* PA^{FT} *over PA can be verified in PRA.*[15]

Remark 5.3 The verification of the conservativity of $\mathsf{PA}^{\mathsf{FT}}$ over PA within PRA was first claimed by Halbach in (Halbach 1999), using cut-elimination.[16] Later, Fischer (2009) gave a proof, based on the cut-elimination argument in (Halbach 1999), to show that $\mathsf{PA}^{\mathsf{FT}}$ is interpretable in PA. Unfortunately, a gap was discovered recently

[12] Indeed B^{FS} turns out to be interpretable in B for all base theories B that have access to the full scheme of induction over their ambient 'numbers'. In particular, $\mathsf{ACA}^{\mathsf{FS}}$ is interpretable in ACA. On the other hand, as shown in (Enayat and Visser, Sect. 8), $\mathsf{ACA}_0^{\mathsf{FS}}$ is *not* interpretable in ACA_0 (more generally, B^{FS} is shown to be *not* interpretable in B, if B is finitely axiomatizable).

[13] LL_1-sets are a special type of 'low' sets.

[14] This classical result was independently established by Mints, Parsons, and Takeuti, using proof-theoretic methods. The work of Paris and Kirby (described in (Simpson 1999, IX.3)), and more recently Avigad (2002) has also provided model-theoretic demonstrations of this conservativity result.

[15] Indeed, by using the technique of Friedman (1999), this conservativity result is already verifiable in the fragment SEFA (Superexponential Arithmetic) of PRA.

[16] Halbach's base theory in his work is the usual version of PA that is formulated in a functional language.

(by Fujimoto) in the cut-elimination argument in (Halbach 1999), which in turn impaired Fischer's interpretability claim. Happily, Leigh (2012) has succeeded in developing a proof-theoretic demonstration of the conservativity of $\mathsf{PA}^{\mathsf{FT}}$ over PA that is implementable in PRA. Moreover, (Leigh 2012, Theorem 1) can be used to verify the interpretability of $\mathsf{PA}^{\mathsf{FT}}$ over PA, by using Fischer's strategy in (Fischer 2009).[17] Therefore, Theorems 5.1 and 5.2 can be arrived at via two completely different routes.

16.6 Further Results

In Sect. 16.4 we saw that the core methodology of Sect 3 can be fine-tuned to build full extensional satisfaction classes. Indeed, as shown in (Enayat and Visser 2012) one can strengthen Theorem 4.7 by imposing further desirable conditions on the satisfaction class S. For example, every model $\mathcal{M}_0$ of PA has an elementary extension $\mathcal{M}$ that carries a full extensional satisfaction class S that satisfies all of the following additional properties:

(1) $\mathsf{Sat}_n^{\mathcal{M}} \subseteq S$ for all $n \in \omega$ (see Theorem 2.5 for Sat_n).
(2) If $c \in \mathsf{Form}^{\mathcal{M}}$ and $\mathcal{M} \models$ "c is an axiom of PA", then S deems c to be 'true'.[18]
(3) If c and c' are $\mathcal{M}$-formulae such that $\mathcal{M} \models$ "c' is an alphabetic variant[19] of c ", then $(c, \alpha) \in S$ iff $(c', \alpha) \in S$.

Furthermore, the third condition above can be strengthened by accomodating a combination of extensional equivalence and alphabetic equivalence, thereby yielding truth classes that are closed under alphabetic equivalence. We have also shown that a small dose of condition (1) can be used to build full truth classes over models of arithmetical theories formulated in *functional* languages (this result will appear in the projected sequel to (Enayat and Visser 2012)).

One can also use the method of Sect. 16.3 to build bizarre satisfaction classes. For example, as shown in (Enayat and Visser), every model $\mathcal{M}_0$ of PA has an elementary extension $\mathcal{M}$ that has a full satisfaction class S that exhibits the following pathology:

$$\{a \in M : (\sigma_a, \alpha_{\mathsf{Null}}) \in S\} = \omega_{\mathcal{M}},$$

where $\omega_{\mathcal{M}}$ is the well-founded initial segment of $\mathcal{M}$ that is isomorphic to ω, and σ_a is defined for all $a \in M$ by a recursion within $\mathcal{M}$ via the following clauses:

- $\sigma_0 := \exists v_0\, (v_0 = v_0)$ (or $\sigma_0 =$ any other logically valid sentence);
- $\sigma_{n+1} := (\sigma_n \vee \sigma_n)$.

---

[17] We are grateful to Graham Leigh for his kind permission to quote his unpublished work here.
[18] As remarked in the last sentence of (Kotlarski et al. 1981), this condition can also be arranged using the machinery of $\mathcal{M}$-logic. Note that 'axioms of PA' in the sense used here do not include the logical axioms.
[19] c' is an alphabetic variant of c if c' is obtainable from c by the usual rules of re-naming the bound variables of c.

References

Avigad, J. (2002). Saturated models of universal theories. *Annals of Pure and Applied Logic,* 118*,* 219–234.

Enayat A., & Visser, A. (2012) Full satisfaction classes in a general setting (Part I), to appear.

Engström, F. (2002). Satisfaction classes in nonstandard models of first-order arithmetic. arXiv.org.math. http://arxiv4.library.cornell.edu/abs/math/0209408v1. Accessed 3 March 2015

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

Friedman, H. (1999). Finitist proofs of conservation. FOM Archives. http://cs.nyu.edu/pipermail/fom/1999-September/003405.html. Accessed 3 March 2015.

Hájek, P., & Pudlák, P. (1993). *Metamathematics of first-order arithmetic*. Springer.

Halbach, V. (1999). Conservative theories of classical truth. *Studia Logica,* 62*,* 353–370.

Halbach, V. (2011). *Axiomatic theories of truth*. Cambridge: Cambridge University Press.

Kaye, R. (1991). *Models of peano arithmetic, oxford logic guides*. Oxford: Oxford University Press.

Kotlarski, H., Krajewski, S., & Lachlan, A. H. (1981). Construction of satisfaction classes for nonstandard models. *Canadian mathematical bulletin,* 24*,* 283–293.

Krajewski, S. (1976). Nonstandard satisfaction classes. In W. Marek Set theory and hierarchy theory: A memorial tribute to Andrzej Mostowski (Vol. 537, pp. 121–144). Berlin: Springer-Verlag.

Leigh, G. E. (2012). Conservativity for theories of compositional truth via cut elimination, to appear in Journal of Symbolic Logic.

McGee, V. (2003). In praise of the free lunch: Why disquotationalists should embrace compositional semantics, in Self-reference, CSLI Lecture Notes, 178, CSLI Publ. Stanford, pp. 95–120.

Simpson, S. (1999). *Subsystems of second order arithmetic, perspectives in mathematical logic*. Berlin: Springer-Verlag.

Smith, S. T. (1984). Non-standard syntax and semantics and full satisfaction classes, Ph.D. thesis, Yale University, New Haven, Connecticut.

Smith, S. T. (1987). Nonstandard characterizations of recursive saturation and resplendency. *Journal of Symbolic Logic,* 52*,* 842–863.

Smith, S. T. (1989). Nonstandard definability.*Annals of Pure and Applied Logic,* 42*,* 21–43.

Part V
Truth Without Paradox

Chapter 17
Truth, Pretense and the Liar Paradox

Bradley Armour-Garb and James A. Woodbridge

Abstract In this paper we explain our pretense account of truth-talk and apply it in a diagnosis and treatment of the Liar Paradox. We begin by assuming that some form of deflationism is the correct approach to the topic of truth. We then briefly motivate the idea that all T-deflationists should endorse a fictionalist view of truth-talk, and, after distinguishing pretense-involving fictionalism (PIF) from error-theoretic fictionalism (ETF), explain the merits of the former over the latter. After presenting the basic framework of our PIF account of truth-talk, we demonstrate a few advantages it offers over T-deflationist accounts that do not explicitly acknowledge pretense at work in the discourse. In turning to the Liar Paradox, we explain how the quasi-anaphoric functioning that our account attributes to truth-talk provides a diagnosis of the Liar Paradox (and other instances of semantic pathology) as having no content—in the sense of not specifying any of what we call *M-conditions*. At the same time, however, we vindicate the intuition that we can understand liar sentences, thereby avoiding one standard objection to "meaningless strategy" responses to the Liar Paradox. With this diagnosis in place, we then, by way of treatment, introduce a new predicate, 'semantically defective', and show how the explanation we give for its application allows for a consistent, yet revenge-immune, (dis)solution of the Liar Paradox, and semantic pathology generally.

Thanks to Graham Priest, Anil Gupta and to the other participants and organizers of both *BW7 | Seventh Barcelona Workshop on Issues in the Theory of Reference: Paradoxes of Truth and Denotation* (LOGOS: Logic, Language, and Cognition Research Group, Universitat de Barcelona) and *Truth at Work* (IHPST: Institut d'Historie et de Philosophie des Sciences et Techniques, Université Paris-1, et Ecole Normale Supérieure) for helpful (and enjoyable) comments and discussion.

Both authors contributed equally to this work and are listed in alphabetical order.

B. Armour-Garb
Department of Philosophy, HU-257 University at Albany-SUNY, 1400 Washington Avenue, Albany, NY, 12222, USA
e-mail: barmour-garb@albany.edu

J. A. Woodbridge
Department of Philosophy, University of Nevada Las Vegas, 4505 S. Maryland Pkwy, Box 455028, Las Vegas, NV, 89154, USA
e-mail: james.woodbridge@unlv.edu

T. Achourioti et al. (Eds.), *Unifying the Philosophy of Truth,* Logic, Epistemology, and the Unity of Science 36, DOI 10.1007/978-94-017-9673-6_17

17.1 Introduction

We accept the following three theses, with respect to truth:

T1: Some form of deflationism about truth (henceforth, 'T-deflationism') is the correct view of truth (or, better, of *truth-talk*, that fragment of discourse that includes the truth-predicate and other alethic predicates) to adopt;
T2: The best way to construe T-deflationism is to see it as understanding truth-talk to involve a sort of fiction; and
T3: The most plausible form of truth-theoretic fictionalism involves semantic pretense.

Hence, our chief claim, with respect to truth-talk, is that it should be understood as operating in terms of semantic pretense.

Our aim in this paper is not to argue for these theses; rather, our intent is to explain and motivate them and to display their merits. In so doing, we will not address T1. Regarding T2 and T3, after briefly explaining why T-deflationists should endorse, or adopt, a fictionalist account of truth-talk, we present the basic elements of a semantic pretense account of that talk. Once we have that on the table, we identify a nice feature of our view—that it affords a consistent solution to the Liar Paradox.[1]

17.2 Why T-Deflationists Should Adopt a Fictionalist Account of Truth-Talk

In order to motivate endorsing a fictionalist view of truth-talk, we begin with a general thesis that has been employed to motivate fictionalism about certain fragments of discourse, that of *expressive indispensability*, viz.,

(EI) We need to enlist certain aspects of X-talk, as a means for expressing certain claims that we could not otherwise express.[2]

In the case of truth-talk, we can see (EI) in effect in our widely acknowledged need to enlist the truth-predicate, which appears to commit us to a property of truth, as a means for expressing certain claims (to be discussed below) that we otherwise could not—or, at least, could not so easily—express.

The impetus for moving to a fictionalist account of truth-talk begins with the T-deflationist's thought that what we are trying to say through our use of truth-talk has nothing to do with truth *per se* and, in fact, but for certain *expressive limitations*, could be expressed without appeal to any such property. Moreover, such expressive

[1] This is in marked contrast with Woodbridge (2005), where a dialetheic approach to the Liar Paradox is endorsed within an earlier version of the sort of pretense account of truth-talk that we currently champion.
[2] While not explicitly formulating the thesis in this way, Yablo (2005) relies on something like (EI) to argue for a particular fictionalist account of mathematical discourse.

needs have nothing to say about truth—its nature or even its existence—or about whether there need be any such property, in order to express what we aim to convey. So, while truth-talk does appear to be *expressively* indispensable, truth, *qua* property, may well be *theoretically dispensable*.

The connection between understanding the notion of truth in this way (as theoretically dispensable but expressively indispensable) and truth-theoretic fictionalism is as follows. Suppose that we can explain the expressive advantages of appealing to truth-talk, and suppose, with the T-deflationist, that these expressive purposes exhaust our use of that talk. Suppose, finally, that what we are trying to get across through our use of truth-talk is not *about* any property of truth, in the sense that what we aim to convey itself has nothing to do with any such property. In that case, because 'true'—the notion of truth, as it occurs in truth-talk—serves essentially in the indirect expression of facts that are not about truth, it simply functions as a *representational aid*.[3] As we understand things, when the central locutions of some fragment of discourse function as representational aids in this way—making as if to talk about one thing for the purposes of talking about something else indirectly—that just *is* for that fragment of discourse to operate *via* some element of fiction.

According to the line of reasoning just sketched, T-deflationists should see truth-talk as operating through some element of fiction. After all, T-deflationists acknowledge the *expressive* indispensability of truth-talk, but they do not then go on to conclude that the truth-predicate is "ontologically serious". Rather, they hold that the truth-predicate functions as a device that allows speakers to talk indirectly about something else, facilitating the expression of facts that are not about truth.

17.3 Fictionalism via Semantic Pretense

We take the above to provide a reason for concluding that T-deflationists should endorse a fictionalist view of truth-talk.[4] Now, there are myriad versions of fictionalism that have been presented, and we do not have space to consider all of the available options here (though for further discussion, see our (2015)). We will point out, though, that not all fictionalist accounts are the same. Elsewhere (2009, 2012, 2014, 2015) we have highlighted an important distinction between fictionalist accounts that are properly described as "error-theoretic fictionalism" (henceforth, *ETF*) and those that are *not* error-theoretic.

The fictionalist approach that we favor is a pretense-involving fictionalism (henceforth, *PIF*) that is non-error-theoretic; it allows that claims from a fragment of discourse understood as involving pretense can still be used to make genuinely true claims. We (2009, 2012, 2015) have argued that certain philosophical problems can

[3] Cf. Yablo (2001, 2005).

[4] For more on this, see Armour-Garb and Woodbridge (2010, 2014).

be solved, by taking PIF to apply to ways of talking where we have not noticed it at work before.[5]

The particular variety of PIF that we favor with respect to truth-talk is a semantic pretense approach.[6] This approach involves postulating a semantic mechanism at work in the linguistic functioning of the relevant fragment of discourse, involving a special, but familiar kind of pretense, viz., *make-believe*. Make-believe games (e.g., the classic children's games of "mudpies", cowboys and Indians, cops and robbers, etc.) involve pretenses of two types. In the first type, certain pretenses are stipulated, or *expressly pretended*—typically about the props that are employed in the game of make-believe (e.g., globs of mud counting as pies, sticks counting as horses, fingers counting as pistols). The second type involves pretenses that are "generated from reality" *via* the game of make-believe's *principles of generation* (e.g., it is to be pretended that someone has put a pie in the oven whenever he has put a glob of mud into the hollow stump). (Cf. Crimmins (1998).) These principles are rules for the make-believe that establish a systematic dependency between some of what is to be pretended—that is, which pretenses are *prescribed*—and real-world conditions that are, as it were, outside of the game.

A semantic pretense account of some fragment of discourse appeals to the structural features that make-believe involves and the kinds of systematic dependencies that games involving make-believe exhibit. Postulating such dependencies as holding for the claims from some discourse can explain how speakers can use utterances from the discourse to say indirectly things that the utterances appear unsuited to say. This is done by making utterances that, in a sense, *belong* to a game of make-believe involving the characteristic locutions of the discourse. A typical merit of the approach is that it allows us to use readily available, familiar linguistic resources—ordinary object-talk, predication, and objectual quantification—in order to make much more complicated and technical claims indirectly.[7]

The semantic pretense approach has been fruitfully applied to many areas of philosophically interesting discourse, including, but not limited to, talk putatively of fictional entities, existence-talk, mathematical discourse, possible-worlds-talk, talk of abstract objects (properties, propositions), moral discourse, talk of unobservable entities in science, etc. We will now explain its application in our pretense account of truth-talk.

---

[5] As seems clear, an ETF account of truth-talk would be intolerable, as it would render false all truth attributions, thereby undermining the status of the T-schema, since not every instance of it would be true.

[6] The source of the semantic pretense approach is Kendall Walton's (1990, 1993) analyses of representation in the arts and of certain kinds of metaphor in terms of make-believe. Walton applies a semantic pretense approach explicitly in his analyses of talk about works of fiction and fictional entities and of existence-talk. See Evans (1982) for a different but related semantic pretense account of existence-talk.

[7] For more on the details of make-believe and its role in semantic pretense, see Richard (2000) and Woodbridge and Armour-Garb (2009).

17.4 A Semantic Pretense Account of Truth-Talk

Our semantic pretense account of truth-talk is sometimes presented with the slogan "Truth is a pretense," the idea being that, really, there is no property of truth; we just talk as if there were.[8] That is, we speak *as if* we are describing things as having or lacking properties called "truth" and "falsity", in order to make claims of other (more complicated) sorts indirectly. On our view, truth-talk relies on a game of make-believe, one that allows us to use familiar linguistic resources in order to make more complex claims that would otherwise involve technical and unfamiliar linguistic and logical devices that ordinary language does not even contain explicitly. The kinds of devices that we have in mind are things like schematic sentence variables and substitutional quantifiers, understood as serving to encode infinite conjunctions or infinite disjunctions.

Understanding truth-talk in terms of semantic pretense is to take it as underwritten by a game of make-believe governed, at least in part, by principles of generation like the following:

(I) It is to be pretended that expressions like 'is true' and 'is false' function predicatively to describe objects as having or lacking certain properties (called "truth" and "falsity").

(II) The pretenses displayed in an utterance of ⌜(The proposition) that p is true⌝ are prescribed if and only if p.

(III) The pretenses displayed in an utterance of ⌜(The proposition) that p is false ⌝ are prescribed if and only if $\sim$ p.

(IV) If S_1 and S_2 are sentences that are alike except (in some transparent context) one has a subsentence ⌜p⌝ where the other has ⌜ $<$ p $>$ is true⌝ then one can directly infer S_1 from S_2 and S_2 from S_1 (where 'p' serves as a variable that can be replaced by a sentence and ' $<$ p $>$ ' stands for a nominalization of such a sentence).

The roles and functions of Rules (I)—(IV) should be understood as follows.

Rule (I) states one of the stipulated, *expressly* made-believe, background pretenses for the relevant game of make-believe. In particular, it specifies certain linguistic expressions as the props for the game and explains what is to be pretended *about* such props. One consequence of this rule is that uses of 'true' and 'false' involve pretense *intrinsically*, which is to say: There are no pretense-free uses of truth-locutions because pretense is invoked in their basic functioning. As a consequence, the only *serious* content (about the real world, outside of the make-believe) that

[8] In order to deflect a possible misinterpretation, we should make clear that we are *not* saying that being true is a matter of being pretended true. There is an important difference between claiming something is true—and the pretenses always involved in such a claim—and *pretending* that something is true. When we claim, or assert, (e.g.) that a given sentence is true, we are *not* pretending that it is true. On our view—and this is in line with T-deflationism—we are indirectly expressing a commitment to what that sentence says.

an instance of truth-talk has (or: possesses) must come from the operation of the make-believe's *principles of generation*—specifically, Rules (II) and (III).

Rules (II) and (III) cover what are arguably the most basic cases of truth-talk, what we call "transparent propositional truth-talk", so an account of them provides a core for our more general account. These principles of generation make the correctness of a putative attribution of truth or falsity, to some nominalized, sentential content-vehicle, a function (possibly negating) of whether the conditions specified by a use of that content-vehicle obtain. This makes the utterance of an instance of truth-talk an indirect means for specifying those very same conditions, thus determining the content of the instances of truth-talk. Since these indirectly specified conditions can actually obtain, this makes it possible for instances of truth-talk to make (what we might, now employing the very pretense being explained, describe as) "genuinely true" claims about the world outside of the pretense.

Rule (IV) satisfies an important condition of adequacy for any T-deflationary theory of truth-talk, as it provides a version of a rule of *intersubstitution*. Such a rule further captures the sense in which the content of a putative ascription of truth to some content-vehicle just is the content of the content-vehicle itself. The general intersubstitution licensed by this rule is integral to a pretense account yielding the right content for the more interesting cases of truth-talk, viz., 'true'-involving generalizations. Since those cases are what give truth (or, more accurately, 'true') its point, it seems to be a fairly central aspect of any adequate account of truth-talk, and it is integral that any pretense account of truth-talk accommodate it. (Cf. Quine (1986) and Field (2008).)

To illustrate how truth-talk functions according to this account, consider a straightforward instance of truth-talk, such as

(1) It is true that crabapples are edible.

For reasons explained elsewhere (2012), we understand (1) to be more perspicuously rendered by

(1') That crabapples are edible is true,

where 'that crabapples are edible' is (in the context of the pretense) a referring expression that picks out the proposition that crabapples are edible. Syntactically speaking, the 'that'-clause is a nominalization of the content-vehicle

(2) Crabapples are edible.

When asserted, a 'true'-involving sentence like (1') presents the pretenses it displays as prescribed, where being prescribed is determined by:

a. the particular principle of generation that governs those pretenses (here, Rule (II)), and
b. whether the conditions, whose obtaining those principles make prescriptive for the pretenses, actually obtain.

Recall that Rule (II) has it that the prescriptive conditions for the pretenses displayed in (1') are those specified by the content-vehicle that is nominalized as the subject

expression of (1')—in this case, (2). In short, by presenting the pretenses it displays as prescribed, (1') specifies indirectly precisely those conditions that (2) specifies directly.

Our talk of the conditions specified by a claim is a central component of what we mean by the *content* of an utterance. However, because, as explained above, a central claim of the pretense account of truth-talk is that, really, there is no property of truth, we do not hold that these conditions can be understood fundamentally as truth-conditions. We call them *M-conditions*. While M-conditions can obtain or fail to obtain, on our view, truth-conditions have only a thin, derivative status, as conditions for the appropriate use of the truth-predicate. The truth-conditions for a sentence are a *by-product* of its meaning, of which M-conditions are a significant component. This thought is in line with the meaning-to-truth conditional,

(MTC) If S means that p, then S is true iff p,

no instance of which we reject.

Now, while some sentences specify M-conditions directly, as is the case with (2), other sentences specify M-conditions only indirectly. Indeed, as should be apparent, one of the consequences of our pretense account of truth-talk is that any specification of M-conditions (that obtain or fail to obtain outside of the pretense) that is accomplished by a 'true'-involving sentence will be accomplished only indirectly, via the operation of the pretense that governs the functioning of the truth-predicate.

The resulting identity of content between an instance of transparent propositional truth-talk of the form *It is true that p* and the content-vehicle nominalized in it (the sentence that goes in for 'p') means that the game of make-believe behind truth-talk generates all the instances of the equivalence schema

(ES) It is true that p iff p.[9]

This is an important result because, as T-deflationists have argued, these equivalences are (some of) the central principles governing truth-talk. Our pretense account has them follow directly from the functioning that truth-talk is given by the game of make-believe that underwrites it.

As is well known, the *real* usefulness of truth-talk is not the kind of use made in (1)/(1'); rather, it is in the use of the predicate 'true' to express generalizations of a certain sort, as might be found in an utterance of

(3) Everything Isabel says is true.

Now, (3) might be expanded to

(4) Everything is such that if Isabel says it, then it is true,

and will ultimately be understood as a means for expressing what would otherwise require something like

[9] For present purposes, this is taken to be equivalent to
(ES*) That p is true iff p.

(5) For all p, if Isabel says that p, then p.

But (5) involves a special "non-objectual" kind of quantification: substitutional (or perhaps propositional, although, as we (2010) have noted, we suspect that the latter actually makes no sense). This involves providing items to fill in dummy-variables that occupy sentence positions. It generalizes on sentence-in-use positions to cover what would otherwise have to be expressed via a conjunction of conditionals like

(6) If Isabel says that crabapples are edible, then crabapples are edible, and if Isabel says that grass is green, then grass is green, and if Isabel says that power corrupts, then power corrupts, and if Isabel says that the moon is made of cheese, then the moon is made of cheese, and if Isabel says that the mass of the Earth is 47 million kilograms, then the mass of the earth is 47 million kilograms, and if Isabel says. . . .

Since (6) must go on to cover everything Isabel might possibly say, and since that is an infinite number of things, it is impossible for us to assert (6) directly. But we can, and do, express a commitment to what (6) would say, given an utterance of (3), together with the rules that govern truth-talk. Truth-talk thus provides a way for us to say what we want, finitely, and in an ordinary language like English.

As we see it (and as T-deflationists would agree), allowing us to generalize in this new way on sentence-in-use positions within claims, without having to incorporate new complicated logical devices into our language, is the main, perhaps the central, purpose of truth-talk. (Cf. Quine (1986) and Field (1994).) Our appeal to pretense explains how truth-talk does this with linguistic resources that seem, on the surface, unsuited to the task (without leaving it a brute, explained fact that it does).

Before moving on to highlight a particular virtue of our pretense account of truth-talk, we should address a question one might have. We have claimed that T-deflationists should be pretense theorists about truth-talk, but given what the pretense account says about the function and purpose of truth-talk, why not just endorse, and adopt, T-deflationism? Why bother also endorsing, or adopting, a *pretense* account of truth-talk? There are two answers to this question.

First, we think that such a question, while reasonable, belies a misunderstanding of what T-deflationism involves.[10] On our view, the pretense approach is correlated with the *genus* of T-deflationism as a whole. The different species of this genus (e.g., disquotationalism, prosententialism, inference-rule deflationism, etc.) should all be considered different attempts at cashing out principles of generation for a game of make-believe that could underwrite truth-talk.[11] Our main reason for claiming this is the recognition that a central thesis of T-deflationism is that truth-talk serves only logical and linguistic *expressive* purposes. The truth-locutions exist in order to

[10] For more on the details behind T-deflationism, see Armour-Garb (2012).

[11] For more on this view, see Armour-Garb and Woodbridge (2010, 2015). We should note that this understanding of the relationship between deflationism and pretense contrasts with that at work in Woodbridge (2005), where the pretense view is presented as a species of deflationism, in competition with other species.

provide a means for talking about other things, unrelated to truth. So T-deflationism takes the truth-locutions to be "representational aides", introduced, not to express something about the world directly, but rather in order to facilitate a certain kind of indirect talk about aspects of the world. Understanding a way of talking in the way T-deflationists view truth-talk just is to see it as involving a kind of pretense.

Second, as we see it, the appeal to pretense does a better job with the generalization problem that confronts T-deflationary accounts of truth-talk, since it makes the logic of 'true'-involving generalizations actual generalizations, logically speaking, instead of just the collection of the instances.[12] Moreover, on our view such generalizations already cover new cases that arise with the expansion of a language, without changing the meaning of 'true' when the substitution class for the otherwise necessary substitutional quantifiers or schematic sentence variables is changed. Thus, our view better accounts for the role that 'true'-involving generalizations can play in explanations, the expression of logical laws, etc.[13]

Having briefly sketched our pretense account of truth-talk, we will now turn to highlight a central virtue of our pretense account of truth-talk, viz., the solution that it offers to the Liar Paradox.

17.5 The Pretense View and the Liar Paradox

Starting from our pretense account of truth-talk, the approach to the Liar Paradox (and to semantic pathology generally) that we now favor is a version of the "meaningless strategy", according to which, in a certain sense, liar sentences turn out to lack content. This is because the pretense account of truth-talk sketched above has an interesting consequence for liar sentences and their kin: They do not specify any M-conditions.

Following the explanation of how the content of any instance of truth-talk is determined, we can see that in the case of a liar sentence, such as

(L) (L) is not true,

any M-conditions that (L) specified would have to be a function of the M-conditions specified by the content-vehicle that this instance of truth-talk putatively denotes. But in this case that is "another" instance of truth-talk (in fact, it is (L) itself). This means that, in order to determine the M-conditions that (L) would specify, we must look to what content-vehicle this "other" instance of truth-talk putatively denotes.

In some cases, this iterated process will "ground out," as it does for

(7) Sentence (1) is not true.

Any M-conditions specified by (7) indirectly would be a (negating) function of the M-conditions specified by (1). Since (1) is itself an instance of truth-talk, any specifying

[12] Cf. Gupta (1993).
[13] For more on this, see Armour-Garb and Woodbridge (2015).

of M-conditions that it accomplished would also happen only indirectly, as a function of the M-conditions specified by the content-vehicle that it putatively denotes—here, the M-conditions that (2) specifies directly. Thus, (7) indirectly specifies the M-conditions of crabapples not being edible.

Unlike in the (7)-(1)-(2) case, however, with a liar sentence like (L), this multi-step determination process repeats endlessly, with the result that (L) never manages to specify any M-conditions. We basically get content-determination instructions that can never be completed.[14] It is in this sense that a liar sentence like (L) can be said to be meaningless. As a result, we presently endorse a version of the meaningless strategy for dealing with the Liar Paradox.

17.5.1 Meaninglessness and Understanding

Any version of the meaningless strategy faces an immediate objection, which arises once we recognize that it seems unequivocal that, in some sense, we *understand* liar sentences. In response, we explain that we do not deny that we understand a liar sentence like (L), but it is important to note that we understand (L) only on a certain notion of understanding. Our claim is that there are (at least) two modes of understanding and that, while we understand (L) on one of them, we do not understand it on the other. Call the sense in which we do *not* understand (L), the sense that would require knowing what M-conditions (L) specifies, *understanding*$_1$. Call the sense in which we do understand (L) *understanding*$_2$.

With understanding$_2$ the mode of understanding that you possess is primarily *metalinguistic*: You know what the sentence appears to "say", for example, that it has a particular form, with a certain expression as the subject, in nominal position, employs a particular predicate, etc. And you know the meaning—the *character*, though not the *content*—of the expressions contained therein. Finally, you know how such a sentence could be used to make a meaningful assertion.

Now, our claim is that if you know the form of the sentence, the meanings of the words that are contained therein and how the sentence could be used to make a genuine assertion, then you can be said to "understand$_2$" the sentence.[15] But insofar as you do not know the M-conditions for some sentence, whether there are any or not, then, while you may—indeed, probably do—understand$_2$ the sentence, you do

[14] The same goes for wide-scope negation liar sentences, e.g., (L*) It is not the case that (L*) is not true.

[15] This allows for a possible contrast between Strawson's (1950) sentence, 'This is a fine red one' and Chomsky's (1957) sentence, 'Colorless green ideas sleep furiously'. The latter sentence is not even understood$_2$, if we insist that to understand$_2$ a sentence, we need to know how that sentence could be used to make a true assertion. We should note that, although we are inclined to accept this condition, we need not insist on it, for the points in this paper to go through.

not understand$_1$ that sentence, since it fails to specify M-conditions and, thus, is meaningless in the way that we have indicated.[16]

17.5.2 Semantic Characterization and S-Defectiveness

Suppose that we are right that liar sentences are meaningless in the sense we have explained and that they therefore cannot be understood$_1$. It follows from this that there is nothing that they express and, hence, nothing they put forward that anyone could accept or reject. Since affirming and denying can be understood as the speech acts that express the mental attitudes of acceptance and rejection (respectively), it thus to follow that, on our view, one can neither affirm nor deny liar sentences. But if one cannot do this (and in particular, do it indirectly by assigning liar sentences truth-values), then we still face the familiar question of how we will characterize such sentences. And, as is familiar from attempted consistent solutions to the liar, it is at this point that revenge problems generally emerge.

We believe that we can address these issues and avoid the usual problems to which they appear to give rise. Although we shall only have space here to sketch a way of dealing with them, our hope is that what we will provide will be sufficient.[17]

We avoid the "first wave" of revenge problems because we take no positive or negative attitude towards liar sentences, and we neither reason to or from them, or evaluate them semantically—in the sense of ascribing them either a *truth-value* or a *logical value* (e.g., 1 or 0). On our pretense account of truth-talk, liar sentences do not admit of these sorts of evaluation. Indeed, because no M-conditions prescriptive for the pretenses displayed in any of these evaluations are ever determined (since (L) itself specifies no M-conditions), there are no M-conditions that would make it correct to utter "(L) is not true", no M-conditions that would make it correct to utter "(L) is true", and, for similar reasons, no M-conditions that would make uttering "(L) is false" correct.

Keeping in mind that liar sentences (and their kin) cannot, in the relevant sense, be understood$_1$ and, thus, cannot be evaluated in the standard ways, what can we say about them? More directly, how will we characterize them? We propose the following.

As a means for characterizing liar sentences, we introduce a new predicate, 'is semantically defective' (henceforth, 's-defective'), which, for present purposes, is to apply to those sentences, which, while perhaps understood$_2$, have no content. More specifically, we are inclined to claim the following, by way of clarifying 's-defective':

---

[16] Analogous to the two modes of understanding, we might also grant two modes of *meaning*. As we would describe things, since liar sentences do not specify any M-conditions, we will not grant that they are meaningful$_1$, though we will, and should, allow that they are meaningful$_2$. For more on this, see Armour-Garb and Woodbridge (2013, 2015).

[17] Indeed, for a consideration of some of the other objections that our solution to the Liar Paradox faces, see Armour-Garb and Woodbridge (2013, 2015).

(i) If a sentence, S, is s-defective, then it has nothing, by way of content, which we can accept or reject.
As a result,
(ii) If S is s-defective, then S is not understood$_1$.
Moreover,
(iii) If S fails to specify any M-conditions—either directly or indirectly—then it is appropriate to attribute *s-defectiveness* to S.
In addition,
(iv) If S is s-defective, then, since S will not be understood$_1$, it is not aletheically evaluable.
If S is not alethically evaluable, it cannot (correctly) be assigned a truth-value.
And, more generally,
(v) S is s-defective only if it is not truth-apt.[18]

By way of shedding further light on the term 's-defective' that is relevant to the Liar Paradox, we might say that for a given sentence, S, S is s-defective at least under the following condition: the process that would determine what M-conditions S specifies never finishes in the case of S, and, thus, S does not specify any M-conditions at all. Of course, this does not count as an *analysis* of the notion of s-defectiveness, as it leaves open the possibility that there are other ways in which a sentence may be deemed s-defective (for example, Strawson's 'This is a fine red one', which includes a demonstrative with no demonstratum), but it will do, for what follows. Let us now apply this approach to liar sentences.

17.5.3 S-Defectiveness and Liars

Consider (L), once again. As we saw, (L) does not specify any M-conditions, which means that, by (iii), (L) is s-defective, in which case

(8) (L) is s-defective

will be true, and, thus, given the relevant identity,

(9)'(L) is not true' is s-defective.

will also be true.

[18] To be sure, there is more that we might say about this notion of *s-defectiveness*, which we are importing into our vocabulary. But what is crucial here is that 's-defective' applies directly to sentences—actually, to sentence tokens, though the view will not end up looking like a tokenist view, at least in any interesting sense—rather than applying to what a given sentence expresses or applying to a sentence in virtue of applying to what it expresses. Moreover, the expression, 's-defective', applies to sentences that do not possess content, even though such sentences will (or, at least, may) be understood$_2$.

The pressing issue is whether our characterization of (L) as s-defective, and the correctness of ascribing truth to a statement of that characterization, generates revenge problems for us. Let us turn, then, to a familiar sort of revenge problem that one might pose for our present proposal. Such a case is found in the sentence

(λ) (λ) is not true or (λ) is s-defective.

In order to see the problem that (λ) appears to present, suppose that (λ) is any of true, false, or not true. For (λ) to be true it must either be not true or be s-defective—both of which are inconsistent with its being true. It is obvious that inconsistency results, if the left-hand disjunct is true. If the right-hand disjunct is true, then, by (ii), (λ) is not understood$_1$. But if (λ) is true then it is understood$_1$. Contradiction. But since a disjunction is true if, and only if, at least one of the disjuncts are true, it follows that (λ) cannot be evaluated as true.

For (λ) to be false, it must fail to be s-defective. But it must also fail to be not true, meaning that, if (λ) were taken to be false, then, *via* some innocuous reasoning, it would seem to follow that it is true, which, again, issues in inconsistency. Similar consequences arise, if we hold that (λ) is not true. After all, if (λ) is not true, then, by enquotation, '(λ) is not true' will be true, and it will follow that '(λ) is not true or (λ) is s-defective' is true. But this mentioned sentence just is (λ), so that would just mean that (λ) is true. Thus, if (λ) is not true, it follows that (λ) is true.

Now, we characterize (λ) as s-defective. But because we make this characterization, further paradox appears immanent. For if we maintain that (λ) is s-defective, then we will also accept that '(λ) is s-defective' is true. But now, by disquotation, or-introduction, and enquotation, we seem to be committed to the truth of '(λ) is not true or (λ) is s-defective', from whence, as we have seen, inconsistency appears to be unavoidable. So, *prima facie*, we too appear to be mired in paradox, having attributed s-defectiveness to (λ).

To be sure, it certainly *seems* that we are mired in paradox. But, in fact, we are not. We would be in trouble if we wanted to hold onto classical logic *and* were to grant that (λ) is a contentful sentence, for then we would seem to be compelled to grant that (λ) is true. But, while we may retain classical logic, on our view, paradox is avoided in the case of (λ), in virtue of the fact that it does not possess any content.

Actually, as we will see, our argument for the claim that, in this case, paradox is avoided, relies on two features, each of which we will motivate. The first is that (λ) is without content. The second is that if a standard, aletheically evaluable sentence is disjoined (or conjoined or otherwise extensionally connected) with a sentence that is without content, then contentfulness cannot be preserved in the resulting complex sentence. We will begin with the first feature, regarding the contentlessness of (λ). We will then turn to the second.

In order for our attribution of s-defectiveness to (λ) to generate paradox, (λ) would have to have content, in the sense of specifying M-conditions. But it does not have content, and here is why. For any content that (λ) would have, both disjuncts are relevant and would have to contribute. This is so because the meaning of a disjunction is a function of the meanings of its parts. So, the meaning—and, thus, the meaningfulness—of (λ) relies, at least in part, on that of its disjuncts. If one of

the disjuncts lacks content, then (λ) itself does too. Accordingly, we will show that (λ) lacks content, by explaining why one of its disjuncts lacks content, where, recall, a given sentence lacks content if it fails to specify M-conditions. In particular, we will show that the left-hand disjunct in (λ), viz., '(λ) is not true', fails to specify M-conditions and, thus, lacks content.

Recall that any M-conditions specified by the left-hand disjunct of (λ) would, as with any instance of truth-talk, have to be a product of M-conditions specified by the putative content-vehicle that the disjunct denotes. However, the supposed content-vehicle that (λ)'s left-hand disjunct denotes is just the whole of (λ) itself. This means that what is relevant to determining M-conditions for the left-hand disjunct just is the M-conditions specified by (λ) as a whole. But the left-hand disjunct is part of (λ), so determining what M-conditions (λ) specifies requires determining the M-conditions that '(λ) is not true' specifies. But that, in turn, requires that we determine what M-conditions (λ) specifies. Accordingly, in order for the left-hand disjunct of (λ) to specify M-conditions, it is required that (λ) already has determined M-conditions. But, of course, M-conditions cannot be settled for (λ) unless, or until, M-conditions are determined for its left-hand disjunct.

So, for any overall M-conditions to get specified by (λ), there would have to be an impossible sort of semantic bootstrapping, which means that the process for determining what M-conditions (λ) specifies never finishes. Since (λ) fails to specify M-conditions, it follows that the left-hand disjunct does not possess any content, and, so, neither does (λ) itself. As such, both (λ) and its left-hand disjunct lack content. (Notice, though, that both are understood$_2$.)

As is evident, our response to the revenge argument relies on a premise, to the effect that only contentful sentences may be disjoined with other contentful sentences to yield a disjunction that is, itself, contentful and, thus, alethically evaluable. We will now provide some support for this premise.

Although a conjunction gets its logical value from its conjunctive parts and a disjunction gets its logical value from at least one of its disjunctive parts, both conjunctions and disjunctions get their *content* from *both* of their respective parts. This is so because the content of a complex sentence—viz., a conjunction, disjunction, etc.—is a function from the contents of its parts. But if the content of a complex sentence is a function from that of its parts, then, if any part of a complex sentence lacks content, the same is true of the sentence as a whole. What this means is that the M-conditions for a disjunctive sentence will be a function of the M-conditions for each of its disjuncts. So, if one of the disjuncts of (λ) lacks M-conditions, then (λ) itself lacks M-conditions.

We have explained why (λ) specifies no M-conditions and thus has no content. This means that any instance of truth-talk (positive or negative) in which (λ) is the supposed content-vehicle putatively denoted will likewise have no content. Since, as a result, the sentence '(λ) is not true' has no content, disjoining it to another sentence yields a disjunctive string with no content. So, even though '(λ) is s-defective' has content and is true, disjoining this sentence with '(λ) is not true', in order to form (λ) itself, yields a sentence that has no content and, thus, is not alethically evaluable.

This has an important consequence for the aforementioned revenge argument. Recall that the revenge argument began by allowing that if '(λ) is s-defective' is true then '(λ) is not true or (λ) is s-defective' will likewise be true, from which paradox appears to follow. But we have argued that, while we grant that the sentence '(λ) is s-defective' is true, one cannot disjoin that sentence with '(λ) is not true' and preserve meaningfulness. In this way, we avoid the putative revenge problem on the table, because (λ)'s lack of content blocks the argument that the would-be revenge requires. In the terminology that we favor, because (λ) fails to specify M-conditions, we claim, by (iii), that (λ) is s-defective. Thus, the revenge argument cannot bootstrap (λ) into contentfulness and thereby make it evaluable as true or false (or even as not true or not false).

17.6 Conclusions

In this paper, after briefly explaining why T-deflationists should be fictionalists about truth-talk, we introduced our preferred fictionalist approach and sketched a semantic pretense account of truth-talk. We then showed how this account provides a diagnosis of the problem with liar sentences that enables us to provide a consistent solution to the Liar Paradox. In closing, we should note that the very same considerations that we employed, in order to "solve" the Liar Paradox, also applies to other cases of apparent semantic pathology.

For example, consider the truth-teller, which is exhibited by a sentence like

(K) (K) is true.

This sentence appears to manifest a kind of indeterminacy, since it seems that (K) can be assigned either the value 'true' or the value 'false', but there seems to be no reason for assigning it one over the other, and thus no fact of the matter as to which value it does, or should, possess. While (K) may provide problems for some extant solutions to the semantic paradoxes,[19] it poses no problem for our proposed account. To see this, notice that, like (L) (and for essentially the same reasons), (K) specifies no M-conditions and, thus, will likewise be characterized as s-defective.

It is also worth noting that, as we can show, the same holds for another host of cases that appear to manifest semantic pathology. We think, for example, of Curry's Paradox, which arises given a sentence like

(C) If (C) is true, then every sentence token is true.

as well as for the "dual symptom" cases that elsewhere (2006 and 2008) we have called *open pairs*, for example

(A) Sentence (B) is not true.
(B) Sentence (A) is not true.

[19] See Armour-Garb and Woodbridge (2006).

This reveals a further merit of the semantic pretense account of truth-talk. It not only provides a way of dealing with apparently inconsistent cases of semantic pathology, such as the Liar Paradox and Curry's Paradox; it also underwrites a unified diagnosis and treatment of apparent semantic pathology more generally, including the putatively indeterminate and dual-symptom varieties. In so doing, it offers a general treatment of the cases of apparent semantic pathology. This is a virtue of the semantic pretense version of PIF that we favor, the version of fictionalism about truth-talk that truth-theorists (as good T-deflationists) should endorse and adopt.

References

Armour-Garb, B. (2012). Deflationism (about theories of truth). *Philosophical Compass, 7,* 267–277.

Armour-Garb, B., & Woodbridge, J. (2006). Dialetheism, semantic pathology, and the open pair. *Australasian Journal of Philosophy, 84,* 395–416.

Armour-Garb, B., & Woodbridge, J. (2010). Why deflationists should be pretense theorists (and perhaps already are). In N. Pedersen & C. Wright (Eds.), *New waves in truth* (pp. 59–77). Basingstoke: Palgrave Macmillan.

Armour-Garb, B., & Woodbridge, J. (2012). The story about propositions. *Noûs, 46,* 635–674.

Armour-Garb, B., & Woodbridge, J. (2013). Semantic defectiveness and the liar. *Philosophical Studies, 164,* 845–863.

Armour-Garb, B., & Woodbridge, J. (2014). From mathematical fictionalism to truth-theoretic fictionalism. *Philosophy and Phenomenological Research, 88,* 93–118.

Armour-Garb, B., & Woodbridge, J. (2015). *Pretense and pathology: philosophical fictionalism and its applications*. Cambridge: Cambridge University Press, (forthcoming).

Brandom, R. (1994). *Making it explicit*. Cambridge: Harvard University Press.

Chomsky, N. (1957). *Syntactic structures*. The Hague: Mouton de Gruyter.

Crimmins, M. (1998). Hesperus and phosphorus: Sense, pretense, and reference. *Philosophical Review, 107,* 1–47.

Evans, G. (1982). *The varieties of reference*.(J. McDowell (Ed.)) Oxford: Oxford University Press.

Field, H. (1994). Deflationist views of meaning and content. *Mind, 103,* 249–285.

Field, H. (2008). *Saving truth from paradox*. Oxford: Oxford University Press.

Gupta, A. (1993). A critique of deflationism. *Philosophical Topics, 21,* 57–81.

Quine, W. (1986). *Philosophy of logic* (2nd ed.). Cambridge: Harvard University Press.

Richard, M. (2000). Semantic Pretense In A. Everett & T. Hofweber (Eds.) *Empty names, fiction, and the puzzles of non-existence* (pp. 205–232). Stanford: CSLI Publications.

Strawson, P. (1950). On referring. *Mind, 59,* 320–344.

Walton, K. (1990). *Mimesis as make-believe*. Cambridge: Harvard University Press.

Walton, K. (1993). Metaphor and prop-oriented make-believe. *European Journal of Philosophy, 1,* 39–56.

Woodbridge, J. (2005). Truth as a pretense. In M. Kalderon (Ed.) *Fictionalism in metaphysics* (pp. 134–177). Oxford: Oxford University Press.

Woodbridge, J., & Armour-Garb, B. (2008). The pathology of validity. *Synthese, 160,* 63–74.

Woodbridge, J., & Armour-Garb, B. (2009). Linguistic puzzles and semantic pretense. In S. Sawyer (Ed.), *New waves in philosophy of language*. Basingstoke: Palgrave Macmillan.

Yablo, S. (2001). Go figure: A path through fictionalism. *Midwest studies in philosophy*. (XXV: Figurative Language), 72–102.

Yablo, S. (2005). The myth of the seven. In M. Kalderon (Ed.) *Fictionalism in metaphysics* (pp. 88–115). Oxford: Oxford University Press.

Chapter 18
Groundedness, Truth and Dependence

Denis Bonnay and Floris Tijmen van Vugt

Abstract Leitgeb (2005) proposes a new approach to semantic paradoxes, based upon a direct definition of the set of grounded sentences in terms of dependence upon non-semantic state of affairs. In the present paper, we account for the extensional disagreement between this dependence approach and more familiar alethic approaches. In order to do so, we study the behavior of dependence jumps and alethic jumps, and provide an equivalence result for the two approaches.

18.1 Immunity to the Semantic Paradoxes

Truth may yield paradoxes, and Tarski showed how widespread the problem is by proving that no sufficiently rich classical language can contain its own truth definition on pain of inconsistency. In order to tackle the semantic paradoxes, one strategy is obviously to try and understand which features of sentences are responsible for the reported misbehavior. Studies of the role played by self-reference (and of phenomena going beyond self-reference, see Yablo 1993) follow this strategy. But, alternatively, one might want to look at the good instead of the bad. Some sentences, including sentences containing the truth predicate, are clearly unproblematic: why is it that they are immune to the semantic paradoxes?

Following a widespread intuition, unproblematic sentences are such that their value ultimately depends on what the world is like. The truth-value of (1) "Emily

Both authors contributed equally to this work and are listed in alphabetical order.

D. Bonnay
Université Paris Ouest, IRePh & IHPST, Nanterre, France
e-mail: denis.bonnay@u-paris10.fr; denis.bonnay@gmail.com

F. Tijmen van Vugt
University of Music, Drama and Media, Hannover, Germany

IMMM, Hannover, Germany

Lyon Neuroscience Research Center, Université Claude Bernard Lyon-1,
Villeurbanne, France
e-mail: f.t.vanVugt@gmail.com

T. Achourioti et al. (Eds.), *Unifying the Philosophy of Truth,* Logic, Epistemology,
and the Unity of Science 36, DOI 10.1007/978-94-017-9673-6_18

Dickinson was born in Amherst" directly depends on where Emily Dickinson was born. When evaluating (2) "The sentence 'Emily Dickinson was born in Amherst' is true", we are led again to evaluate (1). Hence the truth-value of (2) depends, albeit indirectly, on where Emily Dickinson was born, and only on that fact. The same holds for (3) "Either Emily Dickinson is an American poet or (2) is false". No matter how many nested instances of the truth predicate there are, as long as semantic evaluation percolates down, sentences such as (1), (2) and (3) are clearly *grounded* in worldly facts. Rougly, a sentence is considered grounded if its truth value can be established regardless of the truth or falsity of other sentences, or if its truth value can be established based on the truth or falsity of other sentences that are, themselves, grounded.

However, demarcating a class of grounded sentences is not so easy, and some cases are far less clear-cut. What about a sentence such as (4) "Emily Dickinson was born in Amherst or λ" where λ is the liar sentence ? The fact that Emily Dickinson was actually born in Amherst can be construed as sufficient to ensure that (4) will have truth-value true. But on the other hand, that (4) is grounded is contingent on the truth of (1): if Emily Dickinson had been born near Bradford in Yorkshire then the truth of (4) would depend on λ, making it ungrounded. (5) "λ is true or λ is not true" might also be considered as grounded in a classical framework: any sentence of the form $Px \vee \neg Px$ is true by pure logic. Even though (5) would not be grounded in empirical facts, its truth would be secured by logical laws. However, within a different logical framework, say a multi-valued logic, the truth-value of (5) does depend on the truth-value of λ, so that (5) is deemed to be ungrounded.

Despite these difficulties, there are reasons to think that analyzing groundedness is a valuable take on phenomena surrounding semantic paradoxes. First, the class of unproblematic sentences may constitute a more natural kind than the class of problematic sentences. There is more than one way for a sentence to be pathological: consider, for example, the distinction between the straightforwardly paradoxical liar sentence and the non-paradoxical but still pathological truth-teller. The class of grounded sentences might prove to be more homogenous, or at least easier to chart. Second, groundedness is based on a more general notion of dependence, which may provide us with more robust intuitions: analyzing groundedness amounts to analyzing what a particular kind of dependence amounts to.

Two prominent accounts of groundedness have been given by (Kripke 1975) and (Leitgeb 2005). Their approaches to groundedness both seem to rely on the intuition we gave earlier: a sentence is grounded if semantic evaluation eventually percolates down. However, they differ in the way this intuition is captured. The Kripkean strategy, which may be made to work using various semantic backgrounds, consists in gradually defining the extension and anti-extension of the truth-predicate. A sentence is grounded if it is true or false in the least fixed point for the corresponding jump operation. Leitgeb rather proposes to attack groundedness head-on. The idea is to gradually define the set of sentences which depend on non-semantic state of affairs, albeit more and more indirectly. The set of grounded sentences is then again defined as a least fixed point. Kripke and Leitgeb's account are not extensionally equivalent. Some sentences are grounded according to Kripke but not according to Leitgeb and

the other way around. What is revealed by this discrepancy? Does the direct approach in terms of dependence reveal some new intuitions regarding groundedness? Alternatively, it could be that the definition of groundedness is sensitive to some exogenous parameters, as was suggested in the discussion of (5) above, that would account for superficial differences in output. Finally, the right diagnosis could be a mix of the previous two.

Our aim in this paper will be to identify the endogenous or exogenous parameters which may result in different outputs for the fixed point constructions. We shall thus work at bridging the extensional gap between Kripke's and Leitgeb's construction, while assessing the novelty of Leitgeb's approach.

18.2 The Indirect Route to Groundedness

The characterization of grounded sentences is a by-product of Kripke's construction. Kripke defines a gradual process of determination of the extension and anti-extension of the truth-predicate, which is intended to mimic the process by which someone who does not understand the word 'true' gradually comes to understand it. Initially, the extension and anti-extension only contain sentences without nested occurrences of the truth-predicate. These sentences get assigned to the extension or anti-extension of the truth predicate as soon as the subject is told that "we are entitled to assert (or deny) of any sentence that it is true precisely under the circumstances when we can assert the sentence itself" (1975, p. 701). At each new step in the process, the subject is willing to put in the extension or the anti-extension of the truth-predicate the sentences that were true or false at the previous step. Eventually, this process becomes 'saturated'. This corresponds to a stage at which the extension and anti-extension of the truth predicate have reached a stable state, i.e. a fixed point. A sentence is grounded according to Kripke if it is true or false in the least fixed point; that is, if it got its value determined in this process of gradual construction of the extension and anti-extension of the truth-predicate.

Let us investigate this formally. Given a classical language $\mathcal{L}$, rich enough to allow for its own syntax to be expressed in it, let $\mathcal{M}$ be a model for $\mathcal{L}$ with domain D. Tr will be the truth predicate, to be interpreted by means of an extension E and an anti-extension A. In general, the interpretation of the truth predicate will need to be non-classical, that is, some sentence codes will be neither in E nor in A. Given (E, A), one may compute a new interpretation for Tr by means of a jump operator J^K, where K stands for Kripke:

$$J^K(E, A) = \{\phi | \mathcal{M}, (E, A) \models^K \phi\}$$

where $\models^K$ is truth using the strong Kleene scheme (so that ^{K} in J^K could also stand for Kleene). Given any set $E \subseteq \mathcal{L}_{\mathrm{Tr}}$ (where $\mathcal{L}_{\mathrm{Tr}}$ is the language including the truth-predicate) a "set of negatives" is defined: $\neg E \stackrel{\mathrm{def}}{=} \{\phi | \neg\phi \in E\}$. It is not necessary to separately define a 'negative' jump yielding the new anti-extension of Tr, because

that new anti-extension is simply $\neg J^K(E, A)$. In what follows, we shall pretend for simplicity that J^K has only E as an argument, A being assumed without loss of generality to be always $\neg E$.

Generalizing the above procedure, we can build an ordinal sequence $(E^K_\alpha)_{\alpha\in\mathrm{On}}$ as follows: $E^K_0 = \emptyset$, $E^K_{\alpha+1} = J^K(E^K_\alpha)$ and $E^K_\beta = \bigcup_{\alpha<\beta} E^K_\alpha$ for β a limit ordinal. The monotonicity of the sequence $(E^K_\alpha)_{\alpha\in\mathrm{On}}$ together with the fact that the class of sentences is a set yields that there is a fixed point, hence a smallest one (Tarski 1955), which we will call E^K_∞. A sentence ϕ of $\mathcal{L}_{\mathrm{Tr}}$ is defined to be *K–grounded* if it is true or false in that smallest fixed point, i.e. ϕ is grounded according to Kripke (and Kleene) iff $\phi \in E^K_\infty \cup \neg E^K_\infty$.

As Kripke had underlined, the construction is not tied to the strong Kleene scheme as semantic background. Any monotonic jump operator would do. For future reference, we shall now recall the definitions of two of them. Let $\mathrm{Val}_\Phi\phi$ be the truth-value of ϕ in the *fully classical* model $\mathcal{M}, \Phi$ for $\mathcal{L}_{\mathrm{Tr}}$, which is $\mathcal{M}$ expanded with Φ as the extension of Tr (no anti-extension here: the expanded model is classical). We give the definition of the supervaluational jump J^V, which uses van Fraasen's supervaluations as semantic background:

$$J^V(E) = \{\phi | \text{ for all } \Phi \text{ s.t. } E \subseteq \Phi \text{ and } \Phi \cap \neg E = \emptyset, Val_\Phi\phi = 1\}$$

The supervaluation jump requires that $\mathrm{Val}_E\phi = 1$, just like J^K did. In a classical setting, this is not enough, because the jump would not be monotonic. So it is also required that ϕ is true for all interpretations of Tr that respect what is presently known about it (which is that sentences in E are true, and sentences in $\neg E$ are not true).

For future reference again, here is finally a variation on van Fraassen's jump due to Cantini (1990). Cantini's jump J^C is defined by

$$J^C(E) = \{\phi | \text{ for all consistent } \Phi \text{ s.t. } E \subseteq \Phi, \mathrm{Val}_\Phi\phi = 1\}$$

where Φ is consistent if it does not contain the codes of a sentence and its negation.

Our two supervaluational jumps give rise to ordinal sequences (E^V_α) and (E^C_α) analogous to (E^K_α). Since J^V and J^C are monotonic, there are least fixed points E^V_∞ and E^C_∞. A sentence ϕ of $\mathcal{L}_{\mathrm{Tr}}$ is *V–grounded* if $\phi \in E^V_\infty \cup \neg E^V_\infty$. Similarly, it is *C–grounded* if $\phi \in E^C_\infty \cup \neg E^C_\infty$. Thus, the indirect route gives rise to different sets of grounded sentences. The differences between these are extrinsic. Groundedness is captured in the same way, relying on the same intuition of a gradual fixation of truth-values for sentences in which the truth predicate occur. The sets of grounded sentences differ only because different jumps are used, that is because different semantic backgrounds (Kleene logic or supervaluations) are used.

18.3 The Direct Route

Kripke arrives at his definition of groundedness only after he has defined the truth predicate: it is in no way implicated in or required during the construction. By contrast, groundedness is the starting point of Leitgeb's approach, the reason being that Leitgeb's aim is to isolate a class of sentences which can safely occur in the

T-scheme.[1] He defines a notion of dependence, according to which the truth-value of a sentence ϕ is said to depend on the extension of the truth-predicate for a set of sentences Φ. As the basic case, there are the sentences in which the truth predicate does not occur: they do not depend on other sentences at all (but only directly on non-semantic states of affairs). A process somewhat parallel to the previous one gradually defines larger and larger sets of sentences which depend on non-semantic states of affairs, albeit more and more indirectly. Again, this process eventually becomes saturated, and this corresponds to a minimal fixed point gathering all sentences which are grounded according to Leitgeb. In that sense, Leitgeb aims at treating groundedness as a primitive notion: it is not to be derived from being true or false in some condition, being grounded just means depending, directly or indirectly, on non-semantic state of affairs.

A relation of *dependence* between a sentence ϕ and a set of sentences Φ will be defined. Intuitively, the idea is that ϕ depends on Φ if there is no change in the truth value of ϕ without a change in which of the sentences in Φ are considered to be true. For instance, the truth-value of $\mathrm{Tr}\ulcorner 2+2=4\urcorner$ supervenes on the truth-value of $2+2=4$. Thus, $\mathrm{Tr}\ulcorner 2+2=4\urcorner$ depends on $\{2+2=4\}$. By contrast, $\mathrm{Tr}\ulcorner 2+2=4\urcorner \vee \neg\mathrm{Tr}\ulcorner 2+2=4\urcorner$ will be true regardless of whether $\ulcorner 2+2=4\urcorner$ belongs to the extension of the truth predicate, or any other sentence for that matter. Therefore, one can say it depends on no sentences at all.[2]

We shall now define this idea formally: ϕ depends on $\Phi \subseteq \mathcal{L}_{\mathrm{Tr}}$ iff for all $\Psi_1, \Psi_2 \subseteq \mathcal{L}_{\mathrm{Tr}}$, we have that if $\mathrm{Val}_{\Psi_1}\phi \neq \mathrm{Val}_{\Psi_2}\phi$ then $\Psi_1 \cap \Phi \neq \Psi_2 \cap \Phi$. As shown in (Leitgeb 2005), $D(\phi) \overset{\text{def}}{=} \{\Phi \subseteq \mathcal{L}_{\mathrm{Tr}} | \phi \text{ depends on } \Phi\}$ is a filter. If $D(\phi)$ has a least element Φ, we say ϕ depends *essentially* on Φ. The definition of the set of grounded sentences is now based on a dependence jump, as opposed to the alethic jumps presented in the previous section. The dependence jump D^{-1} is defined by:

$$D^{-1}(E) = \{\phi \in \mathcal{L}_{\mathrm{Tr}} | \phi \text{ depends on } E\}$$

Just as in the case of alethic jumps, an ordinal sequence $(\Phi_\alpha)_{\alpha \in \mathrm{On}}$ is defined: $\Phi_0 = \emptyset$, $\Phi_{\alpha+1} = D^{-1}(\Phi_\alpha)$ and $\Phi_\beta = \bigcup_{\alpha<\beta}\Phi_\alpha$. D^{-1} is monotone so, again, there is a least fixed point of this sequence that we will call Φ_{lf}. A sentence ϕ is *L-grounded* iff it belongs to this least fixed point, i.e. iff $\phi \in \Phi_{\mathrm{lf}}$.[3]

---

[1] There is no way to do without a principled criterion that decides which T–equivalences hold. Looking just at maximal coherent classes will not do. McGee has shown that taking simply this restriction, the kinds of candidate sets is virtually unrestricted (McGee 1992)

[2] Interestingly, the same is not true of $\mathrm{Tr}\ulcorner 2+2=4\urcorner \vee \mathrm{Tr}\ulcorner \neg 2+2=4\urcorner$ since a priori both $2+2=4$ and $2+2 \neq 4$ may at the same time be absent from the truth predicate.

[3] Leitgeb shows that this definition confirms our intuition that groundedness means referring (in)directly to the world, or, in his terminology, non–semantic states of affairs: ϕ is ungrounded if, but not only if, there exists a sequence $(\psi_n)_{n\in\mathbb{N}^*}$ with $\psi_n \in \mathcal{L}_{\mathrm{Tr}}$; $\psi_0 = \phi$ and for every $n \in \mathbb{N}$, there is a set Ψ_{n+1} such that ψ_n depends on Ψ_{n+1} essentially and $\psi_{n+1} \in \Psi_{n+1}$ (Leitgeb 2005, p. 169).

Clearly, the sequence $(\Phi_\alpha)_{\alpha\in\mathrm{On}}$ is at every step a combination of true and false sentences taken together; for instance $\ulcorner 2+2=4\urcorner \in \Phi_1$ but equally $\ulcorner 2+2\neq 4\urcorner \in \Phi_1$. A derived sequence containing only those sentences that are true may defined: $\Gamma_0 = \emptyset$, $\Gamma_{\alpha+1} = \{\phi \in \Phi_{\alpha+1} | \mathrm{Val}_{\Gamma_\alpha}\phi = 1\}$, and $\Gamma_\beta = \bigcup_{\alpha<\beta}\Gamma_\alpha$ (Leitgeb 2005, p. 171). At this point, we have obtained a solution to the problem that paradoxes pose to the definition of a truth predicate, because in the fixed point, T-equivalences hold for all sentences which are grounded (that is, belong to Φ_{lf}, see Leitgeb 2005, Theorem 17).

18.4 Parting of the Ways

The set of K–grounded sentences and the set of L–grounded sentences are incomparable: neither one includes the other (Leitgeb 2005, p. 185). First, $\theta \stackrel{\mathrm{def}}{=} \mathrm{Tr}\ulcorner\lambda\urcorner \vee \neg\mathrm{Tr}\ulcorner\lambda\urcorner$ is grounded according to Leitgeb but not according to Kripke. Clearly θ is grounded (and true) in Leitgeb's scheme, since for any predicate P in a two–valued scheme $Px \vee \neg Px$ holds, hence for any Φ we have $\mathrm{Val}_\Phi\theta = 1$. Therefore θ depends on $\emptyset$, which implies that it belongs to Φ_{lf}. Kripke's approach with the strong Kleene scheme fails to make it grounded, since λ cannot be in the extension or the anti-extension of Tr in a fixed point.

Second, $\theta' \stackrel{\mathrm{def}}{=} \mathrm{Tr}\ulcorner 2+2=4\urcorner \vee \lambda$ is grounded according to Kripke but not according to Leitgeb. K–groundedness of θ' is due to the use of a *strong* Kleene scheme.[4] But now θ' does not belong to Φ_{lf}. The point is that θ' depends on $\{2+2=4, \lambda\}$ and essentially so: clearly $\mathrm{Val}_{\Phi\cap\{2+2=4,\lambda\}}\theta' = \mathrm{Val}_\Phi\theta'$ for any Φ and essentiality follows from the fact that if θ' depends on Φ, then $\ulcorner 2+2=4\urcorner \in \Phi$ and similarly $\lambda \in \Phi$. But since $\lambda \notin \Phi_{\mathrm{lf}}$, θ' does not depend on Φ_{lf} and is hence ungrounded.

It seems that the proposals of Kripke and Leitgeb simply disagree about the groundedness of θ because the former operates in three–valued and the latter in two–valued logic. Arguably, the question whether truth is bi– or trivalent is orthogonal to the question of what groundedness is. In any case, it is expected that using J^V or J^C instead of J^K will bring us closer to L–groundedness. However, what about sentences such as θ'? What are exactly the factors that are responsible for the discrepancy between $E^K_\infty \cup \neg E^K_\infty$ and Φ_{lf}? We shall turn to this question in the next section.

18.5 An Alethic Path Along the Dependence Route

In the spirit of Leitgeb's paper, we have taken for granted that the indirect and direct approaches are substantially different: the Kripkean one is indirect and recursively defines truth, Leitgeb's is direct and recursively defines groundedness. However,

[4] Note that using a weak Kleene scheme would not make things really better, since it would not ground $2+2=4 \vee \lambda$, which is L–grounded. More on this below.

calling this difference 'substantial' may be a bit of an overstatement. In order to compare not just the end product, but the underlying mechanisms, it is useful to ask whether the dependence approach can be phrased in alethic terms, that is whether it can be defined along the more familiar route of a recursive definition of truth.

We shall say that two candidate extensions Φ_1 and Φ_2 agree on a set E (notation: $\Phi_1 =_E \Phi_2$) iff $\Phi_1 \cap (E \cup \neg E) = \Phi_2 \cap (E \cup \neg E)$. Leitgeb's alethic jump J^L is defined by

$$J^L(E) = \{\phi | \text{ for all } \Phi_1, \Phi_2 \text{ with } \Phi_1 =_E \Phi_2, \text{Val}_{\Phi_1}\phi = \text{Val}_{\Phi_2}\phi \text{ and } \text{Val}_E\phi = 1\}$$

As before, J^L gives rise to an ordinal sequence $E_0 = \emptyset$, $E^L_{\alpha+1} = J^L(E_\alpha)$, and $E^L_\beta = \bigcup_{\alpha<\beta} E_\alpha$ for β a limit ordinal. J^L is monotonic, so there is a least fixed point E^L_∞, and the set of sentences that are grounded according to J^L is $E^L_\infty \cup \neg E^L_\infty$. Concerning groundedness, this is nothing but an alternative phrasing of Leitgeb's original definition. A sentence is grounded according to Leitgeb's alethic jump iff it is L–grounded, as we shall now see.

Actually, *at each stage* in the recursion, Leitgeb's alethic jump J^L and the dependence operator D^{-1} have similar effects, the difference being that J^L deals with sentences that have been determined to be true when D^{-1} deals with sentences that have been determined to be true or whose negation has been determined to be true.

Lemma 1 $J^L(E) \cup \neg J^L(E) = D^{-1}(E \cup \neg E)$.

Proof The condition on Φ_1 and Φ_2 in the definition of $J^L(E)$ is the contrapositive of the definition of D^{-1} applied to $E \cup \neg E$, except for the extra requirement that $\text{Val}_E\phi = 1$. This implies the direction from left to right, given that for any F, if $\phi \in D^{-1}(F)$, $\neg\phi \in D^{-1}(F)$ as well. From right to left, sentences in $D^{-1}(E \cup \neg E)$ are either true or false when E is the extension of Tr. If they are true, they end up in $J^L(E)$, and if they are false in $\neg J^L(E)$. □

Equality between the alethic fixed point and the dependence fixed point follows.

Proposition 2 $E^L_\infty \cup \neg E^L_\infty$ *is equal to* Φ_{lf}.

Proof By Lemma 1 and transfinite induction, $E^L_\alpha \cup \neg E^L_\alpha = \Phi_\alpha$ for any α. It is then sufficient to show that E_α is a fixed point for E^L iff Φ_α is a fixed point for D^{-1}. By Lemma 1 alone, it is obvious that a fixed point for J^L is also a fixed point for D^{-1}. We need to show the other direction. Let γ be an ordinal at which a fixed point for D^{-1} is reached. By Lemma 1 again, $E^L_\gamma \cup \neg E^L_\gamma = E^L_{\gamma+1} \cup \neg E^L_{\gamma+1}$. Since J^L is monotonic, $E^L_{\gamma+1} \supsetneq E^L_\gamma$ would imply that the intersection of $E^L_{\gamma+1}$ and $\neg E^L_{\gamma+1}$ is not empty, which is impossible, since E^L_α is coherent for all α. □

A comparison between Leitgeb's alethic jump J^L and van Fraassen's jump J^V will help us understand the pecularities of L-groundedness. In order for ϕ to be in $J^L(E)$ or $J^V(E)$, ϕ must be true when the extension of the truth predicate is E but ϕ must also satisfy some stability condition. It must have the same value in all possible extensions for Tr that somehow preserve the information given by E. The difference between J^L and J^V consists in what this is exactly taken to mean. In the case of Leitgeb's

alethic jump, the condition is that $\mathrm{Val}_{\Phi_1}\phi = \mathrm{Val}_{\Phi_2}\phi$ for allΦ_1, Φ_2 such that

$$\Phi_1 \cap (E \cup \neg E) = \Phi_2 \cap (E \cup \neg E) \quad \textit{condition (L)}$$

In the case of van Fraassen's jump, the condition is that $\mathrm{Val}_{\Phi_1}\phi = \mathrm{Val}_{\Phi_2}\phi$ for all Φ_1, Φ_2 such that

$$E \subseteq \Phi_1, \Phi_2 \text{ and } \neg E \cap \Phi_1 = \neg E \cap \Phi_2 = \emptyset \quad \textit{condition (V)}$$

Condition (V) implies condition (L),[5] because if condition (V) holds, $\Phi_1 \cap (E \cup \neg E) = E = \Phi_2 \cap (E \cup \neg E)$. We now have a simple picture of the relationships between L–groundedness and V–groundedness:

Proposition 3 *Φ_{lf} is a proper subset of $E^V_\infty \cup \neg E^V_\infty$.*

Proof Since condition (V) implies condition (L), $J^L(E) \subseteq J^V(E)$, hence if ϕ is in the least fixed point for the ordinal sequence defined from J^L, it is also in the least fixed point for the ordinal sequence defined from J^V, that is, it is V–grounded. This inclusion is strict. By Proposition 2, ϕ is in the least fixed point for the ordinal sequence defined from J^L iff it is L–grounded. The formula $\mathrm{Tr}\ulcorner 2 + 2 = 4\urcorner \vee \lambda$ is V–grounded since it will be determined as true as soon as $2 + 2 = 4$ enters the extension of Tr. But, as we have seen earlier, it is not L–grounded. □

We have thus identified one parameter which is responsible for the discrepancy between L–groundedness and K–groundedness. It is the *semantic parameter* which consists in choosing the underlying semantics for the truth-predicate (classical or partial valuations). Supervaluations bring the direct and the indirect route closer by making the indirect route more classical. Arguably, the semantic parameter is extrinsic: groundedness judgements might differ depending on the underlying logic one favors, but this is independent of what groundedness really amounts to. However, unification of the sets of grounded sentences is yet to be obtained, for counterexamples such as $\theta' = \mathrm{Tr}\ulcorner 2 + 2 = 4\urcorner \vee \lambda$ subsist. This means that other parameters remain to be identified.

Proposition 3 may be compared with Leitgeb's result (Leitgeb 2005) about groundedness for supervaluations 'à la Cantini'. Leitgeb shows that Φ_{lf} is a proper subset of $E^C_\infty \cup \neg E^C_\infty$. Cantini's jump J^C, just like Leitgeb's and van Fraassen's jumps can be analyzed into two requirements, first that the sentence be true when E is the extension of Tr and second that its truth-value is the same for various alternative pairs of extensions. In the case of Cantini's jump, the second requirement is that $\mathrm{Val}_{\Phi_1}\phi$ must be equal to $\mathrm{Val}_{\Phi_2}\phi$ for all Φ_1, Φ_2 such that

$$\Phi_1 \text{ is consistent}, \Phi_2 \text{ is consistent and } E \subseteq \Phi_1, \Phi_2 \quad \textit{condition (C)}$$

Condition (C) implies condition (V), therefore $E^V_\infty \cup \neg E^V_\infty \subseteq E^C_\infty \cup \neg E^C_\infty$, and then Leitgeb's result follows from Proposition 3. Moreover, an inspection of the proof of

[5] We are particularly indebted to Øystein Linnebo for pointing out to us that this simple fact was the crucial feature involved in the comparison between Leitgeb's approach and familiar alethic jumps.

Theorem 21 in (Leitgeb 2005) reveals that consistency of Φ_1 and Φ_2 is not really needed and that the weaker assumption in condition (V) would have been sufficient.

Comparing conditions (V) and (L) will not help us understand why formulas such as $\theta' \stackrel{\text{def}}{=} \text{Tr}\ulcorner 2 + 2 = 4\urcorner \vee \lambda$ do not get classified in the same way according to L-groundedness and groundedness in supervaluational schemes. When condition (V) is adopted, the information that some sentences have been determined to be true (those in E) or false (those in $\neg E$) is secured: stability is required in all possible extensions containing that information. When condition (L) is adopted, things are very different. Extensions for Tr come into play that are not supersets of E. This is why θ' is not L–grounded: even though we know that $2 + 2 = 4$ is true, we do not allow that information to influence our decisions about groundedness, so we are not in a position to see that $\text{Tr}\ulcorner 2 + 2 = 4\urcorner \vee \lambda$ should be true, so to speak. What kind of information do we allow ourselves? Condition (L) says that possible alternatives agree on E and $\neg E$. So the information is that sentences in E and $\neg E$ are determined, but we do not know how. Thus, for example, since $2 + 2 = 4$ gets in the extension of Tr after the first application of J^L, $\text{Tr}\ulcorner 2 + 2 = 4\urcorner$ then gets in the extension of Tr after the second application of J^L. When we know that $\text{Tr}\ulcorner 2 + 2 = 4\urcorner$ is fixed, we are in a position to know that $\text{Tr}\ulcorner\text{Tr}\ulcorner 2 + 2 = 4\urcorner\urcorner$ is fixed as well, no matter whether $\text{Tr}\ulcorner 2 + 2 = 4\urcorner$ has been fixed as true or as false. But things are different with $\text{Tr}\ulcorner 2 + 2 = 4\urcorner \vee \lambda$. When all we know is that $\text{Tr}\ulcorner 2 + 2 = 4\urcorner$ is fixed, and we do not know whether it is fixed as true or fixed as false, we are not in position to know whether $\text{Tr}\ulcorner 2 + 2 = 4\urcorner \vee \lambda$ is fixed as well. This is why $\text{Tr}\ulcorner 2 + 2 = 4\urcorner \vee \lambda$ does not get in the extension of Tr after the second application of J^L, and actually never gets in it.

$\text{Tr}\ulcorner 2 + 2 = 4\urcorner \vee \lambda$ is not L–grounded because the process by which sentences are determined to be L–grounded discards alethic information and keeps only information about groundedness. This explains why L–groundedness appeared as a mysterious mix between a strong and a weak Kleene scheme with respect to sentences of the form $\phi \vee \lambda$. When ϕ belongs to the base language, as in $2 + 2 = 4 \vee \lambda$, $2 + 2 = 4$ takes over because the alethic information about sentences of the base language is presupposed in the whole construction. When ϕ does not belong to the base language, but ends up as grounded and true, as in $\text{Tr}\ulcorner 2 + 2 = 4\urcorner \vee \lambda$, λ takes over because the alethic information about ϕ is discarded in the construction of Φ_{lf} and the information that ϕ is grounded does not suffice to ground $\phi \vee \lambda$. The rationale for the asymmetry regarding groundedness verdicts for $2 + 2 = 4 \vee \lambda$ versus $\text{Tr}\ulcorner 2 + 2 = 4\urcorner \vee \lambda$ originates in an asymmetry regarding the use of alethic information: alethic information for the base language is used, just like it is in probably all fixed-point definitions of truth, but alethic information for sentences in which the truth predicate occurs is not used, which is congenial to the dependence approach.

18.6 At the Crossroads

Understanding the mysterious mix behind L–groundedness does not constitute a vindication of the asymmetry we have pinpointed. In his own discussion of the problem, Leitgeb says, when ϕ belongs to the base language, that "from a purely semantic viewpoint, the sentence $\phi \vee \lambda$ is indistinguishable from ϕ" (Leitgeb 2005). But why should things be any different when ϕ is $\mathrm{Tr}\ulcorner 2+2=4\urcorner$? Leitgeb does not fundamentally address the core issue of this asymmetry. In the absence of any principled reason for a different treatment of alethic information, it is then tempting to restore the broken symmetry, either by totally discarding alethic information or by not discarding alethic information at all. At least at first sight, the first route looks like a dead end, because alethic information regarding the base language is obviously needed at some point to fix the extension of the truth predicate or to determine the set of grounded sentences. The second option is suggested, though not studied in detail, by (Leitgeb 2005), through a notion of conditional dependence, to which we shall now turn.

Leitgeb's suggestion consists in the following definition:

Definition 4 [Leitgeb's conditional dependence] ϕ depends on Φ given Σ iff for all $\Psi_1, \Psi_2 \subseteq \mathcal{L}_{\mathrm{Tr}}$ s.t. $\Sigma \subseteq \Psi_1, \Psi_2$ it holds that if $\mathrm{Val}_{\Psi_1}\phi \neq \mathrm{Val}_{\Psi_2}\phi$ then $\Psi_1 \cap \Phi \neq \Psi_2 \cap \Phi$.

In order to put conditional dependence to good use, we will need to define the corresponding ordinal sequence. But let us postpone spelling out the exact definition and first give a closer look at how conditional dependence works. Given $\{2+2=4\}$, $\mathrm{Tr}\ulcorner 2+2=4\urcorner \vee \lambda$ depends on $\emptyset$. As a consequence, if the fixed point construction of the set of grounded sentences eventually conditionalizes on the information that $2+2=4$, as it should, $\mathrm{Tr}\ulcorner 2+2=4\urcorner \vee \lambda$ will be declared to be grounded, restoring the symmetry with $2+2=4 \vee \lambda$. However, the symmetry is not fully restored. What about sentences such as $2+2 \neq 4 \wedge \lambda$ and $\mathrm{Tr}\ulcorner 2+2 \neq 4\urcorner \wedge \lambda$? In Φ_{lf}, their treatment suffers from the same asymmetry. The former belongs to Φ_{lf}, because the construction of Φ_{lf} builds up on the base language information that $2+2 \neq 4$ is false. The latter does not belong to Φ_{lf}, because the information that $\mathrm{Tr}\ulcorner 2+2 \neq 4\urcorner$ is grounded is not sufficient to ground it. We would expect conditional dependence to fix this asymmetry, just as it fixes the one we first discussed. But this is not the case. Given $\{2+2=4\}$, $\mathrm{Tr}\ulcorner 2+2 \neq 4\urcorner \wedge \lambda$ essentially depends on $\{2+2 \neq 4, \lambda\}$. As a consequence, $\mathrm{Tr}\ulcorner 2+2 \neq 4\urcorner \wedge \lambda$ will not be declared to be grounded, and the symmetry with $2+2 \neq 4 \wedge \lambda$ will not be restored. This problem can be addressed quite simply by treating uniformly the information about true and false sentences.

Definition 5 [conditional dependence, revised definition] ϕ depends on Φ given Σ – notation: $\phi \mathrm{dep}_{\Sigma}(\Phi)$ – iff for all $\Psi_1, \Psi_2 \subseteq \mathcal{L}_{\mathrm{Tr}}$ s.t. $\Sigma \subseteq \Psi_1, \Psi_2$ and $\Sigma \cap \neg\Psi_1 = \Sigma \cap \neg\Psi_2 = \emptyset$, it holds that if $\mathrm{Val}_{\Psi_1}\phi \neq \mathrm{Val}_{\Psi_2}\phi$ then $\Psi_1 \cap \Phi \neq \Psi_2 \cap \Phi$.

Given $\{2+2=4\}$, $\mathrm{Tr}\ulcorner 2+2 \neq 4\urcorner \wedge \lambda$ now depends on $\emptyset$. As a consequence, if the fixed point construction of the set of grounded sentences eventually conditionalizes on the information that $2+2=4$, as it should, $\mathrm{Tr}\ulcorner 2+2 \neq 4\urcorner \wedge \lambda$ will be declared to be grounded, restoring the symmetry with $2+2 \neq 4 \wedge \lambda$.

From now on, we shall work with this revised notion of conditional dependence. Analogous to the sequence Φ_α defined by Leitgeb for his original notion of dependence, one can introduce a parallel sequence $({\Phi^C}_\alpha)_{\alpha\in On}$ and $({\Gamma^C}_\alpha)_{\alpha\in On}$. The idea is again that one begins with the empty set, then takes all sentences that depend on it, and so on, but the difference with before is that at every step we presuppose (i.e. conditionalize upon) all grounded sentences that were true in the previous step. Thus, we shall use a conditional version of D^{-1}, defined by ${D^{-1}}_\Sigma(\Phi) \stackrel{\text{def}}{=} \{\phi \in \mathcal{L}_{\text{Tr}} | \phi \ \text{dep}_\Sigma(\Phi)\}$.

Definition 6 The sequences $({\Phi^C}_\alpha)_{\alpha\in On}$ and $({\Gamma^C}_\alpha)_{\alpha\in On}$ are defined co-recursively in the following way:
${\Phi^C}_0 = \emptyset$, ${\Gamma^C}_0 = \emptyset$,
${\Phi^C}_{\alpha+1} = {D^{-1}}_{{\Gamma^C}_\alpha}({\Phi^C}_\alpha)$, ${\Gamma^C}_{\alpha+1} = \{\phi \in {\Phi^C}_{\alpha+1} | \text{Val}_{{\Gamma^C}_\alpha}\phi = 1\}$,
${\Phi^C}_\beta = \bigcup_{\alpha<\beta}{\Phi^C}_\alpha$, ${\Gamma^C}_\beta = \bigcup_{\alpha<\beta}{\Gamma^C}_\alpha$.

The sequence ${\Phi^C}_\alpha$ is monotone increasing, so it has a fixed point, and a least one which we note ${\Phi_{\text{lf}}}^C$. Interestingly, one cannot do without the recursive definition of Γ_α. At every step in the expansion of the dependence set ${\Phi^C}_\alpha$, one needs to presuppose all the truths of the previous step. Choosing a conditional set Σ, say the set of arithmetical truths, and keeping it fixed throughout the recursion, would not do. The problem is that one would handle well the problematic $\text{Tr}\ulcorner 2+2=4\urcorner \vee \lambda$ but not nested instances like $\text{Tr}\ulcorner\text{Tr}\ulcorner 2+2=4\urcorner\urcorner \vee \lambda$.

Lemma 7 $J^V(E) \cup \neg J^V(E) = D_E^{-1}(E \cup \neg E)$.

Proof The condition on Φ_1 and Φ_2 in the definition of D_E^{-1} applied to $E \cup \neg E$ is the conjunction of conditions (L) and (V). But condition (V) implies condition (L), so the conjunction of conditions (L) and (V) is equivalent to condition (V). The rest of the proof is the same as for Lemma 1. □
It follows that conditional L–groundedness is equivalent to V–groundedness.

Proposition 8 $E^V_\infty \cup \neg E^V_\infty$ *is equal to* ${\Phi_{lf}}^C$.

Proof The proof is essentially the same as for Proposition 2, plus the observation that Lemma 7 and the definition of $\Gamma^C_{\alpha+1}$ imply that $\Gamma^C_{\alpha+1} = J^V(\Gamma^C_\alpha)$. □

A similar result is stated in (Leitgeb 2008), and was stated and proven independently in (van Vugt 2009). In (Leitgeb 2008) and (van Vugt 2009), the equality holds between Leitgeb's original notion of conditional dependence restricted to consistent sets and Cantini's version of supervaluations. (van Vugt 2009) also proves that the equality holds at every stage in the construction. Given the initial situation, where $\Phi_{\text{lf}} \subsetneq E^V_\infty \cup \neg E^V_\infty \subsetneq E^C_\infty \cup \neg E^C_\infty$, Proposition 8 slightly improves on these results by equating the dependence approach and supervaluational approaches closer to Φ_{lf}. The departure from Φ_{lf} to $E^V_\infty \cup \neg E^V_\infty$ is smaller than to $E^C_\infty \cup \neg E^C_\infty$. Proposition 8 is proven independently by (Meadows 2013).[6]

[6] We came to know (Meadows 2013) only after finishing a first version of the present paper. The main difference with the present work is that Meadows (2013) does not arrive at the equivalence result by an analysis of conditions (V) and (L) in alethic jumps. As a consequence, the lesson he draws

18.7 The Future of Dependence Jumps

The gist of our analysis has consisted in accounting for the divergence between the alethic approach and the dependence approach to groundedness. The 'directness' of the dependence approach amounts to focusing only on the incremental determination of the set of grounded sentences while ignoring the information about which truth-value grounded sentences will eventually receive. By contrast, the 'indirectness' of the alethic approaches amounts to fully using the information about which truth-value grounded sentences have received when they got grounded. Setting aside extrinsic semantic features, such as the choice of a supervaluational scheme or a three-valued logic, this seems to us to be the main conceptual difference between Leitgeb's approach and more familiar alethic definitions of the set of grounded sentences.

Asking whether the direct and the indirect approach to groundedness may converge then amounts to asking what happens to the direct approach if the alethic information is not thrown away. However the answer to this question is disarmingly simple: not only the two approaches yield the same result, but the difference between them vanishes. From a technical viewpoint, this shows in the proof of Lemma 7. Making dependence conditional simply makes the originality of the alethic jump generated from the dependence jump go away, because the conjunction of condition (V) and condition (L) is equivalent to condition (V) alone.

However, what is original in the dependence approach is also problematic: as the discussion of particular groundedness verdicts has shown, L-groundedness does not treat information about truth-values uniformly. It is only the alethic information about sentences containing the truth predicate which is sidestepped. But then, there seems to be no good conceptual reason not to treat on a par $2 + 2 = 4 \vee \lambda$ and $\mathrm{Tr}\ulcorner 2 + 2 = 4\urcorner \vee \lambda$ as far as groundedness is concerned. The technical reason why they are treated differently in the definition of Φ_{lf} is clear – discarding alethic information about the ground language looks as a non-starter. However, we do not see how this technical point could count as a conceptual argument in favor of Φ_{lf}. Conditional dependence, at least in its revised version, does provide a fix, but, again, the fix consists in going back to a familiar supervaluational alethic jump.

A complete assessment of the dependence approach would require to have at least two more questions answered, that we would like to briefly present here as perspectives for further work. First, is there really no way to make the dependence approach uniform by treating alethic information about the base language in the same way as alethic information about sentences containing the truth predicate is treated? An option that we have not pursued here would be to take into account alternative

from the equivalence is different: because of the asymmetry problem, Meadows only concludes that properly spelling out the dependence approach makes it equivalent to the supervaluational approach. Our conclusions are less pessimistic. We agree with Meadows that Φ_{lf} is not satisfactory. But as discussed in sect. 18.7, we consider that the question whether there exists a conceptually well-motivated dependence fixed point different from supervaluational fixed point is still open. The upshot of the present paper consists in elucidating what is the specific difference between alethic jumps and dependence jumps.

models for the base language, so that $2+2 = 4 \vee \lambda$ would end up not being grounded, because even though $2 + 2 = 4$ is grounded, we do not allow ourselves to use the information that the interpretation of 2, 4 and $+$ is such that $2 + 2 = 4$ is true.

Second, we have seen that the dependence approach can be mimicked along the lines of the alethic approach. Φ_{lf} can be defined as $E^X_\infty \cup \neg E^X_\infty$ for some alethic jump X. This does not really come as a surprise, given that the dependence approach uses less information than alethic approaches. For at least one alethic jump, namely J^L, the corresponding set of groundedness can be directly obtained as the smallest fixed point of a dependence jump. But does this work with alethic jumps other than J^L? J^V does not seem to count as another positive instance, because the dependence jump used to get to $E^V_\infty \cup \neg E^V_\infty$ essentially relies on an alethic jump, as the co-recursion shows. Is it possible to prove that, in some sense to be made fully precise, $E^V_\infty \cup \neg E^V_\infty$ or $E^K_\infty \cup \neg E^K_\infty$ cannot be directly obtained as the least fixed points of a 'pure' dependence jump? Such a proof would aim at explaining what is special with Φ_{lf}.

Finally, J^L, the alethic jump corresponding to the dependence jump D^{-1}, appears as an alternative supervaluational jump, based on a condition which is more demanding than the conditions used by van Fraassen's jump J^V or Cantini's jump J^C. This suggests another line of research, which would aim at studying possible stability conditions for supervaluational jumps. Condition (L) is weaker than than conditions (V) and (C), are there interesting weakest or strongest conditions?

Acknowledgements We would like to thank Serge Bozon, Paul Égré, Hannes Leitgeb, Øystein Linnebo, Philippe de Rouilhan, and Jönne Speck for their insightful comments on earlier versions of this research. We are particularly grateful to Øystein Linnebo who greatly helped us clarify the rationale behind results proven in (van Vugt 2009). We also need to thank various audiences in Gothenburg, London and Paris for their helpful feedback. The present work originates in the second author's master thesis under the supervision of the first author (van Vugt 2009). Results in the present work were obtained independently of (Meadows 2013), which proves Proposition 8 *supra* as its main result (see Footnote 6 for a more detailed discussion). The work of the first author was partly supported by the ESF-funded project 'Logic for Interaction', a Collaborative Research Project under the Eurocores program LogICCC.

References

Cantini, A. (1990). A theory of formal truth arithmetically equivalent to id_1. *Journal of Symbolic Logic,* 55, 244–259.

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

Leitgeb, H. (2005). What truth depends on. *Journal of Philosophical Logic,* 34, 155–192.

Leitgeb, H. (2008). Towards a logic of type-free modality and truth. In D. Costas et al. *Logic colloquium 2005* (pp. 68–85) Dimitracopoulos: Cambridge University Press.

McGee, V. (1992). Maximal consistent sets of instances of Tarski's schema (T). *Journal of Philosophical Logic,* 21, 235–241.

Meadows, T. (2013). Truth, dependence and supervaluation: Living with the ghost. *Journal of Philosophical Logic,* 42 (2), 221–240.

Tarski, A. (1955). A lattice–theoretical fixpoint theorem and its applications. *Pacific journal of Mathematics* ,5(2), 285–309. http://projecteuclid.org/euclid.pjm/1103044538.

van Vugt, F. (2009). What makes a sentence be about the world? Master's thesis, Cogmaster, Ecole Normale Supérieure, Paris (France).

Yablo, S. (1993). Paradox without self-reference. *Analysis,* 53(4), 251–252.

Chapter 19
On Stratified Truth

Andrea Cantini

Abstract Is there a consistent axiomatization of a stratified form of the Tarskian hierarchy, where stratification is meant in the sense of Quine's *New Foundations* NF? In the following we propose a system of truth and abstraction, which might be regarded as an answer to the problem.

19.1 Stratified Truth: Introduction

Feferman recently[1] raised the question of finding a consistent axiomatization of the Tarskian hierarchy, where stratification is understood in Quine's sense. Some years ago in Cantini 2004 we sketched a reconstruction of a truth theory, based on the stratification discipline and aimed at a discussion of the Russellian paradox about propositions and sets.[2]

In the following we propose a modification of Cantini 2004, which might be regarded as partially approaching a solution to Feferman's problem. The basic idea is to devise a theory of truth SFT (= Stratified Fregean Truth) with strong expressive power, which is based upon *stratification as a means to achieve consistency*. As we shall see, stratification is exactly specified by the syntax of our system and validated by a set–theoretic interpretation in Quine's NF.

Let us anticipate a few informal considerations. First of all, the inspiring idea is to define a sort of analogue to the notion of Frege structure (in the sense of Aczel

This paper arises from the slides for the talk *Marginalia to self-referential truth*, presented at the Conference on *Axiomatic Theories of Truth* (New College, University of Oxford, September 19–20, 2011). We wish to thank the organizers for the nice hospitality and the stimulating environment. The research is supported by MIUR, under the project *Thinking and Computing, PRIN 2008* and within the frame of the University of Florence local research unit, sub-project *Abstraction and computation: logical and epistemological aspects*.

[1] In his talk at the Princeton Conference *Pillars of Truth*, April 8–10, 2011.

[2] See Russell's Principles of Mathematics, appendix B; the paradox is also known as the Russell–Myhill paradox.

A. Cantini
Dipartimento di Lettere e Filosofia, Sezione di Filosofia,
Università degli studi di Firenze via Bolognese, 52 50139 Firenze, Italy
e-mail: andrea.cantini@unifi.it

T. Achourioti et al. (Eds.), *Unifying the Philosophy of Truth,* Logic, Epistemology,
and the Unity of Science 36, DOI 10.1007/978-94-017-9673-6_19

Aczel 1980), whose existence does not depend on lambda calculus, but on assuming a version of the discipline of types as codified by the practice of stratification. The universe is a kind of *abstract logical system* with a built-in *reflection mechanism*: if A represents a given proposition, there will be an object represented by $[A]$, which can be transformed into *a statement of a higher level* by applying the truth predicate T to $[A]$. But, in order to achieve consistency, we are forced to keep track of this natural level stratification: if $[A]$ is given type level i, $T([A])$ is assigned type $i+1$. We underline that truth is here regarded as a predicate T having a wider domain than usual truth predicates: T applies *not only to sentences of an inductively defined formal language*, but in general to objects of the given universe, which will possibly model – or play the role of – propositions (henceforth termed as *propositional objects*). Properly speaking, propositions remain defined *after* the semantical notion of truth is available (see definition 2.1), but we identify certain structures as possible propositional objects (see below definition 1.5). In this connection, the universe can be regarded as a sort of abstract syntax, where objects can be used to convey meanings. Thus we assume that the universe is closed under constructors, that are intended to represent *logical operations* and build *propositional objects*, and under an abstraction operator forming *predicative objects*. In particular, the universe includes objects encoding formulas of our truth language $\mathcal{L}_T$ and definable predicates thereof: if A is a sentence of $\mathcal{L}_T$, a map $A \mapsto [A]$ for forming propositional objects and an abstraction operation $x, A \mapsto [x|A]$ for building predicates are available, in such a way that the free variables of $[A]$ are the same as the free variables of A, and the free variables of $[x|A]$ are the free variables of A minus x (in short, $FV(A) = FV([A])$ and $FV([x|A]) = FV(A) - \{x\}$).

No type restriction is imposed in forming $[A]$ and $[x|A]$; but, as we shall see in the next subsection, the *use* of these expressions has to be suitably restricted according to the initial type-theoretic intuition.

If we exclude equality which is assumed as given, truth T is the only predicate of our language. But we also assume that there is a primitive operation $pred$, such that, roughly, $pred(t, s)$ expresses the fact that the object represented by s falls under the concept represented by t. T and $pred$ allow to express predication: if $[x|A]$ represents a predicate P defined by a given formula A, the result of the application of P to a, is rendered by $pred([x|A]), a)$. Hence the claim that a falls under P simply becomes the claim that $pred([x|A], a)$ *is true*, i.e. $T(pred([x|A], a))$.

One may wonder if self-referential constructions are, to a limited extent, allowed in the present framework. The answer is positive, but we stress that the present treatment is different from self-reference in the case of standard formal languages $\mathcal{L}$ (e.g. first order Peano arithmetic or ZFC), where one makes use of a substitution operation acting on Gödel numbers of $\mathcal{L}$-formulas and $\mathcal{L}$-terms. In the present framework, we assume that the universe is closed under a fixed point operator acting on terms, which depend *extensionally* on their parameters and are *homogeneously stratified*, i.e. arguments and values are assigned the same type. Roughly, the idea is that, if a propositional function F is extensional in a parameter x of given type i, then there

is a fixed point c of type i, i.e. such that $F(c) = c$. Of course, one has to clarify what is the meaning of *extensional in* x and the notion of type assignment in a formally untyped framework. Once clarified, these requirements are sufficient to sterilize self-reference and Liar's arguments: according to the basic intuition, if $tr(y)$ is the object representing the truth of y, $tr(y)$ is *type raising*, i.e. is assigned type one greater than y itself. Hence no fixed point of $tr(y)$ or of its negation will arise.

As to the justification, we shall see that self-reference follows by the set theoretic representation of logical constructors and as a consequence of (a variant of) the Knaster–Tarski theorem.

19.1.1 The Language and Its Stratification

Definition 1.1 The language $\mathcal{L}_T$ and its syntax. $\mathcal{L}_T$ includes:

- a unary predicate T for truth, a binary predicate $=$ for equality;
- binary function symbols id, $pred$, and; unary function symbols tr, neg, all;
- the binding operators $[- \mid -]$ (abstraction) and μ (fixed point).

id, $pred$, tr internally represent *basic constructors* (i.e. constructors for atomic formulas), while neg, and, all internally represent *logical constructors*.

Recall that, if E is an expression (i.e. term or formula) of the language, $FV(E)$ denotes the set of free variables of E.

Definition 1.2 We present a simultaneous inductive definition of the notions of (i) *term*, (ii) *term operative in a list* $\vec{x}$ *of parameters*, and (iii) *formula*:

- variables (and possibly individual constants when available) are terms;
- if x is a variable and $\vec{y}$ is a list of variables possible including x, then x is operative in $\vec{y}$;
- if t, s are terms, and f (g) is a binary (unary) function symbol, then $f(t, s)$ ($g(t)$) is a term;
- if t operative in $\vec{y}$ and s is operative in $\vec{z}$, then $and(t, s)$, $id(t, s)$ are operative in $\vec{y}, \vec{z}$, and $neg(t)$, $all(t)$ are operative in $\vec{y}$;
- if t is operative in $\vec{y}$, $pred(t, s)$ is operative in $\vec{y}$;
- if A is a formula, $[x|A]$ is a term such that $FV([x|A]) = FV(A) - \{x\}$;
- if $t(y, \vec{x})$ is operative in $y, \vec{x}$, then $\mu y t(y, \vec{x})$ is a term operative in $\vec{x}$, such that $FV(\mu y t(y, \vec{x})) = FV(t(y, \vec{x})) - \{y\}$;
- if t, s are terms, $T(t)$, $t = s$ are formulas; if A, B are formulas, then $\neg A$, $A \wedge B$, $\forall x A$ are formulas, and $FV(\forall x A]) = FV(A) - \{x\}$.

NB. The notion of *operative in* $\vec{x}$ corresponds to being extensional in $\vec{x}$, as informally hinted at in the introduction. As to specific examples, note that $neg(pred(y, x))$ is operative in y but not in x; $tr(y)$, $neg(tr(y))$ are not operative in y.

Of course, the language is suspiciously Fregean in the widest sense, and it is to be expected that we need some sort of restriction for governing truth and predication.

Indeed, in order to state the T-schema and the comprehension schema, we extend *the discipline of types* to arbitrary expressions E of the new language. The basic idea is that predication makes sense only in agreement with a suitable modification of Quine's stratification: informally, a predicate (represented by) t truly applies to s, where s is assigned type i, only if t is assigned type $i + 1$. Similarly, truth has an *implicit hierarchical structure*: when we apply the predicate T to (a propositional object represented by) the term t, T must be assigned a level higher than the type assigned to t.

Definition 1.3 (Stratification of terms and formulas.) If E is an expression, E is *stratified* iff it is possible to assign a natural number (type in short) to each term occurrence and to each T-occurrence of E, so that:

1. all free occurrences of the same variable in any subexpression of E have the same type;
2. in each expression of the form $pred(t, s)$ the type of t is one greater than the type of its argument s; $pred(t, s)$ is assigned the type of t;
3. each expression of the form $tr(t)$ is assigned a type one greater than the type of t; in each expression of the form $T(t)$ T is assigned a type one greater than the type of t;
4. in each expression of the form $t = s$, $id(t, s)$ the type of t is the same type as s; $id(t, s)$ is assigned the same type of t (and hence of s);
5. each expression of the form $neg(t)$, $all(t)$ is assigned the same type of t;
6. each expression of the form $and(t, s)$ is assigned the same type as the type of t, s (that must have received the same type);
7. each term of the form $[x \mid C]$ is assigned a type one greater than the type assigned to x, and all the free occurrences of x in C receive the same type;
8. in each expression of the form $\forall x A$, if x is free in A, then the free occurrences of x in A and the occurrence of x in $\forall x$ receive the same type;
9. each term of the form $\mu y t(y, \vec{x})$ is assigned the same type as y and t, and all the free occurrences of $\vec{x}$ in t receive the same type.

NB. Within the same statement, different occurrence of T can be assigned different type labels and this makes sense of the idea of *typical ambiguity* in the semantical framework we are dealing with. Observe also that the definition of stratification imposes a *homogeneity condition* on $and(t, s)$; a semantical justification is to be found in the Quinean interpretation developed in Sect. 19.3 below.

Definition 1.4 We then inductively introduce $A \mapsto [A]$ with $FV(A) = FV([A])$:

- $[t = s] := id(t, s)$;
- $[T(t)] := tr(t)$
- $[\neg A] := neg([A])$;
- $[A \wedge B] := and([A], [B])$;
- $[\forall x A] := all([x \mid A])$

NB. It is not always true that, if A is stratified, then so is $[A]$; for instance, $T(x) \wedge T([Tx])$ is stratified (e.g. assign 0 to x, 1 to the first occurrence of T and 2 to the

second); but $[T(x) \wedge T([Tx])]$ is not, as it fails to meet the homogeneity condition required by *and*. Thus it is the very mechanism of associating propositional objects that in this context can ruin stratification.

Definition 1.5 [P-Form] If an object x is in the range of the logical constructors, then it is called a *P-form*:

$$Pfor(x) \Leftrightarrow \exists y(x = tr(y)) \vee \exists z(x = neg(z) \vee x = all(z)) \vee$$
$$\vee \exists u \exists v(x = id(u, v) \vee x = and(u, v))$$

Roughly, a P-form is an object which is (possibly) apt to represent a proposition. Observe that $Pfor(x)$ is stratified (assign 1 to x, u, v, z, 0 to y).

19.1.2 Axioms of SFT

SFT consists of the classical logical calculus (say, Hilbert-style) with equality and, in addition, the following axioms.

1. Compositional T-axioms:

$$T(id(x, y)) \leftrightarrow x = y;$$
$$T(neg(id(x, y)) \leftrightarrow \neg x = y;$$
$$T(tr(x)) \leftrightarrow T(x);$$
$$T(neg(tr(x)) \leftrightarrow \neg T(x);$$
$$T(neg(neg(x)) \leftrightarrow T(x);$$
$$T(and(x, y)) \leftrightarrow T(x) \wedge T(y);$$
$$T(neg(and(x, y)) \leftrightarrow T(neg(x)) \vee T(neg(y));$$
$$T(all(f)) \leftrightarrow \forall x T(pred(f, x));$$
$$T(neg(all(f)) \leftrightarrow \exists x T(neg(pred(f, x)))$$

2. T-consistency:

$$\neg(T(a) \wedge T(neg(a)))$$

3. T is well-defined on predication:

$$T(pred(f, x)) \vee T(neg(pred(f, x)))$$

4. Stratified β-conversion: if A is stratified,

$$T(pred([x|A], u)) \leftrightarrow T([A[x := u]])$$
$$T(neg(pred([x|A], u))) \leftrightarrow T([\neg A[x := u]])$$

Roughly, this schema states that, insofar as stratified conditions and truth contexts are involved, predicate abstraction and predicate application behave as inverse to each other.[3]

5. Self-reference: if t is operative in the list $y, \vec{x}$ and stratified,

$$\forall \vec{x}(t(\mu y t(y, \vec{x}), \vec{x}) = \mu y t(y, \vec{x}))$$

6. P-form:

$$T(x) \rightarrow Pfor(x)$$
$$\neg Pfor(x) \rightarrow T(neg(x))$$

The P-form axioms grant that true objects lie in the range of logical constructors; furthermore, any object inaccessible to logical constructors is classified as (representing) False.

7. μ-Extensionality: if two terms operative in $\vec{x}, y$ and stratified, are pointwise equal, then the respective fixed points coincide:

$$\forall \vec{x} \forall y(t(y, \vec{x}) = s(y, \vec{x})) \rightarrow \forall \vec{x}(\mu y t(y, \vec{x}) = \mu y s(y, \vec{x}))$$

8. Basic constructors and logical constructors are injective but not surjective, and their images are disjoint. In details, if f, g are distinct basic or logical constructors, f unary and g binary, then:

$$f(x) = f(y) \rightarrow x = y$$
$$g(x, y) = g(u, v) \rightarrow x = u \wedge y = v$$
$$\forall x \forall y \forall z(f(x) \neq g(y, z))$$
$$\exists x \neg Pfor(x)$$

Remark 1 The equivalence between $T(neg(tr(x)))$ and $\neg T(x)$ is *strongly non-kripkean* and makes the truth predicate closer to its classical counterpart. A similar comment holds for the clause involving predication.

19.1.3 Stratified Truth in SFT

The truth predicate is not only provably partial:

---

[3] Of course, one might simply postulate β-conversion at the object level, i.e. if A is stratified,

$$pred([x|A], u) = [A[x := u]]$$

Then the schemata would be trivially derivable. The reason is that we do not know how to prove its consistency.

Proposition 1.6 SFT *proves, for some closed term L:*

$$\neg T(L) \wedge \neg T(neg(L))$$

Moreover:

$$T(\neg T(L) \wedge \neg T(neg(L)))$$

Proof By self-reference choose $L = neg(L) = \mu y.neg(y)$. Then apply logic, T-consistency and the axioms relating T with tr, neg and and. □

Hence, not surprisingly, T is provably internally undefined on (the simplest variant of) the Liar; but, interestingly, T *internally believes* this fact.

Lemma 1.7 *(Compositional schemata) If A and B are arbitrary,*

(i) $T[\neg A] \leftrightarrow \neg T[A]$*;*
(ii) $T[A \wedge B] \leftrightarrow T[A] \wedge T[B]$*;*
(iii) $T[\forall x A] \leftrightarrow \forall x T[A(x)]$*, provided A is stratified.*

Proof As to (i), proceed by induction on A using the fact that T is well-defined on identities, predication and truth. (ii): it follows from the sixth axiom. (iii): apply stratified β-conversion. □

Proposition 1.8 *(Uniform stratified T-schema). If A is stratified,* SFT *proves:*

$$\forall x(T([A(\vec{x})]) \leftrightarrow A(\vec{x})) \tag{19.1}$$

$$T[\forall x(T([A(\vec{x})]) \leftrightarrow A(\vec{x}))] \tag{19.2}$$

Proof We check by simultaneous induction on A

$$(T([A]) \leftrightarrow A) \wedge (T([\neg A]) \leftrightarrow \neg A)$$

If A is of the form $t = s$, $T(t)$, apply the corresponding axioms of SFT.

If A is of the form $B \wedge C$, even if $A \wedge B$ is stratified, $[A \wedge B]$ may be not. However, by $\forall$-instantiation of the compositional axiom about T and $\wedge$, we obtain

$$T([A \wedge B]) \leftrightarrow T([A]) \wedge T([B])$$

where the left hand side is not stratified in general, while the right hand side is stratified. Then we apply IH.[4]

If A is of the form $\neg B$, apply the compositional SFT-axioms involving negated $\wedge$, double negation, and IH.

Let us consider the case of a negated universal quantifier. Then we use the axioms relating T, $\neg\forall$, together with β-conversion and IH in the final step:

$$T([\neg\forall x A]) \leftrightarrow T(neg(all([x|A])))$$

[4] Henceforth IH stands for induction hypothesis in short.

$$\leftrightarrow \exists u T(neg(pred([x|A], u)))$$
$$\leftrightarrow \exists u T([\neg A[x := u]])$$
$$\leftrightarrow \exists u(\neg A[x := u]) \equiv \neg \forall x A$$

The case of positive $\forall$ is similar. □

The stratified T-schema implies that T strongly deviates from the behaviour of self-referential truth predicates à la Kripke–Feferman, which cannot in general be applied to the truth axioms themselves, nor to *arbitrary* logical axioms. On the contrary, T *provably believes that it is two-valued and consistent*; further, it recognizes that each closure condition is also internally true.

Corollary 1.9

(i) SFT *proves:*

$$T([T(a) \vee \neg T(a)]);$$
$$T([\neg(T(a) \wedge T(neg(a)))])$$

(ii) Moreover, if Axiom *is an instance of a compositional T-axiom or T-welldefinedness,* SFT *proves* $T([\mathsf{Axiom}])$.

Proof Observe that the consistency statement as well as *tertium non datur* for T and the compositional axioms are stratified; hence the claim is a consequence of the stratified truth schema. □

Remark 2 One may wonder whether the fixed point property can be extended, e.g. up to include the constructors tr, $pred$ and combinations thereof. It is immediate to see that the answer is negative. Indeed, assume that there exists e such that

$$e = neg(tr(e))$$

Then $T(e) \leftrightarrow T(neg(tr(e))) \leftrightarrow \neg T(e)$: contradiction! The reason is that, roughly, as we shall see in the model construction, neither $x \mapsto tr(x)$ nor $x \mapsto pred(y, x)$ are monotone (in the sense of set theoretic inclusion) with respect to x.

We conclude by showing that SFT proves that its truth predicate is indeed the fixed point of a natural positive operator. Let $\mathcal{V}(x, T)$ be the formula:

$$\begin{aligned}
&\exists v(\neg Pfor(v) \wedge x = neg(v)) \vee \\
\vee\ &\exists w_1((x = [T(w_1)] \wedge T(w_1)) \vee \\
&\vee (x = [\neg T(w_1)] \wedge \neg T(w_1))) \vee \\
\vee\ &\exists w_3(x = neg(neg(w_3)) \wedge\ T(w_3)) \vee \\
\vee\ &\exists w_4 \exists w_5((x = id(w_4, w_5) \wedge\ w_4 = w_5) \vee \\
&\vee (x = neg(id(w_4, w_5)) \wedge w_4 \neq w_5)) \vee \\
\vee\ &\exists w_6 \exists w_7(((x = and(w_6, w_7)) \wedge\ T(w_6) \wedge\ T(w_7)) \vee
\end{aligned}$$

$$\vee\,(x = neg(and(w_6, w_7)) \wedge\ (T(neg(w_6)) \vee\ T(neg(w_7))))) \vee$$
$$\vee\quad \exists f((x = all(f) \wedge \forall z T(pred(f, z))) \vee$$
$$\vee\,(x = neg(all(f)) \wedge \exists z T(neg(pred(f, z)))))$$

Theorem 1.10 *(Fixed point principle)*

$$\forall x(T(x) \leftrightarrow \mathcal{V}(x, T))$$

Proof $\Rightarrow$: let $T(x)$. Then $Pfor(x)$. If $x = neg(y)$ and not $Pfor(y)$, clearly $\mathcal{V}(x, T)$. If $x = neg(y)$ but $Pfor(y)$, we distinguish several cases and we apply the T-compositional axioms from left to right. E.g. if $x = neg(id(u, v))$, then $\neg u = v$ and we conclude $\mathcal{V}(x, T)$.

$\Leftarrow$: if $\mathcal{V}(x, T)$, we again argue by cases using T-axioms from right to left. □

Corollary 1.11 *(Internal fixed point principle)*

$$T[\forall x(T(x) \leftrightarrow \mathcal{V}(x, T))]$$

Proof The statement of the fixed point theorem is stratified: hence apply the theorem and the stratified T-schema. □

19.2 On a Paradox About Propositions and Truth

The problem at issue is whether the predicate of *being a proposition* actually defines a genuine class (type) in Russell's sense. Assume that the answer is positive. Then, according to Russell 1903, a contradiction with a version of Cantor's theorem could be derived: there would exist an injection (see below definition 2.3 and lemma 2.4) from the collection of classes, whose elements are propositions, into the collection of propositions. It turns out that in the present situation stratification blocks the diagonalization leading to contradiction.

Definition 2.1

(i) $P(a) :\Leftrightarrow T(a) \vee T(neg(a))$;
(ii) $V := [x|x = x]$;
(iii) $imp(a, b) := neg(and(a, neg(b)))$;
(iv) $or(a, b) := neg(and(neg(a), neg(b)))$.

Pa formally represents the predicate "a is a proposition". We also define $a \subseteq b$ for $\forall u(T(pred(a, x)) \rightarrow T(pred(b, x)))$.

Proposition 2.2 *(SFT). The collection of all propositions is a proper subset of the universe:*

$$[x \mid P(x)] \subset V$$

Moreover P has the following closure properties:

$$P(id(y,x)) \wedge P(pred(y,x)) \wedge P(tr(x));$$
$$P(a) \wedge (T(a) \rightarrow P(b)) \rightarrow P(imp(a,b));$$
$$P(a) \wedge P(b) \rightarrow P(and(a,b)) \wedge P(or(a,b));$$
$$P(a) \rightarrow T([Pa]);$$
$$P(all(f)).$$

The first claim is a consequence of proposition 1.6. As to the remaining properties, apply the T-compositional axioms.

Note also that $P(all(f))$ implies $\forall x P(pred(f,x))$, i.e. *every set defines a propositional function*.

We now conclude by representing Russell's contradiction of appendix B (see Russell 1903) within the theory of propositions and truth.

Definition 2.3

$$\tau(f) := [P(all(f))]$$

By definition of the map $A \mapsto [A]$, the axiom that logical operators are injective, and proposition 2.2, we obtain:

Lemma 2.4 SFT proves:

$$P(\tau(f)) \wedge T(\tau(f));$$
$$\tau(f) = \tau(g) \rightarrow f = g$$

Informally, the operation τ is a well-defined injective map from sets into truths (and propositions). Also, observe that τ is a type-raising operation: if f is assigned type 0, $\tau(f)$ must receive type 1 (since it contains the operator tr). As a consequence, the formula

$$(\exists f \subseteq P)(\neg T(pred(f,x)) \wedge x = \tau(f))$$

cannot be stratified.

Proposition 2.5 SFT *proves:*

$$\neg\, \exists d \forall x (T(pred(d,x)) \leftrightarrow (\exists f \subseteq P)(\neg T(pred(f,x)) \wedge x = \tau(f))$$

Proof Assume by contradiction that there exists d such that

$$\forall x (T(pred(d,x)) \leftrightarrow (\exists f \subseteq P)(\neg T(pred(f,x)) \wedge x = \tau(f)))$$

If we choose $x := \tau(d)$, we get

$$T(pred(d, \tau(d))) \leftrightarrow (\exists f \subseteq P)(\neg T(pred(f, \tau(d))) \wedge \tau(d) = \tau(f)))$$

Assume $T(pred(d, \tau(d)))$. Then, as τ is injective, $d = f$ and hence $\neg T(pred(d, \tau(d)))$. Hence by minimal logic $\neg T(pred(d, \tau(d)))$. Also $d \subseteq P$: indeed, if x is arbitrary and $T(pred(d, x))$, then $x = \tau(f)$ for some f, whence by lemma 2.4, $P(\tau(f))$, i.e. $P(x)$. Hence by assumption on d, the contradiction $T(pred(d, \tau(d)))$. □

This is the solution of the paradox in appendix B: the diagnosis is that the paradoxical set d does not exist according to the discipline of types even in a liberalized Quinean sense.[5]

19.3 Embedding Stratified Truth in NF

Let $\mathcal{L}_s$ be the elementary set theoretic language, which comprises the binary predicate symbol $\in$. $\mathcal{L}_s$–terms are simply individual variables ($x, y, z, \ldots$); prime formulas (atoms) have the form $t \in s$, $t = s$ (t, s terms). $\mathcal{L}_s$–formulas are inductively generated from prime formulas by means of sentential connectives and quantifiers. The elementary set theoretic language $\mathcal{L}_s^+$ is obtained by adding to $\mathcal{L}_s$ the abstraction operator $\{- \mid -\}$; $\mathcal{L}_s^+$–terms and formulas are then simultaneously generated. The clause for introducing class terms has the form: if φ is a formula, then $\{x \mid \varphi\}$ is a term where $FV(\{x \mid \varphi\}) = FV(\varphi) - \{x\}$ ($FV(E)$ is the set of free variables occurring in the expression E). Two terms (formulas) are called α–congruent if they only differ by renaming of bound variables; we identify α–congruent terms (formulas).

19.3.1 Stratified Comprehension

As usual for Quine's systems, we need *stratification*; we also define a restricted notion thereof, which is motivated by the consideration of "loosely predicative" class existence axioms.

---

[5] We underline that our formalization does not literally represent the paradox of the final section 500 of 2.2. For the reader's sake here is Russell's text:

> If m be a class of propositions, the proposition *"every m is true"* may or not be itself an m. But there is a one-one-relation of this proposition to m: if n be different from m, *"every n is true"* is not the same proposition as as *"every m is true"*. Consider now the whole class of propositions of the form *"every m is true"*, and having the property of not being members of their respective m's. Let this class be w, and let p be the proposition *"every w is true"*. If p is a w, it must possess the defining property of w; but this property demands that p should not be a w. On the other hand, if p is not a w, then p does possess the defining property of w, and therefore is a w. Thus the contradiction appears unavoidable.

(i) φ is *stratified* iff it is possible to assign a natural number (type in short) to each term occurrence [6] of φ in such a way that
- if $t \in s$ is a subformula of φ, the type of s is one greater than the type of t; if $t = s$ is a subformula of φ, the type of s is the same as the type of t;
- all free occurrences of the same variable in any subformula of φ have the same type;
- if x is free in ψ and $\forall x \psi$ is a subformula of φ, then the 'x' in $\forall x$ and the free occurrences of x in ψ receive the same type;
- if $t := \{x \mid \beta\}$ occurs in φ, x is free in β, then t is assigned a type one greater than the type assigned to x, and all the free occurrences of x in β receive the same type.

(ii) $\{x \mid \varphi\}$ is stratified if φ is stratified;

(iii) a stratified term $\{x \mid \varphi(x, \vec{y})\}$ is *loosely predicative* iff for some type $i \in \omega$, $\{x \mid \varphi(x, \vec{y})\}$ has type $i + 1$, no (free or bound) variable of $\varphi(x, \vec{y})$ is assigned type greater than $i+1$; a stratified term $\{x \mid \varphi(x, \vec{y})\}$ is *predicative* iff $\{x \mid \varphi(x, \vec{y})\}$ is loosely predicative and in addition no quantified variable of $\varphi(x, \vec{y})$ is assigned the same type as $\{x \mid \varphi(x, \vec{y})\}$ itself.

(iv) φ is $n + 1$–*stratified* iff φ is stratified by means of $0, \ldots, n$.

For instance, $\bigcup a = \{x \mid (\exists y \in a)(x \in y)\}$ is not loosely predicative, since it requires type 2, but $\bigcup a$ itself has type 1; $a \cap b = \{x \mid x \in a \wedge x \in b\}$ is predicative.

Definition 3.1 The system NF comprises:

(i) predicate logic for the extended language[7];

(ii) class extensionality: $\forall x \forall y (x =_e y \rightarrow x = y)$, where

$$t =_e s :\Leftrightarrow \forall x (x \in t \leftrightarrow x \in s)$$

(iii) stratified explicit comprehension SCA: if φ is stratified, then

$$\forall u (u \in \{x \mid \varphi(x, \vec{y})\} \leftrightarrow \varphi(u, \vec{y}))$$

[6] Individual constants included; these can be given any type compatible with the clauses below.

[7] If the abstraction operator is assumed as primitive, the extended logic contains the schema

$$\forall u (\varphi(u) \leftrightarrow \psi(u)) \rightarrow \{x \mid \varphi(x)\} = \{x \mid \psi(x)\}$$

Other systems

(a) NFP (NFI) is the subsystem of NF, where SCA is restricted to (loosely) predicative abstracts.
(b) NF_k (NFI_k, NFP_k) is the subsystem of NF (NFI, NFP), where (at most) k types are allowed for stratification.

Remark 3 By a theorem of Crabbè 1982, NFI is provably consistent in third order arithmetic. The details of the (different) consistency proofs for NFI can be found in Crabbè 1982 and Holmes 1995.
In order to carry out a Kripke-like construction in the NF-systems and to represent the syntax, we shall essentially exploit Quine's homogeneous pairing operation, which *does require extensionality* and the existence of a copy of the natural numbers. But it is not difficult to check that Quine's pairing is indeed well-defined already in NFI. First of all, the collection of Fregean natural numbers is a set in NFI. Define:

$$\begin{aligned} \emptyset &= \{x \mid x \neq x\} \\ V &= \{x \mid x = x\} \\ 0 &= \{\emptyset\} \\ a+1 &= \{x \cup \{y\} \mid x \in a \wedge y \notin x\} \\ Cl_N(y) &\Leftrightarrow 0 \in y \wedge \forall x(x \in y \rightarrow (x+1) \in y) \\ \mathcal{N} &= \{x \mid \forall y(Cl_N(y) \rightarrow x \in y)\} \end{aligned}$$

NFI proves the existence of $\mathcal{N}$; in fact, by inspection, all the above sets above are loosely predicative. Furthermore, we have, provably in NFI:

Lemma 3.2 *(NFI)*

$$Cl_N(\{x \mid \varphi(x)\}) \rightarrow \mathcal{N} \subseteq \{x \mid \varphi(x)\} \tag{19.3}$$

$$(\forall x)(x \in \mathcal{N} \leftrightarrow x = 0 \vee (\exists y \in \mathcal{N})(x = y+1)) \tag{19.4}$$

$$\emptyset \notin \mathcal{N} \wedge (\forall x \in \mathcal{N})(V \notin x) \tag{19.5}$$

$$(\forall x \in \mathcal{N})(x+1 \neq 0) \tag{19.6}$$

$$(\forall x \in \mathcal{N})(\forall y \in \mathcal{N})(x+1 = y+1 \rightarrow x = y) \tag{19.7}$$

(In (19.3) $\{x \mid \varphi(x)\}$ must be loosely predicative).

Clearly $\mathcal{N}$ is infinite by (19.5) above. As to the proof, (19.5) holds in NFI + Union, as NFI+Union $\equiv$ NF, and NF proves (19.5) according to a famous result of Specker (1953). On the other hand, NFI + ¬Union implies (19.5) by Crabbè 1982. The claims (19.4), (19.3) with the Peano axioms are provable in NFI ((19.7) requires the second part of (19.5)).

Definition 3.3 (Homogeneous pairing; Rosser 1953)

$$\phi(a) = \{y \mid y \in a \wedge y \notin \mathcal{N}\} \cup \{y+1 \mid y \in a \wedge y \in \mathcal{N}\};$$

$$\theta_1(a) = \{\phi(x) \mid x \in a\};$$
$$\theta_2(a) = \{\phi(x) \cup \{0\} \mid x \in a\};$$
$$(a, b) = \theta_1(a) \cup \theta_2(b);$$
$$Q_1(a) = \{z \mid \phi(z) \in a\};$$
$$Q_2(a) = \{z \mid \phi(z) \cup \{0\} \in a\}$$

The definitions above are (at most) loosely predicative and hence the universe of sets is closed under the corresponding operations, provably in NFI.

We below exploit the fact that Quine's pairing operation is $\subseteq$-monotone in both arguments: indeed, the definition of (a, b) is positive in a, b[8].

Lemma 3.4 *We have, provably in NFI:*

(i) $\phi(a) = \phi(b) \to a = b$;
(ii) $0 \notin \phi(a)$;
(iii) $\theta_i(a) = \theta_i(b) \to a = b$, *where* $i = 1, 2$;
(iv) $(x, y) = (u, v) \to x = u \wedge y = v$.
(v) the map $x, y \longmapsto (x, y)$ *is surjective and* $\subseteq$*–monotone in each variable, i.e.*

$$x \subseteq u \wedge y \subseteq w \ \to \ (x, y) \subseteq (u, w) \tag{19.8}$$

The proof hinges upon the properties of $\mathcal{N}$ and the successor operation (Rosser 1953).

Lemma 3.5 *(Fixed point) Let* $A(x, a)$ *be a formula which is positive in* a. *Assume that*

$$\Gamma_A(a) = \{x \mid A(x, a)\}$$

is loosely predicative, where x, a *are given types* i, $i + 1$ *respectively. Then* NFI *proves the existence of a set* c *of type* $i + 1$, *such that*:

- $\Gamma_A(c) \subseteq c$;
- $\Gamma_A(a) \subseteq a \ \Rightarrow c \subseteq a$.

The proof is standard: observe that the set

$$c := \{x \mid \forall d(\Gamma_A(d) \subseteq d \to x \in d)\}$$

is loosely predicative.

[8] We recall that a formula $A(x, a)$ is positive in a if every free occurrence of a in the negation normal form of A is located in atoms of the form $t \in a$, which are prefixed by an even number of negations and where $a \notin FV(t)$.

19.3.2 Generating Truth

We use Quine's pairing for representing logical constructors, and the fixed point lemma 3.5 to interpret the truth predicate.

Definition 3.6

$$\begin{aligned}
\dot{\neg}x &:= (0, x);\\
x\dot{\wedge}y &:= (1, (x, y));\\
\dot{\forall}f &:= (2, f);\\
\dot{\in}xy &:= (3, (x, y));\\
\dot{=}xy &:= (4, (x, y))
\end{aligned}$$

Of course, the number labels above are natural numbers in the sense of lemma 3.2. We also write $[x = y]$ for $(\dot{=}xy)$. If $\{x\}$ denotes the singleton, we let

$$[x \in y] := \dot{\in}\{x\}y = y \cdot x$$

Under the dot-application, the universe of sets becomes an applicative structure. $y \cdot x$ is stratified only if y and x are given the types $i + 1$ and i (respectively), and the result of applying y to x is one greater than the type of x.

We now model the Kripke–Feferman notion of self-referential truth within the abstract framework of Quine's set theory. First of all, in analogy with the notion of P-form, define

$$\begin{aligned}
Pfr(x) \Leftrightarrow \exists u \exists v (x = [u \in v]) \vee \exists z (x = \dot{\neg} z \vee x = \dot{\forall} z) \vee\\
\vee \exists w_1 \exists w_2 ((x = [w_1 = w_2]) \vee (x = w_1 \dot{\wedge} w_2))
\end{aligned}$$

$Pfr(x)$ is stratified (assign 1 to x, v, z, w_1, w_2 and 0 to u). The truth predicate W is introduced as the fixed point of a stratified positive (in a) operator $\mathcal{T}(x, a)$, which encodes the recursive clauses for partial self-referential truth and is given by the formula

$$\begin{aligned}
&\exists y\ (x = \dot{\neg} y \wedge \neg Pfr(y)) \vee\\
\exists u \exists v \exists w\ &[(x = [u \in v] \wedge u \in v) \vee\\
&\vee (x = \dot{\neg}[u \in v] \wedge \neg u \in v) \vee\\
&\vee (x = [v = w] \wedge\ v = w) \vee\\
&\vee (x = [\neg v = w] \wedge\ \neg v = w) \vee\\
&\vee (x = \dot{\neg}\dot{\neg} v \wedge\ v \in a) \vee\\
&\vee (x = v \dot{\wedge} w \wedge\ v \in a \wedge w \in a) \vee\\
&\vee (x = \dot{\neg}(v \dot{\wedge} w) \wedge\ (\dot{\neg} v \in a \vee\ \dot{\neg} w \in a)) \vee
\end{aligned}$$

$$\vee (x = \dot{\forall} v \wedge \forall z(v \cdot z \in a)) \vee$$
$$\vee (x = \dot{\neg}\dot{\forall} v \wedge \exists z(\dot{\neg} v \cdot z \in a))]$$

Clearly $\Psi(a) := \{x \mid \mathcal{T}(x, a)\}$ is $\subseteq$-monotone in a and is predicative: it receives type 2 once we assign type 0 to u, z, type 1 to x, v, w, type 2 to a.

Definition 3.7

$$Cl_T(a) := \forall x(\mathcal{T}(x, a) \to x \in a)$$
$$W := \{x \mid \forall a(Cl_T(a) \to x \in a)\}$$

The fixed point lemma 3.5 immediately implies:

Proposition 3.8 NFI *proves:*

1. $\exists y(y = W)$;
2. $\forall a(\mathcal{T}(a, W) \to a \in W)$;
3. $Cl_T(a) \to W \subseteq a$.

Remark 4 The interpretation of the truth predicate thus requires *an inductive definition over the universe which still yields a set, i.e. an object of the universe*. This makes essential use of the *peculiar impredicative features* of NFI. If we should try to carry out such definition over the standard set theoretic universe of ZFC, T would result in a *proper* class, and hence we should be forced to apply an impredicative theory of classes à la Morse-Kelley.

Definition 3.9 We inductively (and simultaneously) specify a translation $(\text{–}) \mapsto (\text{–})^\nu$ of terms and formulas of SFT into NF:

$$x^\nu := x$$
$$pred(t, s) = [s^\nu \in t^\nu]$$
$$T(t)^\nu = t^\nu \in W$$
$$tr(t)^\nu = [t^\nu \in W]$$
$$(t = s)^\nu = (t^\nu = s^\nu)$$
$$(id(t, s))^\nu = [t^\nu = s^\nu]$$
$$(A \wedge B)^\nu = A^\nu \wedge B^\nu$$
$$and(t, s)^\nu = t^\nu \dot{\wedge} s^\nu$$
$$(\neg A)^\nu = \neg A^\nu$$
$$neg(t)^\nu = \dot{\neg} t^\nu$$
$$(\forall x A)^\nu = (\forall x A^\nu)$$
$$[x|A]^\nu = \{x|A^\nu\}$$
$$\mu y.r(y, \vec{x})^\nu = \{u|\forall z(r^\nu(z, \vec{x}) \subseteq z \to u \in z)\}$$

$$all(t)^\nu = \dot{\forall} t^\nu$$

If $\vec{x} := x_1, \ldots, x_n$, $\vec{y} := y_1, \ldots, y_n$, $\vec{x} \subseteq \vec{y}$ means $x_1 \subseteq y_1, \ldots, x_n \subseteq y_n$. A term $t(\vec{x})$ is monotone if $\vec{x} \subseteq \vec{y}$ implies $t(\vec{x}) \subseteq t(\vec{y})$.

Lemma 3.10

(i) If A (t) is a stratified formula (term) of SFT*, then A^ν (t^ν) is a stratified formula (term) of* NF*, under the same type assignment to variables and terms of A. Furthermore, if $t(\vec{x})$ is stratified operative in $\vec{x}$, then $t(\vec{x})^\nu$ is stratified monotone in $\vec{x}$.*

(ii) If A is a stratified formula of SFT*,* NF *proves:*

$$A^\nu[x := u] \leftrightarrow [A^\nu[x := u]] \in W \tag{19.9}$$

Proof As to (i), proceed by simultaneous induction on the definition of *term*, *term operative in a given parameter* and *formula*.

If t is a variable, the claim is trivial. If $A := T(t)$ is stratified, then so is t; hence by IH t^ν is stratified as well as $t^\nu \in W$.

Let $A := t = s$ be stratified. Then so are t^ν and s^ν. But this implies that $(t = s)^\nu \equiv t^\nu = s^\nu$ is stratified.

Let $t^\nu := (pred(s, r))^\nu$ be stratified. Then $t^\nu = [r^\nu \in s^\nu]$ is stratified too, since by IH the type assignment is preserved by IH. The cases where t (A) is built up by means of and, id, all, tr ($\wedge, \forall$) are straightforward by IH.

If $t := [x|A]$, then A^ν is stratified, whence $t := \{x|A^\nu\}$ is stratified.

Let $t := \mu y.r(y, \vec{x})$ be stratified operative in $\vec{x}$. Then $r(y, \vec{x})$ is stratified and operative in $y, \vec{x}$. Hence $(\mu y r(y, \vec{x}))^\nu = \{u|\forall z(r^\nu(z, \vec{x}) \subseteq z \rightarrow u \in z)\}$ is stratified and by IH $r^\nu(y, \vec{x})$ is monotone in y and $\vec{x}$. Hence $(\mu y r(y, \vec{x}))^\nu$ is monotone in $\vec{x}$, i.e., if $\vec{a} \subseteq \vec{b}$

$$(\mu y r(y, \vec{a}))^\nu \subseteq (\mu y r(y, \vec{b}))^\nu.$$

If $t(\vec{x}) := pred(r, s)$ is stratified operative in $\vec{x}$, r has type one greater than the type of s and r is operative in $\vec{x}$. Hence by IH r^ν is stratified monotone in $\vec{x}$ with type one greater than the type of s^ν, which is also stratified. It follows by definition of the ν-translation, the property 19.8 and Quine's pairing that $t(\vec{x})^\nu$ is stratified and monotone in $\vec{x}$.

Let us check the case where $t(\vec{x})$ is stratified operative in $\vec{x}$ and

$$t(\vec{x}) = id(s(\vec{x}), r(\vec{x}))$$

Then we have to show that, if $\vec{x} \subseteq \vec{y}$, $t(\vec{x})^\nu \subseteq t(\vec{y})^\nu$. By IH we have

$$\vec{x} \subseteq \vec{y} \rightarrow s(\vec{x}) \subseteq s(\vec{y})$$
$$\vec{x} \subseteq \vec{y} \rightarrow r(\vec{x}) \subseteq r(\vec{y})$$

The conclusion follows by 19.8, definition of id and Quine's pairing. The remaining cases when t is built up by means of neg, all are similar.

(ii): by induction on A, applying part (i) and proposition 3.8 on W. We only consider the case of the universal quantifier. Then by applying stratified comprehension in the last step:

$$\begin{aligned}[\forall x A]^\nu \in W &\leftrightarrow \forall u([u \in \{x|A^\nu\}] \in W)\\ &\leftrightarrow \forall u(u \in \{x|A^\nu\})\\ &\leftrightarrow \forall u A^\nu[x := u] \equiv (\forall x A)^\nu\end{aligned}$$

□

The ν-translation induces an interpretation into NF:

Theorem 3.11 If SFT $\vdash A$, then NF $\vdash A^\nu$.

Proof It is enough to prove the ν-translation of the SFT-axioms. We repeatedly use proposition 3.8 and the independence (or injectivity) of the chosen representation for the logical and descriptive symbols (see definition 3.6).

(i) T is total on predication. Consider e.g. the ν-translation of

$$T(pred(y,x)) \vee T(neg(pred(y,x))$$

This amounts to verify

$$([x \in y] \in W \leftrightarrow x \in y) \wedge ([\neg x \in y] \in W \leftrightarrow \neg(x \in y)), \tag{19.10}$$

which in turn follows from the second and third clauses of the inductive definition of W. By 19.10 also

$$[x \in y] \in W \vee [\neg x \in y] \in W \tag{19.11}$$

for every x, y. Hence if we choose $y := W$, we obtain:

$$[x \in W] \in W \vee [\neg x \in W] \in W \tag{19.12}$$

But 19.12 implies the ν-translation of the T-axioms involving tr. The verification of the extant cases ($=$, $\wedge$, $\forall$) is also routine.

(ii) T-consistency. Then we must prove the corresponding ν-translation, i.e.

$$\neg(x \in W \wedge (\dot{\neg}x) \in W)$$

Choose $\psi(x) := \neg((\dot{\neg}x) \in W)$. Then $\{x|\psi(x)\}$ is a set in NFI and it is easy to check:

$$\forall x(\mathcal{T}(x, \{x|\psi(x)\}) \rightarrow \psi(x))$$

The conclusion is a consequence of proposition 3.8, item 3.

(iii) Stratified β-conversion: we want, if A is stratified,

$$(T(pred([x|A], u)))^\nu \leftrightarrow (T[A[x := u]])^\nu$$

By definition of W with proposition 3.8, stratified comprehension and lemma 3.10, we have:

$$\begin{aligned}(T(pred([x|A],u))^\nu &\leftrightarrow [u \in \{x|A^\nu\}] \in W)\\ &\leftrightarrow u \in \{x|A^\nu\}\\ &\leftrightarrow A^\nu[x := u]\\ &\leftrightarrow [A^\nu[x := u]] \in W\\ &\leftrightarrow (T[A[x := u]])^\nu\end{aligned}$$

The remaining β conversion schema is similar.

(iv) Self-reference: let $t(y,\vec{x})$ be stratified operative in $y,\vec{x}$. Then by the lemma 3.10, $t^\nu(y,\vec{x})$ is stratified monotone in y, $\vec{x}$. Hence $\mu y t(y,\vec{x})^\nu = \{u|\forall z(t^\nu(z,\vec{x}) \subseteq z \rightarrow u \in z)\}$ satisfies the due fixed point equation by lemma 3.5.

(v) μ-extensionality: straightforward.

(vi) Logical operators are injective: the ν-translation of the corresponding axioms is sound, simply because the logical operators act as ordered sequence operators, built upon Quine's ordered pair.

(vii) Logical operators are not surjective: in fact there are objects (e.g.$(4,a)$) which differ from $\dot{\neg}x$, $x\dot{\wedge}y$, $\dot{\forall}$ and $\dot{\in}xy$ (use lemma 3.2). The images of the logical operators are trivially disjoint (we use distinct Fregean numbers as labels), and there are objects which are not P-forms, e.g. any ordered pair $(\emptyset,a)$.

□

Remark 5 Observe that the full strength of stratified comprehension is exploited in interpreting predication. Once predication is restricted to loosely stratified (or predicative) formulas, the resulting version of SFT becomes reducible to a consistent subsystem of NF.

Remark 6 Clearly we can try to reverse the embedding: one can define $x \in y$ as $T(pred(y,x))$. Then by the theorem 1.8 it is possible to prove the translation of the NF-stratified comprehension in SFT. However, there is by no means guarantee that SFT proves the translation of the extensionality axiom, and this raises the problem of the consistency strength of SFT.

19.4 Conclusion: Stratified Truth?

Let us try to assess some limits of the theory.

Why stratification? On one hand, that there is a type raising when we move from the mere claim of A to the claim of $T[A]$, can on intuitive grounds be conceded (at least according to the present author). And this is a good reason to pursue the

typed theories of truth[9]. On the other hand, our practice with natural language tends to support the idea that we have to deal with *the* truth predicate, without any further type qualification; hence types ought to be left implicit or possibly avoided. Stratification can be regarded as a way to make both sides coexist. Of course, the awkward aspect is that we do appeal to a theory, which has certain unnatural features. Moreover, the consistency of SFT relies in its full strength upon a discipline – stratification – which is not fully understood, as shown by the yet unsolved problem whether NF be consistent or not.

An additional unsatisfactory point is that the syntactical apparatus of SFT is, at the present stage of formalization, rather complex, and the stratification device is not so transparent as the corresponding explicit typed versions of truth.

Nevertheless, though type-theoretic in essence, SFT allows limited, yet non-trivial forms of self-reference, which are based after all on a semantical construction. And these limitations are apparently essential, in order to preserve consistency. In contrast with usual formal theories of truth, a distinctive feature of SFT is that it allows forms of *direct* self-reference (to make this clear with an example from recursion theory, the second recursion theorem instantiates indirect self-reference, while the first recursion theorem typically supports direct self-reference).

A positive interesting point might be that the compositional axioms of truth receive an unrestricted formalization in SFT, and the truth predicate believes that they are true (in sharp contrast, say, with Kripke-like systems). If we compare SFT with other strong axiomatic systems of truth, we must stress that a high degree of impredicativity is gained. The ground for it is the idea that the basic membership relation is well-defined and given, as made clear by the axiom of well-definedness for truth, and by the semantical clauses in NF governing the operator for inductively defining truth.

As to the relation with the literature, Holmes 2001 explores the possibility that formal semantics is expressed in Quine's NFU, i.e. NFU with urelemente. In particular he shows that the reason why Tarski's argument fails, is not the undefinability of truth, but that the quotation operation becomes type-raising, causing the predicate needed for the 'Tarski sentence' to be unstratified and blocking diagonalization. Now, as already seen, something related happens in our case: the operation for encoding formulas of the form $T(x)$, $\neg T(x)$ is also type raising, and this forbids a form of the Liar leading to inconsistency. Of course, this is at present only a surface analogy. Indeed, a comparative look at Holmes 2001 makes clear a specific limitation of SFT: its truth predicate T is not intended for metamathematical applications, as it is not defined on the inductively defined set of (codes of) sentences of the given *formal language*, say, of NF itself. T can only be applied to objects of the intended universe, which stand for propositions, whatever this means. In other words, the truth notion of SFT is an ontological notion, and is alien to standard semantical arguments, which make use of truth or satisfaction for inductively testing some form of (partial) soundness of the provability tools.

[9] For a thorough critical discussion of the distinction between typed and type-free theories of truth, we send the reader to Halbach 2011, especially part II, and Chaps. 10–11 in part III.

Just as the study of axiomatic theories of truth over standard set theory ZFC has been recently developed (see Fujimoto 2012), the investigation of axiomatic notions of truth over non-standard set theories like NF might be the next reasonable step to the present work.

References

Aczel, P. (1980). Frege structures and the notions of proposition, truth and set. In J. Barwise, H. J. Keisler, K. Kunen (Eds.), *The Kleene Symposium* (pp. 31–59). Amsterdam: North Holland.

Cantini, A. (2004). On a Russellian paradox about propositions and truth. In G. Link (Ed.), *One Hundred Years of Russell's Paradox. Mathematics, logic and philosophy* (pp. 259–284) Berlin: Walter de Gruyter.

Crabbè, M. (1982). On the consistency of an impredicative subsystem of Quine's NF. *The Journal of Symbolic Logic, 47,* 131–136.

Fujimoto, K. (2012). Classes and truths in set theory. *Annals of Pure and Applied Logic, 163*(11), 1484–1523.

Halbach, V. (2011). *Axiomatic theories of truth*. Cambridge: Cambridge University Press.

Holmes, M. R. (1995). The equivalence of NF-style set theories with "tangled" type theories: the construction of ω–models of predicative NF (and more). *The Journal of Symbolic Logic, 60,* 178–190.

Holmes, M. R. (2001). Tarski's theorem and NFU. In C. A. Anderson & M. Zelëny (Eds.), *Logic, meaning and computation. Esssays in memory of Alonzo Church, Synthese Library* (Vol. 305, pp. 469–478). Dordrecht: Kluwer Academic Publishers.

Rosser, J. B. (1953). *Logic for mathematicians*. New York: Mc Graw-Hill.

Specker, E. (1953). The axiom of choice in Quine's New Foundations for Mathematical Logic. *Proceedings of the National Academy of Sciences of the U.S.A., 39,* 972–975. Cambridge: Cambridge University Press.

Russell, B. (1903). *The Principles of Mathematics*. London (reprinted by Routledge, London 1997).

Part VI
Inferentialism and Revisionary Approaches

Chapter 20
Truth, Signification and Paradox

Stephen Read

Abstract Thomas Bradwardine's solution to the semantic paradoxes, presented in his *Insolubilia* written in Oxford in the early 1320s, turns on two main principles: that a proposition is true only if things are wholly as it signifies; and that signification is closed under consequence. After exploring the background in Walter Burley's account of the signification of propositions, I consider the extent to which Bradwardine's theory is compatible with the compositional principles of the distribution of truth over conjunction, disjunction, negation and the conditional.

20.1 A Sophism about Saying That

We find an intriguing sophism about signification, or saying that, in an oration delivered at the University of Cambridge in 1660:

> I wonder whether the proverb 'The donkey and the lyre' is not coined for you alone, else you really are not worthy to be proctor in the schools of the sophists. For a sophist attacks the Proctor like this: 'Whoever says you are an animal says something true; and whoever says you are an ass says you are an animal; so whoever says you are an ass says something true.' 'I fully grant it,' says the Proctor: 'For my auricles' sake I wouldn't dare deny it.' See, then, the Proctor confesses himself to be an ass by auricular confession.[1]

Presented at the Conference on 'Truth at Work', Paris 20–23 June 2011. This work is supported by Research Grant AH/F018398/1 (Foundations of Logical Consequence) from the Arts and Humanities Research Council, UK.

[1] (Raine 1843), p. xii: "Dubito certe annon in te solum cudatur proverbium Asinus ad lyram. At profecto tu non dignus es qui esses procurator in sophistarum scholis: sic enim insurgit quidam sophista contra Procuratorem; 'Qui dicit te esse animal dicit verum; at qui dicit te esse asinum dicit te esse animal ergo, qui dicit te esse asinum dicit verum.' 'Concedo totum,' inquit Procurator: 'non ausus sum negare pro auribus.' Videtis, itaque, Procurator fatetur se esse asinum per confessionem auricularem." I am grateful to my colleague, Professor Sarah Broadie, for her help in trying to

S. Read
University of St Andrews, St Andrews, Scotland
e-mail: slr@st-andrews.ac.uk

T. Achourioti et al. (Eds.), *Unifying the Philosophy of Truth,* Logic, Epistemology, and the Unity of Science 36, DOI 10.1007/978-94-017-9673-6_20

The oration is a typical, if unusually rich, example of academic humour, with its allusion to Aesop's fables, its elaborate pun on auricular confession, and its citation of a sophismatic argument dating back to the thirteenth century. The sophism, in fact, appeared shortly afterwards in Geulincx (1662, IV.ii.16.8; Land 1891, I pp. 451–452), called by him *sophisma splendida*, the splendid or brilliant sophism. Perhaps the anonymous author of the oration had heard Geulincx's lectures in Louvain or Leiden, possibly as a royalist exile during the Commonwealth.[2] The sophism also appeared in the Port-Royal Logic of (Arnauld and Nicole 1662, III 12, p. 279), with a goose (oison, literally 'gosling') taking the place of the ass, and Geulincx himself comments that any falsehood could take the place of 'You are an ass', e.g., 'A white thing is black'.

Apart from this flurry of popularity in the mid-seventeenth century, the sophism appears to survive in few medieval treatises: we find it in the *Abstractiones* of around 1240 of Ricardus Sophista (sometimes therefore referred to as the 'Magister Abstractionum', known only through this work);[3] in the Shorter Treatise of Walter Burley's *De Puritate Artis Logicae* ('On the Essence of the Art of Logic'), written in Paris around 1323;[4] and in John Buridan's *Sophismata* (Buridan 2004, p. 51, 2001, p. 864), written in Paris about a generation later. In the first two cases, it is given as a putative counter-example to what has more recently been entitled 'Suffixing': that whatever follows from the consequent follows from the antecedent.[5]

Burley's response is to distinguish two senses of 'saying that', depending on whether what is said, the *dictum* (e.g., 'that you are an animal'), supposits for an utterance or for a thing. In other words, the *dictum* can be taken in material supposition or in personal supposition, where supposition was, along with signification, one of the medievals' main semantic concepts by which to articulate their theories of meaning, truth and consequence. For example, if I say, 'I am looking at Burley', this is true taking 'Burley' in material supposition, for I am looking at the word 'Burley' as I write it, but it is false in personal supposition, for Burley is long dead and no images of him remain.

---

capture the subtlety of the original Latin in English. The proverb comes from Aesop's Fables. 'Auricular confession' apparently means confession made vocally by the penitent to a priest, as distinct from silently addressed to God. Presumably there's a supposed threat that if he says he's not a donkey then they'll make it true that he's not a donkey by cutting off his (donkey's) ears.

[2] (Geulincx 1662) actually uses the example 'Charles is now king of England': see (Nuchelmans 1988, p. 269).

[3] The *Abstractiones* are as yet unpublished; a preliminary edition can be found online at (Ricardus n.d.).

[4] Burley's treatise was edited in (Burley 1955), and translated into English in (Burley 2000). The counter-example also appears in the treatise on 'Consequences' of 1302 attributed to Burley and edited in (Green-Pedersen 1980, p. 113), but the later passage of the Shorter Treatise is little more than a verbatim repetition of the earlier text. The example is discussed, rather inconsequentially, in (Jacquette 2007a) (which is a shorter version of (Jacquette 2007b)).

[5] (Burley 1955, p. 200 ff.): "Quidquid sequitur ad consequens, sequitur ad antecedens."

Applying the distinction to the sophism, we note that, from the fact that I say, that is, utter the words, 'You are an ass', it does not follow that I utter the words 'You are an animal', so in material supposition the major premise of the argument is false. On the other hand, in personal supposition, just because I say you are an animal (which I might do by saying you are an ass) it does not follow that I say something true:

> But if [the act of saying] takes the *dictum* as its object with reference to things, then the inference 'I say that you are an animal; therefore, I say something true' is not valid, because the antecedent can be true without the consequent. For if I say that you are an ass, I say that you are an animal, insofar as the act of saying takes the *dictum* as its object with reference to things. And yet in saying that you are an ass I am not saying something true.[6]

Setting aside the material interpretation where the *dictum* supposits for an utterance, and concentrating on the case where the *dictum* refers to things (as Burley puts it), it is clear that Burley accepts the inference 'If I say you are an ass, I say you are an animal', and rejects the subsequent inference, 'If I say you are an animal, I say something true'. The principle of Suffixing has been saved, but to understand the response, we need shortly to look more closely at Burley's theories of signification and of truth.

We find a similar diagnosis in Geulincx, in the Magister Abstractionum and in Buridan. Geulincx distinguishes two senses of 'saying that', *dicere formaliter vel expresse*, corresponding to direct speech, where the exact words are used; and *dicere consequenter vel implicite*, roughly indirect speech, committing oneself to every proposition entailed by the original.[7] Then the major premise, 'If I say you are an animal I say something true', is only true if 'say' is taken in direct speech (*formaliter*), in which case the minor premise, 'If I say you are an ass I say you are an animal', is false, true only *dicens consequenter*.[8]

Similarly, the Magister Abstractionum defends suffixing as a necessary maxim, insisting that the inference 'If someone says you are an animal he says something true, so if someone says you are an ass he says something true' is necessary, "since it is superior to say you are an animal than to say you are an ass."[9] But he denies that everyone saying you are an animal says something true, "for this is not valid: that you are an animal is true, so everyone saying you are an animal says something true. It is a fallacy of figure of speech, for there is a change from one mode of supposition into another."[10] This particular fallacy was noted by Aristotle in Chap. 24 of his *De Sophisticis Elenchis*, a very broad class of fallacy. Peter of Spain (De Rijk 1972, pp. 144–145), writing around the same time as the Magister Abstractionum, distinguishes three modes of the fallacy of figure of speech, the third where there is

---

[6] (Burley 1955, p. 205), my own translation.
[7] (Geulincx 1662, II.i.3.2; Land 1891, I p. 238). Cf. (Nuchelmans 1994, p. 94).
[8] (Geulincx 1662, IV.ii.16.9-10) (Land 1891, I pp. 452–453).
[9] "Cum superius sit dicere te esse animal quam dicere te esse asinum."
[10] "Et non ualet: te esse animal est uerum, ergo omnis dicens te esse animal dicit uerum; sed est fallacia figurae dictionis; commutatur enim unus modus supponendi in alium."

a change of supposition. The Magister Abstractionum concedes, as does Burley, that if someone says you are an ass, he says you are an animal. But if he says you are an ass, he says something false. So sometimes if someone says you are an animal, he does not say something true.

To explain this phenomenon, Burley distinguishes subjective truth from objective truth: "for I say that truth in as much as it is subjectively in the mind is none other than some equating (*adaequatio*) of the mind to a true proposition which only has objective being in the mind" (Brown 1973, § 1.27). These terms 'subjective' and 'objective' need to be approached with care when used in medieval texts. As Sir William Hamilton (1863, pp. 806 ff.) noted, the terms underwent a nearly complete reversal of sense during the eighteenth century. For the medievals, something is subjectively in the mind when it is in the mind as a (real) quality of the subject. In contrast, something is objectively in the mind, or has objective being, when it is an object of thought—indeed, that is its etymology, as something "thrown in the way of" thought (*ob-iacere*), as, e.g., the moon is literally thrown in the way of the sun in a solar eclipse. So for Burley, a thought (a mental proposition, existing as a quality subjectively in the mind) is true if it corresponds to a real proposition, a *propositio in re*, existing only objectively in the mind. Indeed, for him, the notion of proposition was four-fold: there is the written proposition, itself a sign of a spoken proposition (writing is a way of recording speech); the spoken proposition is a conventional sign of a mental proposition, from which it derives its signification; but the ultimate significate of the spoken proposition is the real proposition.[11] Whereas the mental proposition is a compounding of concepts (if affirmative, or dividing them if negative), and the spoken proposition is a compounding of spoken terms, the real proposition is a compounding or dividing of real things. Burley cites Averroes with approval when he wrote: "Things are made true by the mind when it divides them from one another or compounds them with one another."[12] For example, consider 'A man is an animal' (*Homo est animal*). There are numerous subjective mental propositions compounding the concepts of man and animal, all of which are true by their correspondence to the one true real proposition which identifies man and animal. It is this real proposition which is signified by the spoken proposition 'A man is an animal', just as the spoken term 'man' signifies man (the animal) and 'animal' signifies animal (the universal):

> Hence I say that the thing signified by 'A man is an animal' does not depend on the mind nor does the truth of this thing, for it would be true even if no mind thought about it . . . I say that to the truth 'A man is an animal' having being outside the mind there correspond many truths having subjective being in the mind, for many thoughts can correspond to the same thing. (Brown 1973, § 1.27)

[11] See, e.g., (Conti 2011, § 4; Cesalli 2007, pp. 190 ff.).

[12] (Brown 1973, § 1.03): "Oratio in mente componitur ex rebus patet per Commentatorem, VI *Metaphysicae*, in fine, qui dicit quod entia vera, cuiusmodi sunt propositiones, facta sunt ab intellectu quando dividit ea ab invicem vel componit ea ad invicem."

One may well be reminded here of Bertrand Russell's early theory of propositions. (Russell 1903, p. 47) wrote: "A proposition, unless it happens to be linguistic (i.e., to be about words) does not contain words: it contains the entities indicated by words." At that time, Russell held an identity theory of truth. Such a theory rejects any correspondence between thought and reality, exemplified by Frege's remark (Frege 1997, p. 327) that if facts and thoughts "did correspond perfectly . . . they would coincide"; "a fact," he says, "is a thought that is true" (p. 342). In his rejection of idealism, Russell proclaimed that in thought we directly apprehend the fact containing the objects in question.

However, in both Frege and Russell, what we apprehend is in the modern sense objective and mind-independent. Burley's account of the real proposition is closer to that articulated recently by Jeffrey King (2007, Chap. 2). King locates the proposition in the relation we create between objects by constructing a sentence in some language whose terms refer to them. Thus for King, as for Burley, the proposition results from a semantic act of ours, compounding or dividing objects with or from one another:

> The facts that are propositions are facts of there being a context and there being *some words* in *some language L* whose semantic values relative to the context are so-and-so occurring in such-and-such way in so-and-so sentential relation that in *L* encodes such-and-such. (King 2007, p. 45)

On Burley's account, what makes a subjective proposition (a thought) true is its correspondence to a true objective, that is, real proposition; and the truth of the real proposition itself is nothing other than things' being as the mind considers them to be. Burley elaborates on Averroes' remark, cited above:

> The mind makes things true by compounding those with one another which are in reality united or dividing those from one another which are in reality divided. For if the mind asserts some things to be the same, then it compounds them with one another; but if it asserts them to be divided then it divides them from one another . . . For when the mind compounds correctly or divides correctly, then there is truth in the mind, and when the mind does not compound correctly or does not divide correctly, as when it compounds those which are in reality divided or divides from one another those which are in reality the same, then there is falsehood in the mind. (Brown 1973, § 1.24)

To return to our sophism: Burley accepts the inference 'If I say you are an ass, I say you are an animal' (talking of things, not of words), but rejects the inference 'If I say you are an animal, I say something true'. We can now understand why he says this. What 'You are an ass' signifies is the real proposition (*propositio in re*) which compounds you and being an ass together. But being an ass necessitates being an animal, as part of its form. Being an animal is a formal consequence of being an ass. (Burley 1955, p. 86) writes in *De Puritate*:

> Formal consequence is of two kinds: one kind holds by reason of the form of the whole structure (*complexio*), . . . Another kind of formal consequence holds by reason of the form of the constituent terms (*incomplexa*), e.g., an affirmative consequence from an inferior to its superior is formal, but holds by reason of the terms.

Hence Burley accepts that signification is closed under consequence, at least, formal consequence. If the mind is compounding (incorrectly) you and being an ass, it is

inevitably compounding you and being an animal, since animal is superior to ass, and being an animal is formally included in being an ass.

It follows, as Burley notes, that it is incorrect to infer from my saying you are an animal that what I am saying is true and that I say something true. For my saying you are an animal may be merely consequent on my incorrectly compounding things which are not united, as when I say you are an ass. Not everyone saying you are an animal says something true. I must compound and divide correctly if I am to say something true. In other words, everything I say must be that way in reality for my (subjective) proposition to be true.

20.2 Truth and the Liar

We find these two claims, that truth requires that things are only or wholly as signified and that signification is closed under consequence, utilized in Thomas Bradwardine's proposed solution to the semantic paradoxes in his *Insolubilia*, composed in Oxford a year or two after the Shorter Treatise of Burley's *De Puritate* was composed in Paris. Bradwardine's main object of attack in this work is Burley's restrictivist solution to the paradoxes, but in other respects, Bradwardine seems to share similar views to Burley's.[13] He sets out two definitions, six postulates and two theorems, among which we read:

> First Definition (D1): A true proposition is an utterance signifying only as things are.
> Second Definition (D2): A false proposition is an utterance signifying other than things are.
> . . .
> Second Postulate (P2): Every proposition signifies or means . . . everything which follows from it . . .
> . . .
> Second Theorem (T2): If a proposition signifies itself not to be true or itself to be false, it signifies itself to be true and is false. (Bradwardine 2010, ¶¶ 6.2–6.4])

Theorem (T2) depends on the fact that propositions can signify (conjunctively) several things; e.g., if a proposition signifies that it itself is not true, it will also signify itself to be true. If a proposition signifies that you are an ass, by (P2) it will also signify that you are an animal. But by (D1), a proposition is true only if things are altogether as it signifies, that is, everything it signifies obtains. Though a proposition might signify that you are an animal, it does not follow that, since you are an animal, it is true. It may also signify that you are an ass, and so not everything it signifies obtains. Similarly, an insoluble such as the Liar, e.g., 'This proposition is false', or Bradwardine's favourite example, Socrates' utterance (only) of 'Socrates says something false', is false. But it does not follow that it is true, since although that is what it signifies, it is not all it signifies (by T2), for it also signifies that it is true, and things cannot be altogether as it signifies, since no proposition is both true and false (by P1: 'Every proposition is true or false,' and not both).

[13] See (Bradwardine 2010, 'Introduction' § 4).

Bradwardine's argument for (T2) runs as follows: suppose some proposition signifies itself not to be true. Then either that is all it signifies, or not. First, suppose that is all it signifies; and suppose that it is not true. Then by (D1), things are not as it signifies, namely, not true, so it is true. That is, its being true follows from its not being true. But it signifies that it is not true, so by (P2) it follows that it signifies that it is true.

So its not being true is not all it signifies: suppose it also signifies that q, say. (There may be many other things it signifies. Let q be their conjunction.) Again, suppose it is not true. Then by (D1), things are not wholly as it signifies, that is, by a De Morgan conversion (P4), either it is true or not-q. So again by (P2), it signifies that either it is true or not-q. But we supposed that it signified that q, and from its either being true or not-q, and q, it follows by Disjunctive Syllogism (P5) that it is true. So once again, by (P2), it follows that it signifies that it is true.

Next, suppose some proposition signifies that it itself is false. If it is false, it's not true, by (P1). So by (P2), it follows that it signifies that it is not true, and so by the above argument, it signifies that it is true.

Finally, we have already noted that things cannot be wholly as any proposition signifies which signifies both that it is not true and that it is true. So by (D2), any such proposition is false. Thus any paradoxical proposition, such as the Liar, is false and not true. So too with Socrates' utterance of 'Socrates says something false'. For if that is his only utterance, then by (P2) it signifies that his only utterance is false, that is, that it itself is false. So by (T2), it also signifies that it is true, and is false.

The use of Disjunctive Syllogism (P5) in Bradwardine's argument might make one think it depends on taking the disjunction truth-functionally. Not so; suppose s signifies that s is not true, and perhaps other things too, call them q. It follows by (D2) that if s is false, then if it's not q that fails to obtain, it must be the falsity of s that fails, that is, if s is false and q holds, s must be true. If the conditionals here are strict or relevant conditionals, for example, the inference follows.[14] So since s signifies that s is false and q, it follows that s signifies that s is true, by (P2).

Bradwardine has made great play here with his second postulate, (P2). I've interpreted it as a closure postulate, that signification is closed under consequence, and that is certainly how Bradwardine repeatedly uses it. But it is not quite how he states it. He says "a proposition signifies everything which follows from it," not "from what it signifies". Paul Spade (1981, p. 120) took him more literally, but found that Bradwardine's own reasoning did not then go through. He attributed to Bradwardine the principle (BP), in the form of a schema:

(BP) If p only if q, then P signifies that q,

where what replaces 'P' names the proposition which replaces 'p'. With a naming function $\ulcorner\ \urcorner$, we could express this as

[14] $p \rightarrow (q \rightarrow r)$ entails $(p \wedge q) \rightarrow r$ (but not vice versa) in the logics of strict implication **S2** (and stronger) and of relevant implication, **R**: on the latter, see, e.g., (Anderson and Belnap 1975, § 29.3.1, R30 and R31).

If p only if q, then $\ulcorner p \urcorner$ signifies that q,

Suppose that 'signifies' is closed under consequence, that is,

(P2) If p only if q, then if s signifies that p then s signifies that q.

As an instance we have:

If p only if q, then if $\ulcorner p \urcorner$ signifies that p then $\ulcorner p \urcorner$ signifies that q.

Then (BP) follows from (P2) together with the plausible assumption that $\ulcorner p \urcorner$ signifies that p. On that assumption, (P2) entails (BP), but is stronger than (BP) since it can be used, and was used, in Bradwardine's proof of (T2), while (BP) cannot. On the other hand, (BP) entails that $\ulcorner p \urcorner$ signifies that p, which (P2) does not. Bradwardine nowhere states this in general, though his practice certainly assumes it.

Bradwardine restricts (T2) to claiming only that every proposition which signifies that it is not true signifies its own truth. John Buridan (in his early writings) and Albert of Saxony, some 20 or 30 years later, claimed that every proposition whatever ("*omnis propositio mundi*") signifies its own truth.[15] (Spade 1981, p. 120 n.17) observes that Burley does so too, not only in the Longer Treatise of his *De Puritate*, written shortly after Bradwardine's *Insolubilia*, but also in his own treatise on insolubles, written in 1302.[16] Spade also finds the claim in John Duns Scotus in his *Quaestiones in duos libros Perihermeneias*, written 30 years earlier, around 1295,[17] and even earlier in Bonaventure:[18]

> An affirmative proposition makes a double assertion: one in which the predicate is affirmed of the subject and the other in which the proposition is asserted to be true. By virtue of the first assertion an affirmative proposition is differentiated from a negative one, which denies the predicate of the subject. So far as the second assertion is concerned, however, affirmative and negative statements agree since they both assert something to be true. Contradiction is concerned not with the second type of assertion but with the first. For when it is stated that no truth exists, insofar as it negates the predicate of the subject this proposition does not imply its opposite, *viz* that some truth exists. But to the extent that it asserts this to be true, it does entail that some truth exists.

---

[15] Buridan famously rejected this claim in his later writings, saying not that every proposition signifies its own truth, but that it virtually implies it. His main reason for doing so was his rejection of the notion of the complexly signifiable (*complexe significabile*), what is signifiable only by a *complexum*, that is, a proposition. Given the similarity between this doctrine and Burley's notion of the real proposition, Bradwardine would not share Buridan's worries, if he did indeed accept Burley's semantic account. See, e.g. (Klima 2009, Chaps. 9–10, esp. § 10.2).

[16] See (Burley 1955, p. 25, 2000, p. 108): "Quaelibet propositio asserit seipsam esse veram," and (Roure 1970, p. 272): "Quilibet dicens asserit suum dictum esse verum".

[17] See Lib. I Questiones 7–9 § 10 (Andrews et al. 2004, I p. 181): "Quaelibet propositio significat se esse veram, ergo ista 'tu eris albus cras' significat se esse veram. Antecedens patet, quia ad omnem propositionem veram sequitur suum dictum fore verum. Similiter contradictorium affirmativae ut ista 'tu non eris albus' infert hanc '"te non fore album" est verum'. Utrumque igitur contradictoriorum in illis de futuro significat se esse determinate veram." This occurs as part of an objection, but in his response Scotus does not question the basic principle.

[18] *Quaestiones disputatae de mysterio Trinitatis*, q1 a1, translated in (Wippel and Wolter 1969, pp. 310–311).

None of them provides much, if anything, in the way of an argument for this claim. Geulincx gives this argument:[19] any proposition says things to be (*dicit esse*), indeed, it says them to be what it says them to be. But things being as it says them to be is for it to be true. So it says itself to be true. (Nuchelmans 1988, pp. 280–281) comments that the sense of 'say' here must be 'dicere consequenter' (see above, § 20.1).

In fact, that every proposition signifies its own truth follows straightforwardly (though unremarked by Bradwardine) from his postulate (P2). Suppose some proposition s signifies that q_1, q_2 and so on. Let q be their conjunction—everything s signifies. Then by (D1), s is true if and only if everything s signifies obtains, that is, iff q. In particular, if q then s is true. But s signifies that q. So by (P2), s signifies that s is true.

One might now worry about a circularity. To show that s is true, we need to check that everything it signifies obtains. One of the things it signifies is that s is true. So to check that s is true we need first to check that s is true. That threatens to open up a vicious regress. The objection is ill-founded, however. (D1) tells us that s is true iff everything s signifies obtains. So to check that s is true we need to check that everything it signifies obtains, and of course, that condition is equivalent to s's being true. So we need to check that s is true. But that is no more than we are doing. There is no regress here, just a repetition of the task we are set.

More worrying, perhaps, is the open-ended nature of the condition: to check that s is true, check that everything s signifies obtains, where what s signifies is everything entailed by what it signifies (by P2). Of course, having checked that one thing that s signifies obtains, one can be sure that everything entailed by that also obtains. But there may well be other things s signifies that are not entailed by what has been checked. For example, suppose s is 'You are an ass'. That signifies that you are an ass, that you are an animal, that you are alive and many other consequences. Checking that you are an animal confirms that you are alive and so on. But it does not confirm that s is true. One has further to check that you can bray, that you have long ears, in short, that you are an ass. As soon as one of these significates is found not to hold, we know by (D2) that s is false. But if s is true, this check could in principle, and perhaps in practice, never be completed. Does Bradwardine's account of truth mean that no proposition is ever true?

First, the objection confuses the ontological criterion for s's being true with the epistemological condition for knowing that s is true. There is nothing problematic about the first being indefinitely, even infinitely, complex. But even the epistemology is confused. We can know, and be certain, that Brownie is a donkey, even if we have not checked implausible subterfuges, that Brownie is not a heavily disguised CIA spy, or a Martian robot or whatever. That Brownie is a donkey entails that he is not a robot or a disguised human being. Knowledge is fallible. If Brownie turns out to be

[19] (Geulincx 1663, Chap. 1; Land 1891, II p. 25): "Sit enunciatio quaecunque, nempe A. Dico quod A dicat se esse veram. Quia A dicit esse, et dicit esse quod dicit esse, sed esse id, quod A esse dicit, est A esse veram. Cum igitur A prius dicat, dicit etiam posterius, seu A dicet A, id est seipsam, veram esse." Cf. (Geulincx 1662, II.i.1.4; Land 1891, I p. 234) and (Nuchelmans 1988, p. 280). Similar proofs are found in Albert of Saxony and John Buridan. See, e.g., (Read 2002, § 3).

a robot, then we were badly misled, did not know he was a donkey, and he wasn't. Fortunately, we are not faced with such tricks too often, and we do know (defeasibly) that Brownie is a donkey.

In claiming that the Liar proposition, indeed, perhaps every proposition, signifies its own truth in addition to what it more obviously signifies, Bradwardine is claiming that such propositions are exponible. Exponible propositions were defined by the medievals as propositions which, though grammatically simple, were logically complex.[20] A standard example was an exclusive proposition such as 'Only a man is running', apparently a simple subject-predicate proposition, but implicitly complex, to be "expounded" or analysed as 'A man is running and nothing other than a man is running'. (Burley 1955, p. 134) points out that in consequence, the negation of an exclusive proposition is implictly disjunctive: 'Not only Socrates is running' is equivalent to the disjunction 'Socrates is not running or something other than Socrates is running'. The same is true of the Liar. Since it is implicitly conjunctive, its negation is implicitly disjunctive: to contradict, e.g., 'This proposition is false', which signifies both that it is false and that it is true, we need to assert either that it is true or that it is false.

20.3 Compositionality

Bradwardine's approach to the Liar belongs to a class of solutions that revise and constrain the theory of truth rather than the underlying logic. Bradwardine's logic is robust and orthodox, endorsing such principles as Bivalence (P1), the De Morgan equivalences (P4) and Disjunctive Syllogism (P5). He is also committed to what Tim Maudlin (2004, p. 112) calls Downward T-Inference and Hartry Field (2008, p. 121) calls T-OUT, embodied in one half of Tarski's T-scheme:

> If $\ulcorner p \urcorner$ is true then p.

This is a special case, given that $\ulcorner p \urcorner$ signifies that p, of the more general rule which Bradwardine accepts, namely, that 's is true' implies that anything that s signifies obtains, which follows from (D1). What Bradwardine rejects is Upward T-Inference (Field's T-IN):

> If p then $\ulcorner p \urcorner$ is true.

It does not follow that a proposition is true merely from the fact that something that it signifies obtains—though one may defeasibly infer it. I may infer 'Brownie is an ass' is true from the fact that Brownie can bray—though I may have to retract the claim when I find a voice recorder hidden under the cleverly life-like fake skin. I may infer that the Liar is true from the fact that it is false—but again, have to retract my claim when I see Bradwardine's demonstration that the Liar signifies both that it is false and that it is true.

[20] See, e.g., (Yrjönsuuri 1993).

Bradwardine's theory shares with other theories which reject T-IN, such as Kripke's and Maudlin's, a difficulty in justifying the standard compositional, or distributive, principles for conjunction and disjunction. (Maudlin 2004, p. 144) notes the problem:

> The absence of the Upward Inferences is a severe constraint. In essence, one loses information when using the Downward Inferences, and has no means of semantic ascent again. For example, whenever it is permissible to assert that a conjunction is true, it is permissible to assert that each conjunct is true, but the system as we have it does not allow this inference. From the claim that the conjunction is true one can assert the conjunction itself (by the Downward T-Inference), and hence can assert each conjunct (by & Elimination), but since there is no Upward T-Inference one cannot assert that the conjunct is true.

His response is bold. He simply adds the requisite compositional principles as an axiom. So did Bradwardine. He writes:

> Sixth Postulate (P6): If a conjunction is true each part is true and conversely; and if it is false, one of its parts is false and conversely. And if a disjunction is true, one of its parts is true and conversely; and if it is false, each part is false and conversely. (Bradwardine 2010, ¶6.3)

But this seems unsatisfactory. The compositional principles should follow from the theory of truth in conjunction with the meaning of the connectives. Indeed, there is a risk of inconsistency in Maudlin's procedure. Consider the corresponding principle for negation:

> (Neg) If a negation is true, its negated part is false and conversely; and if it is false, its negated part is true and conversely.

The Liar is a counter-example to this. Let L be $\ulcorner L$ is not true$\urcorner$. L is false, but the negated part 'L is true' is also false. Consequently, if one were to add (Neg) to Bradwardine's theory, as a new postulate, the theory would be inconsistent. As we noted, L is implicitly contradictory, to be analysed or expounded as a conjunction, so its contradictory is a disjunction. 'L is true' no more contradicts 'L is not true' than, e.g., 'Only Socrates is running' contradicts 'Only Socrates is not running', or 'The King of France is bald' contradicts 'The King of France is not bald'.[21]

Similarly, the corresponding principle for conditionals runs into trouble with Curry's paradox.[22] To adapt an example from Jean de Celaya, writing around 1500:[23] let C be the conditional 'If C is true then you are an ass', and suppose we adopted the principle:

> (Cond) If a conditional is true then either the antecedent is false or the consequent is true, and conversely; and if it is false, then the antecedent is true and the consequent is false, and conversely.

[21] See (Read 2008, p. 217).
[22] So called from (Curry 1942). Cf. (Geach 1955).
[23] See (Roure 1962, p. 262).

If C is true, then by (D1), if C is true you are an ass, so by absorption,[24] if C is true you are an ass. But you are not an ass and could not be (your essence is incompatible with that of an ass), so C is necessarily false. Now apply (Cond): given that C is false, it follows that it is true that C is true and false that you are an ass. No harm in the second conjunct, but the first conjunct entails that C is true, by T-OUT. Contradiction. So we cannot endorse the compositional principle (Cond), at least not in the form given.

(Cond) makes the conditional truth-functional, so one might consider adapting it to treat conditionals non-truth-functionally, for example:

> (Cond′) If a conditional is true then the truth of the antecedent is incompatible with the falsity of the consequent, and conversely; and if it is false, then the truth of the antecedent is compatible with the falsity of the consequent, and conversely.

If we now apply (Cond′) to the fact that C is false, it follows that its being true that C is true is compatible with the falsity of your being an ass. But anything compatible with a truth could be true. So it could be true that C is true, so C could be true. But we showed that C cannot be true, so once again, we have a contradiction. The compositional principle (Cond′) cannot be accepted either.

No such contradiction follows, however, from (P6), the compositional principles for conjunction and disjunction, despite the fact that there are similar paradoxes for 'and' and 'or', as (Bradwardine 2010, ¶¶8.9.1-2) noted. Let D be the disjunction 'You are an ass or D is false'. Suppose you are an ass or D is false. Since you are not an ass, it follows by (P5) that D is false. But D signifies that you are an ass or D is false, so by (P2), D signifies that D is false. Hence by (T2), D signifies that D is true and so is false. By (P6), given that the disjunction D is false, it follows that it is false that D is false. So D is false and it is false that D is false. But that is not a contradiction, though it may seem surprising. The explanation is that the falsity of D does not suffice to make it true that D is false—to repeat, Upward T-Inference fails in general. 'D is false' entails, by (P6), as we just noted, that it is false that D is false. By (BP), 'D is false' signifies that D is false. So by (P2), 'D is false' signifies that it is false that D is false, whence by (T2), 'D is false' also signifies that 'D is false' is true and so 'D is false' is false.

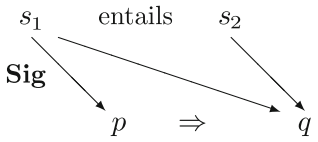
Similarly, let E be the conjunction 'There is a God and E is false'. Then by a similar argument we show that E signifies its own falsehood and so by (T2), E is false. Hence by (P6), one conjunct is false, and it's not the first, so the second, that is, it's false that E is false. It doesn't follow that E is true, for E is false not because it signifies its own falsehood and it's not false, but because it signifies its own truth and it's false that it's true.

One may still be puzzled why (Neg) and (Cond) lead via L and C to contradiction, whereas (P6) does not produce contradiction through D and E. The explanation is that by (P6), the falsehood of a complex proposition implies only the falsehood of one or both components, whereas by (Neg) and (Cond), the falsehood of a complex

[24] So called by (Geach 1955): in symbols, from $p \to (p \to q)$ infer $p \to q$; nowadays usually referred to as "contraction".

proposition implies the truth, or at least the possible truth, of one of its parts, a part that must be false.

Given that they are consistent with Bradwardine's theory, are the compositional principles of the distribution of truth over conjunction and disjunction in fact derivable from the theory together with the standard rules for 'and' and 'or'? First, consider the following diagram:



I have so far interpreted (P2) as saying that if s_1 signifies that p, and p entails q (in symbols, $p \Rightarrow q$), then s_1 signifies that q. Paul of Venice interprets it slightly differently, however. He writes:

> I say that any proposition signifies the significate of any proposition following from it formally ... This is how the common saying, 'Any proposition signifies whatever follows from it', should be understood.[25]

My interpretation follows what we might call the "southern" route in the diagram, Paul's the "northern" route. Arguably, the diagram commutes, and s_1 signifies q whichever route one takes.

Now suppose that some conjunction is true. Then things are however the conjunction signifies. Suppose its first conjunct signifies that, say, p. Then by Paul's principle, the conjunction also signifies that p, since any conjunction entails its first conjunct. But things are however the conjunction signifies. So p. That is, things are however the first conjunct signifies. So the first conjunct is true, and similarly for the second conjunct. Hence, if a conjunction is true, so are each of its conjuncts, and if either is false, and so not true, then the conjunction is not true, but false.

For the converse, we need to generalise (P2) a little further. Recall Bradwardine's proof of (T2). Assuming that the proposition signifying itself not to be true signified something else as well, call it q, Bradwardine showed that it signifies that either it is true or not q. He concluded that it signifies that it is true, since we have assumed that it signifies that q. This does not follow strictly from (P2). Rather, we need to know that if a proposition signifies that p_1 and signifies that p_2, it signifies that p_1 and p_2.[26] This may seem obvious. In the present context, we need a somewhat similar converse principle, namely, that whatever a conjunction signifies is entailed (jointly) by things signified by each conjunct.

Now suppose each conjunct of some conjunction is true. Then by our new principle, whatever the conjunction signifies is entailed by something signified by each

---

[25] (Del Punta and McCord Adams 1978, p. 74), my translation.

[26] We can capture this in a generalisation of (P2): if s signifies that p_1 and signifies that p_2, and p_1 and p_2 (jointly) entail r, then s signifies that r.

conjunct. But since the conjuncts are true, each of those obtains, and so whatever the conjunction signifies must obtain too. So things are however the conjunction signifies, and so a conjunction is true whenever each conjunct is true. Thus we have established the compositionality principle for conjunctions which Bradwardine states in (P6), that if a conjunction is true, each conjunct is true and conversely, and if it is false, at least one conjunct is false and conversely.

What of the distribution of truth over disjunction? Take a disjunction, and suppose one disjunct is true, that is, whatever the disjunct signifies obtains. By Paul's principle, the disjunct signifies whatever the disjunction signifies, since a disjunction is entailed by each disjunct. So whatever the disjunction signifies obtains, and so the disjunction is true.

Conversely, suppose each disjunct is false. Then something each disjunct signifies fails to obtain. It's reasonable to assume that a disjunction signifies the disjunction of anything its disjuncts severally signify. So the disjunction signifies something disjunctive neither part of which obtains, and so which does not obtain as a whole. So the disjunction is also false. Contraposing, if a disjunction is true then one or other disjunct is true. Putting it all together, we have the compositional principle for disjunction that Bradwardine states in (P6): a disjunction is true if at least one disjunct is true and conversely; and a disjunction is false if both disjuncts are false and conversely.

20.4 Conclusion

Implicit in their responses to the sophism 'If I say you are an ass, I say you are an animal, and if I say you are an animal I say something true, so if I say you are an ass I say something true, so you are an ass', is a shared acceptance by the Magister Abstractionum and Walter Burley that saying that, or signification, is closed under at least some form of consequence. That closure principle lies at the heart of Thomas Bradwardine's idea for solving the semantic paradoxes, together with the idea that a proposition is true only if things are wholly as it signifies, that is, only if everything it signifies obtains.

Bradwardine uses the closure principle to show that any proposition which signifies its own falsity also signifies its own truth, and so not everything it signifies can obtain, whence it must be simply false. In fact, it turns out that the closure principle implies that any proposition whatever signifies its own truth. But this does not mean that nothing is true. Establishing that something is true may be defeasible, in that one cannot check that everything it signifies obtain, but knowledge and certainty are still possible. What might at first seem more problematic is the failure of (T-IN), that if p then $\ulcorner p \urcorner$ is true. This must fail in general since the Liar, 'This proposition is false', is an immediate counter-example. As a result, the usual derivation of the compositionality principles distributing truth over negation, conjunction, disjunction and the conditional can no longer be completed. In fact, the Liar shows that truth does not

distribute over negation, and Curry's paradox shows that it does not distribute over the conditional either.

The compositional principles for conjunction and disjunction, however, can be derived by invoking other persuasive principles, including an alternative interpretation of the closure postulate. Bradwardine's solution is thus found to preserve those truth principles which are unaffected by the paradoxes, without sacrificing any logical principles, and so constitutes an attractive and viable solution.

References

Anderson, A., & Belnap, N. (1975). *Entailment: The logic of relevance and necessity* (Vol. 1). Princeton: Princeton University Press.

Andrews, R., Etzkorn, G., Gál, G., Green, G., Noone, T., Plevano, R., Traver, R., & Wood, R. (2004). *B. Ioannis Duns Scoti: Opera Philosophica*. St. Bonaventure: The Franciscan Institute.

Arnauld, A., & Nicole, P. (1662). *La logique ou l'art de penser*. Paris: Charles Savreux.

Bradwardine, T. (2010). *Insolubilia* (trans: S. Read). Leuven: Peeters.

Brown, S. (1973). Walter Burley's middle commentary on Aristotle's *Perihermenias*. *Franciscan Studies, 33,* 42–134.

Buridan, J. (2001). *Summulae de Dialectica* (trans: G. Klima). New Haven: Yale University Press.

Buridan, J. (2004). *Summulae de Practica Sophismatum* (Ed. F. Pironet). Turnhout: Brépols.

Burley, W. (1955). *De Puritate Artis Logicae Tractatus Longior, with a revised edition of the Tractatus Brevior*. St. Bonaventure: The Franciscan Institute.

Burley, W. (2000). *On the purity of the art of logic* (trans: P. V. Spade). New Haven: Yale University Press.

Cesalli, L. (2007). *Le réalisme propositionnel*. Paris: Vrin.

Conti, A. (2011). Walter Burley. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. (Fall 2014 edition) http://plato.stanford.edu/archives/fall2014/entries/burley/.

Curry, H. (1942). On the inconsistency of certain formal logics. *Journal of Symbolic Logic, 7,* 115–117.

De Rijk, L. (1972). *Peter of Spain: Tractatus*. Assen: Van Gorcum.

Del Punta, F., & McCord Adams, M. (1978). *Pauli Veneti Logica Magna: Secunda Pars, Tractatus de Veritate et Falsitate Propositionis, et Tractatus de Significato Propositionis*. Oxford: Oxford University Press for the British Academy.

Field, H. (2008). *Saving truth from paradox*. Oxford: Oxford University Press.

Frege, G. (1997). Thoughts. In M. Beaney (Ed.), *A Frege Reader* (pp. 325–345). Oxford: Blackwell.

Geach, P. (1955). On *insolubilia*. *Analysis, 15,* 71–72.

Geulincx, A. (1662). *Logica fundamentis suis, a quibus hactenus collapsa fuerat, restituta*. Leiden: Henricus Verbiest. (Reprinted in (Land 1891, Vol. I, pp. 165–454)).

Geulincx, A. (1663). *Methodus inveniendi argumenta*. Leiden: Issacus de Wael. (Reprinted in (Land 1891, vol. II, pp. 1–111)).

Green-Pedersen, N. J. (1980). Walter Burley's *De consequentiis*: An edition. *Franciscan Studies, 40,* 102–166.

Hamilton, W. (1863). On presentative and representative knowledge. In W. Hamilton (Ed.), *The works of Thomas Reid* (Vol. 2, pp. 804–815). Edinburgh: MacLachlan and Stewart.

Jacquette, D. (2007a). Burleigh's fallacy. *Philosophy, 82,* 437–448.

Jacquette, D. (2007b). Deductivism and the informal fallacies. *Argumentation, 21,* 335–347.

King, J. (2007). *The nature and structure of content*. Oxford: Oxford University Press.

Klima, G. (2009). *John Buridan*. Oxford: Oxford University Press.

Land, J. (Ed.). (1891). *Arnold Geulincx: Opera Philosophica* (3 volumes). The Hague: Martinus Nijhoff.

Maudlin, T. (2004). *Truth and paradox*. Oxford: Oxford University Press.
Nuchelmans, G. (1988). Geulincx' containment theory of logic. Medelingen van de Afdeling Letterkunde. Amsterdam: Koninklijke Nederlandse Akademie van Wetenschappen, NS51, no.8, pp. 266–317.
Nuchelmans, G. (1994). Walter Burleigh on the conclusion that you are an ass. *Vivarium, 32,* 90–101.
Raine, J. (Ed.). (1843). *The correspondence of Dr. Matthew Hutton, Archbishop of York*. London: Publications of the Surtees Society 17. J.B. Nichols and Son.
Read, S. (2002). The liar paradox from John Buridan back to Thomas Bradwardine. *Vivarium, 40,* 189–218.
Read, S. (2008). Further thoughts on Tarski's T-scheme and the Liar. In J. Rahman, T. Tulenheimo, & E. Genot (Eds.), *Unity, truth and the liar: The modern relevance of medieval solutions to the liar paradox* (pp. 205–225). Berlin: Springer.
Ricardus. (n.d.). Abstractiones (Ed. P. King). http://www.hs-augsburg.de/~harsch/Chronologia/Lspost13/RicardusSophista/ric abst.html.
Roure, M.-L. (1962). Le traité "Des Propositions Insolubles" de Jean de Celaya. *Archives d'Histoire Doctrinale et Littéraire du Moyen Age, 29,* 235–336.
Roure, M.-L. (1970). La problématique des propositions insolubles au XIII^{e} siècle et au début du XIV^{e}, suivie de l'édition des traités de W. Shyreswood, W. Burleigh et Th. Bradwardine. *Archives d'Histoire Doctrinale et Littéraire du Moyen Age, 37,* 205–326.
Russell, B. (1903). *The principles of mathematics*. London: George Allen & Unwin.
Spade, P. (1981). 'Insolubilia' and Bradwardine's theory of signification. *Medioevo, 7,* 115–134.
Wippel, J., & Wolter, A. (Eds.). (1969). *Medieval philosophy*. New York: The Free Press.
Yrjönsuuri, M. (1993). *Expositio* as a method of solving sophisms. In S. Read (Ed.), *Sophisms in medieval logic and grammar* (pp. 202–216). Dordrecht: Kluwer.

Chapter 21
Vagueness, Truth and Permissive Consequence

Pablo Cobreros, Paul Egré, David Ripley and Robert van Rooij

Abstract We say that a sentence A is a *permissive consequence* of a set of premises Γ whenever, if all the premises of Γ hold up to some standard, then A holds to some weaker standard. In this paper, we focus on a three-valued version of this notion, which we call *strict-to-tolerant* consequence, and discuss its fruitfulness toward a unified treatment of the paradoxes of vagueness and self-referential truth. For vagueness, st-consequence supports the principle of tolerance; for truth, it supports the requisit of transparency. Permissive consequence is non-transitive, however, but this feature is argued to be an essential component to the understanding of paradoxical reasoning in cases involving vagueness or self-reference.

21.1 Introduction

According to the standard view of logical consequence, a sentence A is said to follow from a set of premises Γ if it is impossible for all the premises of Γ to be true together and for the conclusion A *not to be true*. Alternatively, a sentence A may be said to follow from a set of premises Γ if it is impossible for all of the premises of Γ to be true together and for the conclusion A to be *false*. In a bivalent setting, the foregoing definitions coincide, because falsity and non-truth coincide. When the underlying space of truth values is larger, however, these two definitions can come apart, and the

P. Egré
Institut Jean Nicod, Ecole Normale Supérieure-PSL Research University, Paris, France
e-mail: paul.egre@ens.fr

P. Cobreros
Universidad de Navarra, Pamplona, Spain
e-mail: pcobreros@unav.es

D. Ripley
University of Connecticut, Storrs, CT, USA
e-mail: davewripley@gmail.com

R. van Rooij
University of Amsterdam, Amsterdam, Netherlands
e-mail: R.A.M.vanRooij@uva.nl

T. Achourioti et al. (Eds.), *Unifying the Philosophy of Truth,* Logic, Epistemology, and the Unity of Science 36, DOI 10.1007/978-94-017-9673-6_21

latter definition, in particular, becomes potentially more permissive than the former, allowing for more schemata to count as valid inference patterns.

Our goal in this paper is to show the interest of such a notion of permissive consequence, whereby consequence is no longer defined as the preservation of some designated truth-value (or set thereof) from premises to conclusion, but rather, as the enlargement of the set of designated truth-values, or as a weakening of standards when going from premises to conclusion (see, in order of appearance, (Nait-Abdallah 1995), (Bennett 1998), (Frankowski 2004), (Zardini 2008), (Smith 2008), (van Rooij 2012), (Cobreros et al. 2012b)). More specifically, we intend to show the fruitfulness of this notion for the prospect of getting a unified treatment of the paradoxes of vagueness and of the paradoxes of self-referential truth. The notion of permissive consequence we are concerned with was originally introduced with an aim to solving the sorites paradox (see (Zardini 2008)), and in order to account for the semantics and pragmatics of vague predicates more generally (see (vanRooij 2012), (Cobreros et al. 2012b), (Cobreros et al. 2012a)). It soon became apparent, however, that it could be applied in a natural way to the treatment of the semantic paradoxes and in particular to the Liar Paradox (see (Ripley 2012), (Cobreros et al. 2013)).

The quest for a unified treatment of vagueness and self-referential truth has been viewed as both natural and desirable by several authors before us (see in particular (McGee 1991), (Tappenden 1993), (Field 2003), (Colyvan 2009), (Priest 2010)). One of the reasons for that is that both the paradoxes of vagueness and the semantic paradoxes appear to put into question two central laws of classical logic, namely the law of excluded middle and the principle of non-contradiction. In the case of vagueness, borderline cases often appear as semantically indeterminate cases, cases of which it is neither determinately true to say that the predicate holds, nor determinately false to assert it. The same appears to hold of our use of the word 'true'. As McGee put it, there are sentences "that the rules of our language, together with the empirical facts, determine to be definitely true; sentences that the rules of our language, together with the empirical facts, determine to be definitely not true; and sentences that are left unsettled" (McGee 1991, p. 6). Among those, Liars and Truth-Tellers figure most prominently.

For vagueness as well as for truth, consequently, three-valued logic appears as a natural and well-motivated framework. The addition of a third truth value to deal with vague predicates or with the truth predicate leaves a number of issues open, however, starting with the interpretation of the third truth value, and with the choice of an appropriate consequence relation. Let us agree to call the value 1 'true-only', value 0 'false-only', and leave open what to call the value $\frac{1}{2}$ (see (Priest 1979), (Lewis 1982)). Depending on the view, $\frac{1}{2}$ may be called 'neither true nor false', or 'both true and false'. On paracomplete approaches, the Liar sentence is fundamentally viewed as neither true nor false, and borderline cases of vague predicates are cases for which it is neither true nor false that the predicate applies. On the dual, paraconsistent approaches to vagueness and the Liar paradox (Priest 1979), the Liar sentence is fundamentally viewed as both true and false, and similarly borderline cases of vague predicates are cases for which it is both true and false that the predicate applies. Importantly, distinct logics result depending on which interpretation of the third truth value is favored, and on whether logical consequence is defined as the preservation

of the value 1 (the true-only), or as the preservation of non-0 values (the non-(false-only)).

As it turns out, however, the duality between paracomplete and paraconsistent logics is such that the relative merits that one logic may claim over the other can usually be turned into relative limitations, and conversely. We will argue that a promising avenue for the treatment of the paradoxes lies in the definition of a consequence relation that results from a combination of paracomplete and paraconsistent features, rather than in the choice of one approach exclusive of the other. The relation of permissive consequence we have in mind is exactly along those lines, since it requires that when the premises of an argument take value 1 (are true-only), the conclusion must not take value 0 (is not false-only). A striking feature of this permissive consequence relation is that it exactly coincides with classical logic when no special provisos are made to deal with vagueness or with self-referential truth. When such provisos are included, however, we will see that this notion only departs from classical logic in that it yields a nontransitive consequence relation. This feature, arguably, does not constitute a cost: rather, we will argue that it captures a common and fundamental aspect to both families of paradoxes.

The paper is structured as follows. In Sect. 2, we give a brief overview of three-valued logic and introduce the notion of permissive consequence or st-consequence we use as our framework. In Sect. 3 we show how to extend the basic framework to accommodate vagueness on the one hand, and self-referential truth on the other, and in particular to deal with the Liar paradox and the sorites paradox. In Sect. 4, finally, we propose an assessment of our approach with regard to two main issues: the nontransitive character of permissive consequence on the one hand, and so-called revenge problems on the other, namely the treatment that we can give in our framework of the strengthened Liar and of higher-order vagueness.

21.2 Permissive Consequence and the Logic ST

21.2.1 The Scope of Permissive Consequence

The shape of the notion of permissive consequence we are about to introduce is by no means specific to the framework of three-valued logic. Also, in the literature the notion comes under various other names, such as *plausible* consequence (Frankowski 2004), *potential* consequence (Nait-Abdallah 1995), *arguable* consequence (Bennett 1998), or *tolerant* consequence (Zardini 2008). It can be defined for any logic in which it is sensible to distinguish a set of designated values and a set of tolerated values, where the set of tolerated values is a superset of the set of designated values. The general form of such a consequence relation was introduced and investigated independently by Frankowski (in (Frankowski 2004), based on earlier work by (Malinowski 1990) on the dual notion of *quasi*-consequence) and by Zardini in (Zardini 2008), with rather minimalist assumptions about the algebra of truth-values in each case. Zardini does not make restrictions, in particular, about the number of

truth-values, nor on whether truth-values should be partially or linearly ordered. Different logics correspond to this notion of permissive consequence depending on the algebra under consideration. Three-valued logic, however, is in a sense the smallest non-trivial framework for the investigation of this notion of permissive consequence, and interestingly, this notion was introduced independently by (Nait-Abdallah 1995) and by (Bennett 1998) in a trivalent setting.[1]

Bennett in particular put forward a notion of 'arguable entailment' for supervaluations, which he defined as follows: "if all of the premises are 'unequivocally' true, then the conclusion is 'in some sense' true". Although Bennett does not present it in that way, the definition can be seen to combine the notions of super-truth and sub-truth that are familiar from the literature on vagueness (see (Hyde 1997), (Cobreros et al. 2012a)). Even closer to our proposal, in an underappreciated book Nait-Abdallah investigated the interplay between two notions of truth in a trivalent setting, which he calls *classical truth* (for the value 1) and *potential truth* (for values > 0), which correspond exactly to the notions of strict truth and tolerant truth we are about to review and that we introduced independently. In different ways, however, Nait-Abdallah's study is both more general and more restricted than ours. It is more general in that it studies a notion of consequence in which premises and conclusions can be interpreted strictly or tolerantly in a non-uniform manner. It is more specific in that Nait-Abdallah limited his study of permissive consequence—consequence from classical to potential truth—to the case of consequence from zero premises.

21.2.2 st-consequence

To make our definitions precise, let us consider the language of first-order logic without identity and function symbols as our basic language, with negation, conjunction, and the universal quantifier as our basic logical connectives (we assume that the symbols $\bot$ and $\top$ are not part of the language). Importantly, we define the conditional $\supset$ as the material conditional in the usual way. We define three-valued models for this language over the set $\{1, \frac{1}{2}, 0\}$ of truth-values, using Kleene's strong schema as our valuation schema. According to Kleene's strong schema, negation maps the value 1 to 0, 0 to 1, and $\frac{1}{2}$ onto itself; conjunction is defined as the minimum of the values of the conjuncts, and universal quantification as the minimum of values over all assignments that differ at most on the value they assign to the variable bound by

[1] (Smith 2008), for instance, defines a notion of permissive consequence for fuzzy logic with continuum many truth values linearly ordered (the real interval $[0, 1]$), whereby a sentence A is said to follow from a set of premises Γ provided in all models in which all formulae in Γ have value strictly greater than half, A gets a value greater or equal than 1/2. This notion of permissive consequence, as fine-grained though it appears, can in fact be shown to be representable without loss of generality in a three-valued framework, and indeed, both Smith's notion and its three-valued version coincide with classical consequence for first-order logic. We refer to (Cobreros et al. Ms.) for details and more ample discussion of this point.

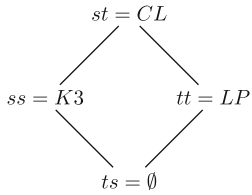


Fig. 21.1 Mixed consequence based on s and t

the quantifier (viz. (Kleene 1952)). A Kleene model for first-order logic is a structure $M = (D, I)$ where D is a set of individuals, and I an interpretation function for the non-logical vocabulary, that maps n-ary predicate symbols to functions from D^n to $\{1, \frac{1}{2}, 0\}$.

Based on this, let us say that a sentence A is *strictly true* or s-true in M, noted $M \models^s A$, provided $I(A) = 1$; we say that it is *tolerantly true* or t-true in M, noted $M \models^t A$, provided $I(A) > 0$. Thus, s-truth corresponds to what we earlier called for a sentence to be true-only, and t-truth for a sentence to be non-(false-only). Clearly, s-truth and t-truth are duals, that is, a sentence is tolerantly true iff its negation is not strictly true, and vice versa. For $n, m \in \{s, t\}$, moreover, the usual notion of logical consequence can be generalized by saying that:[2] $\Gamma \models^{nm} \Delta$ provided there is no model M such that $M \models^n \gamma$ for every $\gamma \in \Gamma$ and $M \models^m \delta$ for no $\delta \in \Delta$. As shown in Fig. 21.1, we thereby get four distinct notions of 'mixed' consequence.

When $n = m = s$, the resulting notion of logical consequence, or preservation of strict truth from premises to conclusions, coincides with Kleene's strong logic K3. When $n = m = t$, logical consequence corresponds to preservation of non-falsity or tolerant truth from premises to conclusions and the resulting system is Priest's Logic of Paradox (LP). When $nm = ts$, this corresponds to a case in which we go from tolerantly true premises to strictly true conclusions. The corresponding relation of consequence can be shown to be empty in this case. Intuitively, this corresponds to a notion of *restraining* consequence, since conclusions have to match a higher standard for truth than the premises. Conversely, the notion of *permissive* consequence we elect is defined in a dual way, namely as st-consequence, in that it asks for strictly true premises to imply conclusions that are tolerantly true.

The remarkable feature of st-consequence is that it coincides with classical consequence. Obviously, a classical countermodel to the entailment from Γ to Δ is an st-countermodel. But conversely, any st-countermodel can be turned into a classical countermodel, basically because reassignments of the values 1 or 0 to subsentences with value $\frac{1}{2}$ in the original model do not alter the value 1 or 0 assigned to the sentences in which they appear.

[2] We state the definition in terms of a multi-premise and multi-conclusion setting, although, for most of the applications we are interested in here, we can limit our perspective to the multi-premise-single-conclusion case.

Let us call ST the logic of st-consequence, and TS the logic of ts-consequence. It is worth stressing two aspects in which ST improves on the non-classical behavior of LP and K3. Both LP and K3 differ from CL in that both lose some classical validities, and both lose the deduction theorem. In particular, although $A \models_{K3} A$, it is not the case that $\models_{K3} A \supset A$. Similarly, although $\models_{LP} (A \wedge A \supset B) \supset B$, it is not the case that $A, A \supset B \models_{LP} B$. Thus, in K3 the loss of the deduction theorem is related to the loss of excluded middle in the same way in which, in LP, the loss of the deduction theorem is related to the loss of modus ponens as a valid inference. Because of that, it is generally agreed that neither K3 nor LP provides a satisfactory analysis of the conditional.

Upon reflection, this deficiency is not surprising. Indeed, it is easy to see that a conditional of the form $A \supset B$ is tolerantly true in a model provided either A is not strictly true, or B is tolerantly true, that is, provided that if A is strictly true, then B is tolerantly true. Dually, a conditional $A \supset B$ is strictly true in a model provided if A is tolerantly true, then B is strictly true. This suggests that in so far as a consequence relation can be expected to mirror the semantic behavior of its object-language conditional, st-consequence is the right correlate of the material conditional of LP, whereas ts-consequence is the right correlate for the material conditional of $K3$. In the case of ST, in particular, we will see in Sect. 3 that the classical behavior of the conditional is a significant advantage when we deal with vagueness and self-referential truth.

21.2.3 What Do 'strict' and 'tolerant' Mean?

A few more words are in order regarding the way in which the notions of strict truth and tolerant truth should be understood. We defined permissive consequence as the entailment from strict truth to tolerant truth. This may raise the legitimate worry that truth becomes an ambiguous notion.

Importantly, talk of tolerant *truth* and strict *truth* is not required to make sense of the notion of permissive consequence. An alternative route consists in linking the semantic values 1 and 0 not to truth proper, but to assertion. If we do so, "tolerant truth" and "strict truth" become essentially a *façon de parler*, and should be understood as shorthand for tolerant assertion and strict assertion. The idea, basically, is that a sentence is assertible strictly when there is non-arbitrary ground for the assertion. It can be denied strictly if there is non-arbitrary ground for denying it. To say that a sentence can be asserted or denied tolerantly means that there is ground for the assertion, but ground that may contain some element of arbitrariness (such as the existence of equal ground for the opposite assertion). Finally, a sentence can be such that it is assertible tolerantly and deniable tolerantly (at the opposite, no sentence can be asserted and denied strictly, but both a sentence and its negation can fail to be assertable strictly). Sentences that fall in that third category, we shall argue in the next section, are best matched by those sentences for which the rules that connect our use of language to empirical facts leave room for unsettledness

(remember the quote by McGee above). In what follows, we will see that we can use the strict/tolerant distinction as a way of classifying problematic sentences involving either vague predicates or self-referential truth.

The interpretation of the strict/tolerant distinction in terms of assertability rather than truth is compatible with an inferentialist interpretation of logical consequence, as opposed to what we might call a referentialist conception, on which truth values essentially reflect the correspondence status between a sentence and a state of affairs. By inferentialism, we mean the view on which linguistic meanings are to be explained by which inferences are valid, and more specifically the *bilateralist* view on which the validity of arguments itself is to be explained by general constraints on the speech acts of assertion and denial, or acceptance and rejection more generally (see (Rumfitt 2000), (Ripley 2013b), and (Malinowski 1990)).[3] This interpretation is arguably the most adequate when it comes to incorporating a transparent truth predicate in the language (see (Ripley 2012) and (Cobreros et al. 2013) and below for more ample discussion). This interpretation is not mandated, however. Some may find more appeal in the distinction between strict truth and tolerant truth as two levels of truth proper. (Smith 2008) for example defends a notion of permissive consequence for fuzzy logic in writing (p. 223):

> a sentence needs to meet more stringent standards of truth if it is to be used as the basis for further argument than if it is merely to be asserted—just as building codes place more stringent standards of load-bearing capacity on foundations than on superstructures.

Given Smith's commitment to degrees of truth, by standards of truth we take Smith to mean that assertability is based on those different levels of truth proper. A more neutral conception is defended by (Zardini 2008) who prefers to talk of truth values as "levels of goodness", whereby goodness is essentially a measure of the normative attitude to take toward a sentence (whether to believe it, assert it, or act upon it, see p. 345), without those attitudes necessarily being called good by reference to the truth of the corresponding sentence. If such levels of goodness are seen as ways of linking assertion to grounds for assertion, then they readily fit a unitary conception of truth, but a dual conception of assertion, on which the latter can come with different force.

21.3 Vagueness and Truth

Over a three-valued architecture, we see that ST allows us to preserve a classical notion of logical consequence. In this section we show how to extend ST to deal specifically with vague predicates on the one hand, and with a truth predicate on the other. In the case of vagueness, we will see that ST allows us to accommodate the

[3] Malinowski's notion of q-consequence is defined exactly in terms of the basic attitudes of acceptance and rejection. A sentence A is a q-consequence of a set of premises Γ whenever A is accepted when all of the premises in Γ are not rejected.

tolerance principle. In the case of truth, it allows us to accommodate a transparent truth predicate. In this section, we briefly review how to extend ST to deal with either kind of predicate. We mostly stress the analogies. We postpone a discussion of the potential disanalogies and limits of our account until the next section.

21.3.1 STVP

The hallmark of vague predicates on our account is the *tolerance principle* (Wright 1976), according to which a sufficiently small shift of the P-relevant respects of an individual should not make a difference as to whether the predicate P can be applied to that individual. For example, if someone 178cm tall is to be considered "tall", then someone only slightly shorter (177cm) can be considered tall too. More generally, we take the principle to be that if P is applicable to x, and x is sufficiently similar to y in the relevant respects, then P is applicable to y as well. In two-valued classical logic, the tolerance principle leads to paradox. In our approach, the tolerance principle can be validated without paradox.

To see this, we proceed to define an extension of ST called STVP (for 'ST with Vague Predicates'). As our language, we consider the language of first-order logic without identity, and containing, for simplicity, only unary predicates.[4] For each unary predicate P, moreover, the language contains a similarity predicate I_P. The formula aI_Pb is to be interpreted as: a and b are indiscriminable or sufficiently similar in P-relevant respects. Given a three-valued model $M = (D, I)$, each such predicate is to be interpreted by a relation $\sim_P$, with the following proviso:

$$M \models^s aI_Pb \text{ iff } M \models^t aI_Pb \text{ iff } |I(Pa) - I(Pb)| < 1 \text{ (closeness)} \qquad (21.1)$$

Three comments can be made on this definition of P-similarity. First, it implies that two individuals are P-similar on that view provided the application of the predicate P yields truth values that are sufficiently close. This interpretation of P-similarity in terms of closeness in truth values is faithful to what Smith calls the *closeness principle*, according to which, if two individuals a and b are sufficiently similar in P-relevant respects, then the degrees of truth of the corresponding sentences Pa and Pb should not be too far apart (Smith 2008). Secondly, the relation of similarity is, for each predicate, reflexive and symmetric, but it need not be transitive. The non-transitivity of indiscriminability or similarity is actually a central aspect to our conception of vagueness, and we share it with significantly different accounts of the logic of vague predicates (see (Williamson 1994) in particular). Thirdly, this relation of P-similarity has a crisp interpretation, in the sense that it makes no difference whether it is interpreted strictly or tolerantly.[5]

[4] The generalization to n-ary predicates presents no special difficulty, and only involves the definition of appropriate similarity relations between n-tuples.

[5] For more rigor, we might have chosen to break the proviso 21.1 into two separate constraints: first, the crispness constraint that $M \models^s aI_Pb$ iff $M \models^t aI_Pb$ (see (Cobreros et al. 2012b)), and

Let us call $\models_{STVP}$ the relation of ST-consequence specific to this language, with the proviso 21.1. It is easy to see that:

$$Pa, aI_Pb \models_{STVP} Pb \text{ (tolerance as a rule)} \tag{21.2}$$

and similarly that:

$$\models_{STVP} \forall xy(Px \wedge xI_Py \supset Py) \text{ (tolerance as an axiom)} \tag{21.3}$$

For this means that if Pa gets value 1 in the model, and the distance in truth values between Pa and Pb is less than 1, then the truth value of Pb is necessarily greater than 0. Because st-valid formulae coincide with tt-valid formulae, note that the validity of the tolerance principle could have been obtained with a tt or tolerant-to-tolerant consequence relation, that is using LP as background logic. However, the same difference between ST augmented with vague predicates and LP augmented with vague predicates remains that we had between ST and LP: in ST for the language with vague predicates, the conditional satisfies modus ponens, and more generally it satisfies the deduction theorem. This feature matters particularly, for as eloquently argued by (Zardini 2008, p. 339), rejection of modus ponens in the case of vagueness seems to "deprive" the tolerance principle, formulated in conditional form, "of its intended force".

More generally, for each of the four logics we considered above, namely ST, TT, SS and TS, the inclusion of similarity predicates with the closeness proviso in 21.1 yields four new consequence relations, which we call STVP, TTVP, SSVP and TSVP. The latter strictly extend the former (see Fig. 21.2 below; for instance TSVP now has validities such as $aI_Pb \supset aI_Pb$), and moreover, these are (model-theoretic) conservative extensions, in the sense that they coincide on the set of I_P-free formulae (we mark this relation of conservative extension with double lines on Fig. 21.2).

Obviously, although ST coincides with CL, STVP no longer coincides with CL, in particular because the tolerance principle is not classically valid. A further central difference between STVP and CL is that STVP does not yield a transitive consequence relation. This explains, in particular, why the sorites paradox can be blocked. For instance, we have that $aI_Pb, Pa \models_{STVP} Pb$ and $bI_Pc, Pb \models_{STVP} Pc$ but $aI_Pb, bI_Pc, Pa \not\models_{STVP} Pc$ (assume that $I(Pa) = 1$, $I(Pb) = 1/2$, $I(Pc) = 0$).

The nontransitive feature of STVP is intuitively faithful to the nontransitive character of indiscriminability in this example. This means that although the tolerance principle is $STVP$-valid, the tolerance step cannot be taken more than once without risk when reasoning with vague predicates. Note that because STVP satisfies the deduction theorem, the two versions of tolerance stated above, 21.2 and 21.3, are equivalent in STVP. This does *not* mean that the principle of tolerance, interpreted strictly, is equivalent to its tolerant interpretation. Indeed, a prima facie

secondly, the closeness constraint proper that if $M \models^{s/t} aI_Pb$, then $|I(Pa) - I(Pb)| < 1$ (not assuming the "only if" part). That way of doing things actually appears preferable to us in general, but we collapse both constraints in 21.1 for the sake of simplicity.

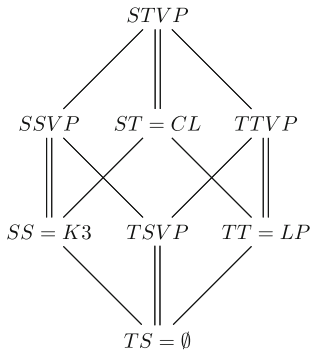


Fig. 21.2 Mixed consequence for vague predicates

counterintuitive consequence of our account is that $Pa_1, a_1 I_P a_2 ... I_P a_n, \forall xy(Px \wedge xI_P y \supset Py) \models_{STVP} Pa_n$. That is, if the tolerance principle is assumed to hold strictly, then the sorites paradox shows its ugly head again. However, the tolerance principle is only tolerantly valid, and cannot be used as a strict premise to derive new consequences. Note that this consequence is as it should be. For the tolerance principle $\forall xy(Px \wedge xI_P y \supset Py)$, interpreted strictly, actually means that if P holds *tolerantly* of x, and x and y are indiscriminable, then P holds *strictly* of y. Viewed in this way, we see that it now is a much stronger principle than when interpreted tolerantly. Only the tolerant interpretation, in our view, captures the adequate pretheoretical meaning attached to the notion of tolerance.

21.3.2 STTT

The hallmark of a truth predicate on our account is the *transparency principle*, according to which a sentence A should be intersubstitutable for $T\langle A\rangle$ in all extensional contexts and in all arguments without change of validity. Our reasons to hold on to transparency are fundamentally to let truth fulfill its expressive function in natural language (see (Field 2008) and (Cobreros et al. 2013) for ampler discussion). In two-valued classical logic, however, and provided the language is sufficiently expressive, the transparency principle leads to the Liar paradox and other related paradoxes, such as the Curry paradox.

Contrary to the tolerance principle in the case of vagueness, which is often viewed with suspicion by supporters of two-valued classical logic, the transparency principle for truth is generally seen as desirable even by supporters of bivalent classical logic. Because of that, a family of responses to the paradoxes of truth consists in typing truth predicates (a move first made by Tarski). This, intuitively, corresponds to one way of limiting the expressiveness of the language: sentences like the Liar or the Curry sentence are not well-formed. An alternative, first explored by (Kripke 1975), consists

in changing the logic, without the need for type-distinctions.[6] Kripke's approach thus succeeds in preserving the transparency principle, but it has several limitations. One of those concerns the fact that the resulting logic, K3TT (for K3 with Transparent Truth), is too weak to validate other principles that seemingly ought to result from transparency, such as the T-equivalence $T\langle A\rangle \equiv A$. In this section we show that by adopting ST as our background logic, we can likewise achieve transparency for truth, but without falling prey to the same limitations. As in the case of the sorites paradox for vagueness, the approach diagnoses the Liar and kindred paradoxes as making illegitimate use of the transitivity of logical consequence.

To see this, we proceed to define the system STTT (for ST with Transparent Truth). As our language in what follows, we assume the language of first-order logic without identity and function symbols, augmented with a distinguished predicate T for truth, and with a quote-name forming operator $\langle\rangle$, such that $\langle A\rangle$ is a name for the sentence A. In this language, in particular, we assume that we can formulate self-referential sentences such as the Liar sentence λ, which by definition is the sentence $\neg T\langle\lambda\rangle$, or the Truth-teller sentence τ such that τ is the sentence $T\langle\tau\rangle$, or the Curry sentence κ identical to $T\langle\kappa\rangle \supset A$ (for A a sentence that may take value 0 on all models). Our models for this language are Kripke-Kleene models, namely three-valued models of the same kind used so far, but with the following two constraints:

$$\begin{aligned} &a.\ \langle A\rangle \text{ always denotes the sentence } A \\ &b.\ A \text{ and } T\langle A\rangle \text{ always have the same truth value (identity of truth)} \end{aligned} \tag{21.4}$$

Note that the identity constraint on truth (also called the *fixed point property*) is an essential component toward transparency in our theory, but that identity and transparency are two independent constraints in general.[7] Likewise, closeness as defined in 21.1 is a component toward tolerance in the present theory of vagueness, but closeness and tolerance too are independent principles.[8] In the architecture of our theory therefore, we may say that the identity constraint on truth has the same

[6] Type-free treatments of the Liar purporting to maintain classical logic ought to be mentioned too. See in particular the contextualist accounts of (Parsons 1974) and (Glanzberg 2004), who both argue that the Liar rests on a phenomenon of variable quantifier domain restriction.

[7] We are indebted to an anonymous reviewer for this important clarification. As pointed out by the reviewer, in Kripke's construction the minimal fixed point V for the supervaluation schema satisfies the fixed point property, but not transparency, since as a consequence of the lack of value-functionality in the supervaluation schema, $V(\lambda \vee T\langle\lambda\rangle) = V(\neg T\langle\lambda\rangle \vee T\langle\lambda\rangle) = 1$, but $V(T\langle\lambda\rangle \vee T\langle\lambda\rangle) = 1/2$. Conversely, a transparent theory of truth may fail identity, for example if you start from a Kripke-Kleene model M and generate a new model M' that assigns to each sentence A the pair $<M(A), A>$ as a value, and then simply ignore the second coordinate of its values when defining validity. This sort of model will yield the same logic as the original models, but without ever assigning the same value to any two distinct sentences, so it will exhibit transparency without identity.

[8] (Smith 2008) presents closeness as an explicit weakening of tolerance in his fuzzy approach, and means to endorse closeness without endorsing tolerance. See (Cobreros et al. (Ms.)) however for a more thorough discussion of the status of both principles in relation to Smith's notion of consequence. Conversely, the theory of vagueness presented in (Cobreros et al. 2012b), which

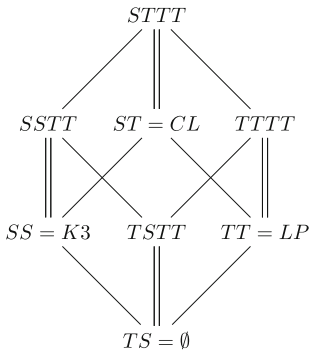


Fig. 21.3 Mixed consequence for transparent truth

priority with regard to transparency as the closeness constraint does with regard to tolerance in the case of vague predicates. That is, identity and closeness are initial model-theoretic postulates governing our special vocabulary, from which we are able to derive transparency and tolerance as general principles governing validity (even though identity and closeness are not meant to be substantially analogous besides this functional level).

That models satisfying (4)-a and (4)-b exist results from Kripke's 1975 fixed-point construction. The main difference with Kripke's approach is that we define logical consequence in terms of strict-to-tolerant consequence, that is, a sentence A follows from a set Γ of formulae provided there is no model where all the formulae of Γ take value 1 and where A takes value 0. Like the Strong Kleene definition of consequence (or indeed the LP one, or the tolerant-to-strict), this notion of consequence supports transparency (see (Ripley 2012) for a proof of this result). One particular consequence of this is the fact that all T -equivalences are STTT-valid. One of the essential benefits of this choice, moreover, which sets it apart from the other schemes, is that if an inference involving a T-free sentence is classically valid, then it remains STTT-valid for any uniform substitution over the full vocabulary (see (Ripley 2012)). Furthermore, the logic is simply better behaved than other three-valued logics in its vicinity, like LPTT, in which transparency and the T-equivalences can be validated, but where the rule of modus ponens is lost. Moreover, with regard to the conditional, STTT satisfies the deduction theorem, which neither K3TT nor LPTT do (compare the situation with the case of vagueness). If we map the extensions of TS, SS, TT and ST that we obtain with the enforcement of transparency, we get a diagram exactly congruent to the one we had for the corresponding extensions with vague predicates, as shown in Fig. 21.3.

As in the previous case, in Fig. 21.3 double lines (read top-down) in the figure indicate (model-theoretic) conservative extensions, and simple lines (top-down) that

involves the notion of classical extension for vague predicates, is one in which tolerance is *st*-valid, but without involving the notion of closeness in truth values. Tolerance *t*-holds in all models despite the existence of elements a and b for which aI_Pb holds (strictly or tolerantly), but such that $|I(Pa) - I(Pb)| = 1$ in the two-valued models used.

one logic contains strictly more validities than the other. Note that in contrast to our treatment of vague predicates, the inclusion of a transparent truth predicate does not necessarily add more validities. For instance, TSTT remains an empty consequence relation, just like TS (contrast this with STTT, in which A follows from $T\langle A\rangle$—a schema that would not be valid over plain ST augmented with T and quote-name operators but without the two provisos on names and identity).

Just like STVP, STTT no longer exactly coincides with two-valued classical consequence, despite preserving so many of the features of the latter. First of all, STTT validates sentences that would not be classically valid, on pain of contradiction. One such validity is the Liar sentence λ. The Liar is a valid sentence in STTT because it can only take the value $\frac{1}{2}$ (as in any Kripke fixed point), and therefore cannot take the value 0. Relatedly, STTT departs from standard classical logic in that it is not a transitive consequence relation. Thus, we have that for every sentence A, $A \models_{STTT} \lambda$ and for any sentence B, $\lambda \models_{STTT} B$, but we do not have $A \models_{STTT} B$ for every sentences A and B. Note that, as in the case of the tolerance principle for STVP, the loss of transitivity in STTT precisely explains why counting the Liar as a valid sentence is compatible with blocking the Liar paradox. In particular, we have that $\models_{STTT} T\langle\lambda\rangle \wedge \neg T\langle\lambda\rangle$. That is, we can infer both that the Liar is true and that it is not true. Likewise, we do have that $T\langle\lambda\rangle \wedge \neg T\langle\lambda\rangle \models_{STTT} A$, for any sentence A. If the Liar is true and not true, then anything follows. But we cannot derive $\models_{STTT} A$, that is, we cannot derive that any formula can be accepted tolerantly. The reason is that we illegitimately chained two valid inferences here, but one in which $\lambda \wedge \neg\lambda$ is accepted tolerantly as a conclusion with one where it is used strictly as a premise for further reasoning.

The analogy with our treatment of tolerance in STVP is worth stressing. Remember that in STVP, tolerance (whether as an axiom, or as a rule) was tolerantly but not strictly valid. Hence, it could not be used as a sound premise in an STVP-valid argument. Similarly here, one can accept tolerantly that the Liar is true and not true, but assuming strictly that it is true and not true (or even just one of them) leads to contradiction. Again, note that this consequence is as it should be. For this means that the Liar is not a sentence one can accept or deny strictly. But it remains a sentence that can be accepted tolerantly.

21.4 Nontransitivity and Revenge

In the previous section we have shown how to deal with the sorites paradox and the Liar paradox by means of strict-to-tolerant consequence. In this section, we propose an assessment of our approach. Two main issues need to be considered. The first concerns the scope of permissive consequence, and the question of how high a cost it is to give up transitivity for consequence in the face of the paradoxes. The second concerns whether we can similarly deal with higher-order versions of the paradoxes, namely with strengthened liars and higher-order vagueness.

21.4.1 Nontransitive Consequence

We have seen that both STTT and STVP are nontransitive consequence relations. One objection that could be made to our approach is that transitivity is too intimately tied to the analytic notion of consequence for our approach of the paradoxes to count as satisfactory.

Our answer to this objection is that in the case of STTT as well as in the case of STVP, transitivity is lost only to the extent that some conclusions can be drawn from strict premises that can only be accepted tolerantly. In other words, transitivity is lost only where it would be illegitimate to feed those sentences which can only be accepted tolerantly as strict premises to further arguments. However, transitivity is retained wherever we can ensure that we go from strictly accepted premises to strictly accepted conclusions. This would typically happen where we deal with non-vague predicates in particular (so in large chunks of science, where we can ensure precision). This also happens wherever we deal with sentences that do not involve the truth predicate at all. Arguably, therefore, the loss of transitivity in our system is quite limited: it only affects those inferences that involve special vocabulary, such as the truth predicate, or similarity predicates.[9]

Of course, the question may be asked of how good it is, then, to confer a special status of tolerant validities to sentences such as the tolerance principle, or such as the Liar. Why not simply work with only one notion of assertion, strict assertion, stick to transitive consequence for it, and rest content with the view that the tolerance principle or the Liar are paradoxical sentences, which should simply be excluded from sound reasoning?

This question raises fundamental issues, probably too fundamental to be answered satisfactorily here. However, one methodological answer is that, in so far as identity (see 21.4) is considered a basic postulate for truth on our account, and likewise in so far as the principle of closeness (see 21.1) is taken to govern our intuitions about similarity for vague predicates, going with strict assertion only would simply prevent us from getting any object-language equivalent of these principles as validities (such as the T-equivalences, or the tolerance principle). A more fundamental intuition we have is that going with strict assertion only would put too high standards on assertion. In the case of vagueness, for example, we do not agree that borderline cases of P, because they are neither strictly P nor strictly not P, should command silence on the part of speakers. Rather, we think that borderline cases are cases for which we have equally good reasons to issue judgments either way (see (Wright 1995), (Raffman 2014), (Cobreros et al. 2012b), (van Rooij 2012), (Egré 2011), (Ripley 2013a) for distinct but compatible justifications for this view). In the case of truth, mutatis mutandis, we consider the Liar to be a sentence for which there are inferential reasons for acceptance as well as rejection. Simply refraining to assert anything of

[9] See also (Cobreros et al. 2013) for a discussion of structural motivations for the admission of non-transitive consequence.

the Liar because of that would seem to us to amputate those inference grounds.[10] Of course, the question can be posed again of why it is not good enough to work with only one notion of assertion, namely the tolerant notion. Our answer in this case is that the logic we get is too weak. Consider vagueness again: in the LP version of our approach, the inference from $Pa, aI_P b$ to Pb is not valid, although the corresponding conditional is. This discrepancy appears to undercut the very motivation for having a tolerance principle.[11]

More specific worries may be expressed still about the failure of transitivity in STVP or STTT. In particular, one way in which the failure of transitivity shows in STTT and STVP concerns the closure of validities under modus ponens. In STTT, since $\neg\lambda$ is tolerantly valid, so is any sentence of the form $\lambda \supset A$, with A an arbitrary sentence. However, we cannot detach A for any such A, as soon as A is a sentence that can be denied strictly. So the Liar is a sentence that gives us as many conditionals as we want, but many of those will be conditionals whose consequent cannot be guaranteed to hold tolerantly (in contrast to the consequent Pb of the non-trivial conditional $Pa \wedge aI_P b \supset Pb$). Likewise, the set of STVP validities is not closed under modus ponens either. For example, $(Pa \vee \neg Pa) \supset ((aI_P b \wedge aI_P c) \supset (Pb \supset Pc))$ is STVP-valid, and so is $Pa \vee \neg Pa$. But $((aI_P b \wedge aI_P c) \supset (Pb \supset Pc))$ is not (let Pc take value 0, Pb take value 1, and Pa take value 1/2). Here too, we get the example of a conditional sentence whose validity can only be useful if the antecedent can be asserted strictly. In case the antecedent is only tolerantly assertable (as stipulated here), the sentence's tolerant validity makes the sentence too fragile for consequences to be detached safely.

A further concern one might have toward our non-transitive notion of consequence is how we can know of an arbitrary sentence that it can only be asserted tolerantly. After all, whether a sentence is Liar-like is contingent (see Kripke's Nixon-Dean example). So when we utter a sentence, how can we know that we are not allowed to assert it strictly? How can we know in particular that we are not allowed to reason transitively with a given sentence? Our answer to this question is that, as far as possible, our commitments with regard to assertion and reasoning should go with strict standards. If we know our grounds are safe enough for a strict assertion, then we can use modus ponens transitively provided we know the corresponding conditional itself to be good enough.[12] In other words, what ST-consequence recommends is:

[10] On the comparison between norms of assertion with regard to theories of truth, see especially (Wintein 2012). Chapter 7 of (Wintein 2012), in particular, presents a theory of truth based on the strict-tolerant distinction, but taking a different perspective on Kripke-Kleene models for truth as well as on assertibility proper.

[11] See, again, (Zardini 2008).

[12] Our view on this should be compared to Priest's original view on the status of modus ponens, a rule that is not LP-valid, but that Priest calls a "quasi-validity", still applicable to sentences that are not paradoxical. In our system, modus ponens is a validity, but wherever Priest talks of quasi-validities that are lost in relation to the conditional, we can speak of corresponding classical metainferences that are lost for consequence in the vicinity of paradoxes.

make sure that your conclusions are sufficiently robust in order to start using them as new premises.

Whether we are justified to assert a sentence strictly may not always be easy to ascertain, however. Because of that, we have to agree that our theory does not provide any a priori characterization of those sentences that can be asserted strictly, as opposed to tolerantly only. Upon reflection, however, the problem may be no more nor less pressing than it is when dealing with a transitive consequence relation. Suppose we had elected K3 as our logic. The choice of K3, a transitive consequence relation, would not make the predicament of determining whether a sentence is grounded (hence assertible strictly, or deniable strictly) easier to solve than it is with ST as our logic.[13] As argued by Kripke, any adequate theory (transitive or not) needs to admit an element of "risk" when dealing with truth and paradoxical sentences. But still, one could argue that the need to care about which sentences are assertible or deniable only matters for the soundness of arguments when our logic can rely on a unique mode of assertion, but that it does not matter for validity. In contrast, in ST we need to make sure that a sentence is assertible with the same force throughout in order to chain inferences.

This is indeed the case, but note that even in the setting of a classical and transitive logic, care needs to be taken in order to avoid ambiguity, and ambiguity is always likely to disrupt the validity of an argument. Here, we are talking of content-level ambiguity. Consider the following argument: "Aristotle is a merchant; if Aristotle is a merchant, then Plato is a slave. Hence Plato is a slave". For this inference to be valid, we need to ascertain that the names "Aristotle", "Plato", and the predicates "merchant" and "slave" get the same meaning in each occurrence. Usually, we assume such content-level ambiguities as already filtered out. But here too, that is with regard to reasoning quite generally, there is an element of risk. As a matter of principle, even for a transitive logic, validity holds only if we are certain to have avoided any equivocations. We take it that the problem is no more dramatic once we introduce two modes of assertion. In that case, the ambiguity concerns the mode rather than the content of sentences, but we view it as a virtue of our logic, rather than a defect, that it makes explicit the way in which ambiguities are likely to affect argument-validity quite generally. (Lewis 1982) famouly writes: "Logic for ambiguity—who needs it? I reply: pessimists." Our enterprise may be called: logic for assertoric ambiguity, but this is not to endorse pessimism about equivocation, since unlike the logics discussed by Lewis, our logic is explicitly committed to a dual theory of assertion.

[13] We are indebted to an anonymous reviewer for drawing attention to this point.

21.4.2 Revenge Issues: Strengthened Liars and Higher-Order Vagueness

We now turn to the second main objection our account needs to face. The objection concerns the recurrence of paradoxes at higher-orders. Whether for vagueness or for truth, this objection is usually pressed against three-valued accounts quite generally, irrespective of how logical consequence is defined in them.

Consider vagueness first. First-order vagueness is the claim that, between the clear instances of a predicate, and the clear counter-instances, there are borderline cases. Second-order vagueness in particular is the claim that there should also be borderline cases of borderline cases. That is, there should not be a sharp cutoff between clear cases of P and borderline cases of P. It may appear, however, that we are committed to such a sharp cutoff by accepting, in our models, the existence of at least two individuals d and d' in a model such that they are P-similar, and yet such that $I(P)(d) = 1$ and $I(P)(d') = 1/2$. Another way to phrase the problem is the following. Introduce an operator D for "determinately", such that $I(DA) = 1$ if $I(A) = 1$ and $I(DA) = 0$ if $I(A) < 1$. Clearly, $M \models^s DA$ iff $M \models^t DA$. Now, take any individual a such that $I(Pa) = 1/2$, that is, a is borderline P. Necessarily, $M \models^{s,t} D(\neg DPa \wedge \neg D\neg Pa)$, that is: a has to be a clear borderline case of P according to that definition.

Similarly, in the case of truth, although the Liar sentence can take on the value $1/2$ without contradiction, and without threatening transparency, a strengthened version of the Liar brings contradiction back in if we accept to enrich our vocabulary. Again, let D be the determinateness operator such that DA gets value 1 if A gets value 1 in the model, and gets value 0 otherwise. Let σ be the sentence such that σ is equivalent to $\neg DT\langle\sigma\rangle$. Thus, σ says of itself that it is not determinately true. We cannot assign σ a coherent truth value in the model while maintaining transparency anymore, if indeed sentences are allowed to take exactly one of the three truth values at our disposal.

When it comes to revenge issues, theories of vagueness as well as truth are usually faced with a dilemma. One horn of this dilemma is to limit the expressive power of the theory, and to deem unnecessary or illegitimate the introduction of such definiteness operators. The other horn is to consider that expressiveness should not be limited, but that such operators should be treated with particular care. To conclude this paper, we wish to explain in what sense we think our theory is compatible with both horns of this dilemma. Nevertheless, there is a sense in which, by referring the strict and tolerant distinction to assertion, rather than truth, our theory fits maybe more naturally with the idea of preventing the expression of revenge.

The issue of expressive limitation invites a more careful examination of determinateness operators in relation to our framework. One important observation to make about our whole approach is that determinateness operators are not part of the content of the sentences we are interested in, although something like determinacy operators is implicitly at play in the strict-tolerant distinction upon which our theory is built. Indeed, consider an atomic sentence like Pa, meaning that "a is rich". If Pa

holds strictly in our model, then this could be taken to mean that a is determinately rich. Dually, for Pa to hold tolerantly could be taken to mean that a is not determinately not rich. As a matter of fact, instead of introducing strict and tolerant levels for the truth values of our sentences, we could decide to work with a single notion of assertion, but to translate the strict and tolerant metalanguage distinctions into appropriate modal sentences of our object-language, enriched with determinateness operators (see (Kooi and Tamminga 2013) for an exact statement of such a modal translation for basic LP and K3 sentences). For example, to say that the tolerance principle is tolerantly valid, in modal terms, would turn out to be equivalent to the observation that the following "gap principle" holds classically in our models:[14]

$$\forall xy(DPx \wedge xI_P y \supset \neg D\neg Py) \tag{21.5}$$

Although such a translation is available and can be used to embed our treatment of vagueness in modal terms, we think that such an interpretation would likely distort the philosophical motivation of our approach. The main reason is that for us, strict and tolerant are primarily modes of assertion or acceptance; they qualify the force rather than the content of an assertion. Because of that, to assert a sentence such as Pa strictly is not analytically equivalent to the assertion that a is determinately P. Rather, asserting strictly is primarily tied to inferential and coherence commitments (such as the impossibility to deny even tolerantly). Asserting tolerantly, on the other hand, is also to take commitments (such as refusing to deny strictly). Because our approach relies primarily on such speech act distinctions, we thus believe the need to deal with determinateness operators in the object-language is less pressing than for other theories in which semantic values are primarily seen as ways of encoding the relation of a sentence with the world (see (Cook 2009) for such a conception).

Besides, the modal translation does not straightforwardly extend to sentences involving self-reference and the truth predicate. Kooi and Tamminga show how every propositional sentence of the basic language of LP or K3 can be translated into a modal sentence of S5 in a way that preserves argument-validity, but they do not provide a similar translation for the extended language with truth predicates. In LPTT, for example, we know that the sentence $T\langle\lambda\rangle \wedge \neg T\langle\lambda\rangle$ is valid. But if we apply the same translation manual to sentences like the Liar, with no special proviso on the truth predicate, we would get $\neg D\neg T\langle\lambda\rangle \wedge \neg DT\langle\lambda\rangle$ as our translation, and the grounds on which such a sentence should come out modally valid remain to be worked out (under a possible worlds semantics for the D operator).

This does not mean that issues of determinateness should be ignored. Consider the problem of higher-order vagueness. We can still accommodate determinateness operators in the object-language of our theory. But importantly, we do not think that "a is determinately P" should then necessarily have the truth conditions given above. In the case of vague predicates, we could accept, in particular, that the D operator does not necessarily map three-valued sentences to 0 or 1, but that it can

[14] See (Wright 1992), (Fara 2003), (Cobreros 2011) on gap principles, and (Egré 2011) and (van Rooij 2012) on the link between tolerance in the strict-tolerant framework and gap principles.

map sentences to 1/2, depending on the case (so that one could strictly assert that someone is bald, and still say something different with the strict assertion that that person is determinately bald). Whether we can accommodate indefinitely iterated borderline cases is an issue that goes beyond the scope of this paper, but our main point, once more, is that the semantics of determinateness is a matter distinct from strict assertion proper.

The situation is more thorny in the case of the strengthened Liar. We have to agree that, from a metatheoretical point of view, the strengthened Liar remains unescapable as soon as the relevant expressive means are available, just as, in recursion theory, any attempt to give a complete specification for the set of recursive functions, not involving partial functions, is threatened by the diagonalization method (Rogers 1987).[15] As (Priest 1984), (Cook 2009), and (Schlenker 2010) have argued, the strengthened Liar may in fact be considered an argument for the idea that truth values are indefinitely extensible, and that working with only three truth values sets an artificial bound on this phenomenon of indefinite extensibility. According to Cook and Schlenker, in particular, given any set S of truth values, the problem will indeed recur as soon as we have a sentence ρ equivalent to "ρ has a truth value in S other than true".

However, we have to emphasize that, in our theory, the natural way to express the sentence saying of itself that it is "other than true" is the standard Liar sentence. The reason is that we do not see the predicate "True" as tied to the value 1. Rather, "True" on our account is a predicate whose function is primarily inferential (as reflected by the identity constraint). In principle, however, we can still build a sentence that says of itself: "I am not strictly assertible", which one may formalize in terms of the sentence σ given above. Once we let sentences such as σ in, what are we to do with them? One possible line of response is to assent to the view of the indefinite extensibility of truth values, but to maintain the principled division between two modes of assertion. To do this, we may use a construction proposed by (Priest 1984).[16] To deal with σ, enlarge the space of truth-values to the power set of $\{0, \frac{1}{2}, 1\}$ (minus the empty set), and repeat the construction at higher levels. Now, consider what happens if σ gets the value 1: then it has to get value 0, and conversely. If it gets the value $\frac{1}{2}$, then it has to get the value 1, and also 0. By this reasoning, it seems the possible values for σ, upon this extended set, are $\{1, 0\}$ and $\{1, \frac{1}{2}, 0\}$. Now, let us say that, relative to this set, a sentence A is tolerantly assertible if the values it gets all contain 1 or $\frac{1}{2}$, tolerantly deniable if the values all contain 0 or $\frac{1}{2}$, strictly assertible if its value is $\{1\}$, and strictly deniable if its value is $\{0\}$. Seen in that way, the sentence "I am not strictly assertible" is therefore tolerantly assertible and deniable, and neither strictly assertible nor strictly deniable. This appears to be

[15] Rogers' emphasis on the use of partial, as opposed to total, functions as a way of blocking diagonalization arguments bears some analogy with the idea of limiting the expressiveness of our language to block the strengthened Liar.

[16] See (Ripley 2013a) for an application of this strategy to deal with a particular version of higher-order vagueness.

a desirable outcome, since by the valuation chosen, σ is not strictly assertible. But because this is what σ says, it appears we should be able to assert it in some sense, which we can do tolerantly. Similarly, consider a sentence ρ saying of itself that it is not tolerantly assertible (hence strictly deniable). This sentence is in fact tolerantly assertible on the valuation chosen. Hence it appears we should be able to deny it in some sense, which we can do tolerantly. Finally, take a sentence ν saying of itself that it is neither strictly assertible nor strictly deniable. If it gets value 1, it has to get value 0, and if it gets value $\frac{1}{2}$, it has to get value 1. However, it can get value 0 without contradiction. So whichever value it takes, it ought to get the value 0. Hence the sentence is strictly deniable. But since the sentence denies this, it seems that one should be able to deny it in some sense, which we can.

Based on this sample of examples, the valuation chosen appears to do justice to intuitions about acceptance and denial based on the inferences we can perform, and in a way that is faithful to inferential practice. Also, we can see that this is a way of replicating the trichotomy between strictly assertible, strictly deniable and tolerantly assertible one level up. Of course, in doing so, we are building a hierarchy for assertibility that parallels the hierarchy of truth values. But the important point is that, for any new predicate that one might introduce in the language to build a new extended Liar, one can come up with a way of understanding strict and tolerant assertion that will make the sentence neither strictly deniable, nor strictly assertible, but tolerantly both. The view of revenge as imposing an indefinite extension of truth-values is therefore compatible with the basic architecture of our theory, that is with the distinction between the mode of assertion (strict or tolerant) and the predicates intended to reflect those properties of assertibility in the object-language.

21.5 Conclusion

Our account of the paradoxes in terms of permissive consequence presents several advantages over other three-valued accounts in its vicinity. Firstly, with regard to other three-valued accounts based either on paracomplete or paraconsistent solutions, it allows us to maintain a simple logic, with a simple conditional, an ingredient that is notoriously missing from standard three-valued theories of either truth or vagueness. This feature, as we have emphasized, is obtained essentially because of the duality between strict and tolerant interpretations in our system, a duality that is missing from Kleene's logic as well as from LP.

Secondly, the framework accounts both for the tolerance of vague predicates, and for truth obeying the T-equivalence (on top of transparency) without paradox. From a philosophical point of view, however, it is particularly important to point out that tolerance and the T-equivalences are only tolerant validities. They are not strict validities, on pain of contradiction. In this, our framework, as the name indicates, is permissive indeed, but it comes with a caveat that inferences based on those principles in conditional form are fragile, because they fundamentally require the antecedent to be accepted strictly in order to go through safely.

The third feature of our account is that, because permissive consequence involves a shift of standard from premises to conclusions, it has to give up on the transitivity of logical consequence. Opponents to our account will likely consider that by validating tolerance or the T-schema without restriction, we have traded a reliable notion of consequence for principles that are fragile and of limited use. But this conclusion would be unfair. By its emphasis on two modes of acceptance, strict and tolerant, our account of consequence simply recognizes that chaining inferences is not an innocent business. As soon as we make room for reasoning with either vague predicates or self-referential sentences, logic, on our view, and logical consequence with it, needs to incorporate more generality, and more complexity with it.

Acknowledgments We are grateful to two anonymous referees for detailed comments, and to the editors for their assistance in the preparation of this paper. We also thank audiences in Amsterdam, Barcelona and Paris. We thank the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) / ERC grant agreement n 229 441-CCC, the ANR project BTAFDOC ("Beyond Truth and Falsity: Degrees of Confidence"), and the program "Non-Transitive Logics" (Ministerio de Economía y Competitividad, Government of Spain, FFI2013-46451-P). We also thank grants ANR-10-LABX-0087 IEC and ANR-10-IDEX-0001-02 PSL*.

References

Bennett, B. (1998). Modal semantics for knowledge bases dealing with vague concepts. In A. G. Cohn, L. Schubert, & S. Shapiro (Eds.), *Principles of Knowledge Representation and Reasoning: Proceedings of the 6th International Conference (KR-98), Morgan Kaufman, 1998*.

Cobreros, P. (2011). Supervaluationism and Fara's argument concerning higher-order vagueness. In P. Egré & N. Klinedinst (eds.), *Vagueness and language use* (pp. 207–221) Palgrave Macmillan.

Cobreros, P., Egré, P., Ripley, D., & van Rooij, R. (2012a). Tolerance and mixed consequence in the s'valuationist setting. *Studia Logica, 100*(4):855–877.

Cobreros, P., Egré, P., Ripley, D., & van Rooij, R. (2012b). Tolerant, classical, strict. *Journal of Philosophical Logic, 41*(2), 347–385.

Cobreros, P., Egré, P., Ripley, D., & van Rooij, R. (Ms). Tolerance and Degrees of Truth. Manuscript, in progress (of a paper presented at Cerisy 2011 and ESSLLI 2012).

Cobreros, P., Egré, P., Ripley, D., & van Rooij, R. (2013). Reaching transparent truth. *Mind. 122*(488):841–866.

Colyvan, M. (2009). Vagueness and truth. In H. Dyke (ed.), *From truth to reality: New essays in logic and metaphysics* (pp. 29–40). Routledge.

Cook, R. (2009). What is a truth value and how many are there? *Studia Logica, 92*(2), 183–201.

Egré, P. (2011). Perceptual ambiguity and the sorites. In R. Nouwen, R. van Rooij, U. Sauerland, & H.-C. Schmitz (eds.), *Vagueness in communication* (pp. 64–90). Springer.

Fara, D. (2003). Gap principles, penumbral consequence, and infinitely higher-order vagueness. *Liars and heaps: New essays on paradox* (pp. 195–222) (Originally published under the name "Delia Graff").

Field, H. (2003). Semantic paradoxes and vagueness paradoxes. In J. Beall (ed.), *Liars and heaps* (pp. 262–311). Oxford: Oxford University Press.

Field, H. (2008). *Saving truth from paradox*. Oxford: Oxford University Press.

Frankowski, S. (2004). Formalization of a plausible inference. *Bulletin of the Section of Logic, 33*(1), 41–52.

Glanzberg, M. (2004). A contextual-hierarchical approach to truth and the liar paradox. *Journal of Philosophical Logic, 33*(1), 27–88.
Hyde, D. (1997). From heaps and gaps to heaps of gluts. *Mind, 106,*641–660.
Kleene, S. (1952). *Introduction to metamathematics*. D. Van Nostrand Company Inc.
Kooi, B., & Tamminga, A. (2013). Three-valued logics in modal logic. *Studia Logica*, *101*(5), 1061–1072.
Kripke, S. (1975). Outline of a theory of truth. *The Journal of Philosophy, 72*(19), 690–716.
Lewis, D. (1982). Logic for equivocators. *Noûs, 16*(3)*,* 431–441.
Malinowski, G. (1990). Q-consequence operation. *Reports on mathematical logic, 24,* 49–54.
McGee, V. (1991). *Truth, vagueness, and paradox: An essay on the logic of truth*. Hackett Pub Co Inc.
Nait-Abdallah, A. (1995). *The logic of partial information*. Springer.
Parsons, C. (1974). The liar paradox. *Journal of Philosophical Logic, 3*(4), 381–412.
Priest, G. (1979). Logic of paradox. *Journal of Philosophical Logic, 8,* 219–241.
Priest, G. (1984). Hyper-contradictions. *Logique et Analyse, 27*(107), 237–243.
Priest, G. (2010). Inclosures, vagueness, and self-reference. *Notre Dame Journal of Formal Logic, 51*(1), 69–84.
Raffman, D. (2014). *Unruly words: A study of vague language*. Oxford University Press.
Ripley, D. (2012). Conservatively extending classical logic with a transparent truth predicate. *Review of Symbolic Logic 5*(2):354–378.
Ripley, D. (2013a). Sorting out the sorites. In F. Berto, E. Mares, & K. Tanaka, (eds.), *Paraconsistency: Logic and applications*. Springer 329–348.
Ripley, D. (2013b). Paradoxes and failures of cut. *Australasian Journal of Philosophy* 91(1): 139–164.
Rogers, H. (1987). *Theory of recursive functions and effective computability*. MIT.
Rumfitt, I. (2000). "Yes and no". *Mind, 109*(436), 781–823.
Schlenker, P. (2010). Super liars. *The Review of Symbolic Logic, 1*(1), 1–41.
Smith, N. J. J. (2008). *Vagueness and degrees of truth*. Oxford: Oxford University Press.
Tappenden, J. (1993). The liar and sorites paradoxes: Toward a unified treatment. *The Journal of Philosophy, 90*(11), 551–577.
van Rooij, R. (2012). Vagueness, tolerance, and non-transitive entailment. In P. Cintula, C. Fermueller, L. Godo, & P. Hàjek (Eds), *Understanding vagueness—Logical, philosophical and linguistic perspectives* (pp. 205–221). King's College Publications.
Williamson, T. (1994). *Vagueness*. London: Routledge.
Wintein, S. (2012). *Playing with truth*. PhD Thesis, Tilburg University.
Wright, C. (1976). Language-mastery and the sorites paradox. In G. Evans, & J. McDowell (Eds.), *Truth and meaning: Essays in Semantics*, pp. 223–247.
Wright, C. (1992). Is higher order vagueness coherent? *Analysis, 52*(3)*,* 129–139.
Wright, C. (1995). The epistemic conception of vagueness. *The Southern journal of philosophy, 33*(S1), 133–160.
Zardini, E. (2008). A model of tolerance. *Studia Logica, 90,* 337–368.

Chapter 22
Validity and Truth-Preservation

Julien Murzi and Lionel Shapiro

Abstract The revisionary approach to semantic paradox is commonly thought to have a somewhat uncomfortable corollary, viz. that, on pain of triviality, we cannot affirm that all valid arguments preserve truth (Beall 2007, 2009; Field 2008, 2009b). We show that the standard arguments for this conclusion all break down once (i) the *structural rule of contraction* is restricted and (ii) how the premises can be aggregated—so that they can be said to *jointly* entail a given conclusion—is appropriately understood. In addition, we briefly rehearse some reasons for restricting structural contraction.

Logical orthodoxy has it that valid arguments preserve truth (see e.g. Etchemendy 1990; Harman 1986, 2009):

(VTP) If an argument is valid, then, if all its premises are true, then its conclusion is also true.

Intuitive as it may seem, this claim, on natural enough interpretations of 'if' and 'true', turns out to be highly problematic. Hartry Field has argued that its most immediate justification requires all the logical and semantic resources that yield the standard semantic version of Curry's Paradox. Worse yet, both Field and Jc Beall have observed that the claim that valid arguments preserve truth almost immediately

Thanks to Jc Beall, Colin Caret, Roy Cook, Charlie Donahue, Ole T. Hjortland, Jeff Ketland, Hannes Leitgeb, Francesco Paoli, Stephen Read, and Greg Restall for helpful discussion on some of the topics discussed herein, and to Dave Ripley and a referee for detailed comments on a previous draft. Julien Murzi warmly thanks the Alexander von Humboldt Foundation, the University of Padua, and the School of European Culture and Languages at the University of Kent for generous financial support. Lionel Shapiro is grateful to the Arché Research Centre at the University of St Andrews for making possible a productive visit.

J. Murzi
University of Salzburg, Franziskanergasse 1, 5020 Salzburg, Austria
e-mail: julien.murzi@sbg.ac.at

L. Shapiro
University of Connecticut, Storrs, CT, USA
e-mail: lionel.shapiro@uconn.edu

T. Achourioti et al. (Eds.), *Unifying the Philosophy of Truth,* Logic, Epistemology, and the Unity of Science 36, DOI 10.1007/978-94-017-9673-6_22

yields absurdity via Curry-like reasoning in most logics (Field 2008; Beall 2007, 2009). Moreover, Field has argued that, by Gödel's Second Incompleteness Theorem, any semantic theory that declares all valid arguments truth-preserving must be inconsistent (Field 2006, 2008, 2009b, 2009a). We can't coherently require that valid arguments preserve truth, or so the thought goes.[1]

Two main ingredients are required for this conclusion: that the conditional occurring in VTP detaches, i.e. satisfies *Modus Ponens*, and the *naïve view of truth*, viz. that (at the very least) the truth predicate must satisfy the (unrestricted) T-Scheme

$$(\textit{T-Scheme})\ Tr(\ulcorner \alpha \urcorner) \leftrightarrow \alpha,$$

where $Tr(...)$ expresses truth, and $\ulcorner \alpha \urcorner$ is a name of α. Both assumptions lie at the heart of the leading contemporary *revisionary approaches* to semantic paradox. These include recent implementations (see e.g. Brady 2006; Field 2003, 2007, 2008; Horsten 2009) of the *paracomplete* approach inspired by Martin and Woodruff (1975) and Kripke (1975), as well as *paraconsistent* approaches (see e.g. Asenjo 1966; Asenjo and Tamburino 1975; Priest 1979, 2006a, 2006b; Beall 2009). Paracomplete approaches solve paradoxes such as the Liar by assigning the Liar sentence a value in between truth and falsity, thus invalidating the Law of Excluded Middle. Paraconsistent approaches solve the Liar by taking the Liar sentence to be both true and false, avoiding absurdity by invalidating the classically and intuitionistically valid principle of *Ex Contradictione Quodlibet*. Both approaches have sought to preserve room for a detaching conditional that underwrites the T-Scheme. And when such a conditional threatens to reintroduce absurdity through Curry's Paradox, both approaches have offered a common diagnosis: they take it to show that this conditional cannot satisfy the law of contraction:

$$(\textit{Contraction})\ (\alpha \rightarrow (\alpha \rightarrow \beta)) \rightarrow (\alpha \rightarrow \beta).$$

More generally, they require that a theory of truth be *robustly contraction free* ('rcf', for short); free, essentially, of a conditional satisfying Contraction and other natural principles such as *Modus Ponens* (Restall 1993).

In this paper, we assume for argument's sake the naïve view of truth, and argue that this view doesn't in fact require rejecting VTP. However, maintaining VTP requires more than revising logic so as to ensure that Contraction is no longer a theorem. Rather, it involves adopting a logic that lacks one or more of the rules usually thought to correspond to basic features of reasoning in the context of assumptions. We will focus on the *structural* rule of contraction

$$(\mathsf{SContr})\ \frac{\Gamma, \alpha, \alpha \vdash \beta}{\Gamma, \alpha \vdash \beta}$$

Once SContr is rejected, we will see, the standard objections against VTP all break down. The standard arguments against VTP at best support the weaker conclusion

[1] Shapiro (2011) refers to the the claim that VTP and the naïve view of truth we introduce in the next paragraph yield triviality as the 'Field-Beall thesis'.

that, given the naïve view of truth, *either* VTP *or* SContr (or perhaps some other structural feature of the consequence relation) should be rejected.

To be sure, rcf theorists, especially Field, are aware of the existence of substructural revisionary approaches. Field dismisses them, though, as "radical" (Field 2008, p. 10), and as "very desperate measures" that are, ultimately, not needed (Field 2009a, p. 350). He writes:

> I haven't seen sufficient reason to explore this kind of approach (which I find very hard to get my head around), since I believe we can do quite well without it. ... [Hence] I will take the standard structural rules for granted. (Field 2008, pp. 10–11; also 283n)

However, while we agree with Field that more work needs to be done to make sense of a failure of SContr, we'd like to stress that giving up VTP is *also* a radical move. What is more, revisionary theorists have at least one powerful reason to reject SContr. Let us assume, as is often done, that the "valid" arguments include those whose goodness depends on rules governing the truth and validity predicates (McGee 1991; Whittle 2004; Priest 2006a, 2006b; Field 2007, 2008; Zardini 2011). Then there exist validity-involving versions of Curry's Paradox which cannot be solved by revising the logic's *operational* rules (those governing the behavior of logical vocabulary) to ensure that the theory is robustly contraction free. This is because the only operational rules these versions of Curry's Paradox employ are a pair of rules governing a validity predicate, rules that are arguably essential to that predicate's expressing validity (Shapiro 2011; Beall and Murzi 2013).

It has long been known that Curry-paradoxical reasoning can be blocked by adopting a "substructural" logic lacking SContr.[2] Yet we're not aware of any detailed examinations of how the various challenges to VTP are affected by adopting such logics.[3] What makes matters delicate is that all the challenges to VTP involve arguments with *multiple premises*. Hence how we may respond to the challenges depends crucially on how we understand what it means for a conclusion to follow validly from all of the premises *taken jointly*. Even stating what truth-preservation amounts to requires us to represent such joint consequence using some logical connective in place of the above informal 'all' or (in the case of arguments with finitely many premises) in place of the corresponding 'and.' Once SContr is rejected, various possibilities open up for the logical behavior of such an 'and', with different choices having different implication for the challenges to VTP. Moreover, the possibility arises that there are *two* suitable connectives, corresponding to different modes in which premises may be understood as taken jointly. Our chief aim is to clarify this poorly understood complex of issues and challenge the received wisdom that VTP is incompatible with revisionary approaches to paradox.

Two final qualifications. The structural feature of validity encapsulated in SContr isn't the only standardly accepted structural feature whose rejection would block the validity-involving versions of Curry's Paradox and allow a defense of VTP against the standard objections. An alternative "substructural" strategy, proposed by Ripley

[2] See Slaney (1990), Restall (1994) and Field (2008, p. 283n).
[3] There is some relevant discussion in Shapiro (2011) and Zardini (2011).

(2013), involves restricting the *transitivity* of validity as reflected in the structural rule of Cut.[4] While we will occasionally remark on this approach, we do not have space to compare it with the strategy of giving up SContr.[5] In what follows, we will assume (as rcf theorists typically do) that validity is transitive. Likewise, we won't here be able to discuss the various ways in which one might try to make sense of and motivate the failure of SContr.[6]

The remainder of this paper is structured thus. §1 introduces the standard arguments in favor of rejecting VTP. §2 observes that VTP follows from what we call the *naïve view of validity*, viz. that the validity predicate satisfies (generalisations of) the Rule of Necessitation and the T axiom. It then rehearses some reasons for thinking that the naïve view of validity is in tension with SContr, and considers a couple of possible objections to this claim. §3 examines various possible interpretations of VTP, interpretations that become available once SContr is rejected. Specifically, it considers different ways of understanding the claim that an argument's premises are *all* true, as one finds in linear logic and what we call dual-bunching logics. It then argues that, once SContr is rejected, the standard objections to VTP are all blocked. §4 offers some concluding remarks.

22.1 Three Challenges to VTP

We focus on *three* challenges to VTP: that the most obvious argument in defense of this principle rests on inconsistent premises, that VTP yields triviality via Curry-like reasoning, and that Gödel-like reasoning shows that no consistent recursively axiomatizable semantic theory can endorse VTP.

22.1.1 The Validity Argument and Curry's Paradox

Field (2008, §2.1, §19.2) considers an argument, which he calls the Validity Argument, to the effect that "an inference is valid if and only if it is logically necessary

[4] Weir (2005) also addresses semantic paradox by restricting the transitivity of validity, though this shows up in his natural deduction system as a structure-based restriction on the use of *operational* rules.

[5] Both of these "substructural" approaches to semantic paradox have an advantage worth mentioning: they allow for a *unified* approach to the paradoxes of self-reference (Weir 2005; Zardini 2011; Ripley 2013), as opposed to the piecemeal approach proposed by current rcf theories, where similar paradoxes, e.g. the Liar and Curry, are treated in radically different ways. In recent unpublished work, Beall uses the desideratum of uniformity as one motivation for a new approach to paradox—one that retains the standardly accepted structural rules but gives up on a detaching conditional altogether. For a sketch of that approach, see Beall (2011).

[6] For discussion of this important topic, see Shapiro (2011), Zardini (2011), Beall and Murzi (2013), Mares and Paoli (2014). [Note added in proof: see also Shapiro (2015).]

that it preserves truth" (Field 2008, p. 284). If sound, the argument for this biconditional's 'only if' direction would seem to establish VTP. However, Field argues, it can't be sound. Let's use $\alpha_1, ..., \alpha_n \vdash \beta$ to mean that "the argument from the premises $\alpha_1, ..., \alpha_n$ to the conclusion β is logically valid" (Field 2008, p. 42). And let *Tr*-I and *Tr*-E, respectively, be the rules that one may infer $Tr(\ulcorner\alpha\urcorner)$ from α in any context of assumptions, and vice versa. Then Field reasons thus (we have adapted his terminology):

> *'Only if' direction*: Suppose $\alpha_1, ..., \alpha_n \vdash \beta$. Then by *Tr*-E, $Tr(\ulcorner\alpha_1\urcorner), ..., Tr(\ulcorner\alpha_n\urcorner) \vdash \beta$; and by *Tr*-I, $Tr(\ulcorner\alpha_1\urcorner), ..., Tr(\ulcorner\alpha_n\urcorner) \vdash Tr(\ulcorner\beta\urcorner)$. By $\wedge$-E, $Tr(\ulcorner\alpha_1\urcorner) \wedge ... \wedge Tr(\ulcorner\alpha_n\urcorner) \vdash Tr(\ulcorner\beta\urcorner)$. So by $\rightarrow$-I, $\vdash Tr(\ulcorner\alpha_1\urcorner) \wedge ... \wedge Tr(\ulcorner\alpha_n\urcorner) \rightarrow Tr(\ulcorner\beta\urcorner)$. That is, the claim that if the premises $\alpha_1, ..., \alpha_n$ are true, so is the conclusion, is valid, i.e. holds of logical necessity. *'If' direction*: Suppose $\vdash Tr(\ulcorner\alpha_1\urcorner) \wedge ... \wedge Tr(\ulcorner\alpha_n\urcorner) \rightarrow Tr(\ulcorner\beta\urcorner)$. By *Modus Ponens*, $Tr(\ulcorner\alpha_1\urcorner) \wedge ... \wedge Tr(\ulcorner\alpha_n\urcorner) \vdash Tr(\ulcorner\beta\urcorner)$. So by $\wedge$-I, $Tr(\ulcorner\alpha_1\urcorner), ..., Tr(\ulcorner\alpha_n\urcorner) \vdash Tr(\ulcorner\beta\urcorner)$. So by *Tr*-I, $\alpha_1, ..., \alpha_n \vdash Tr(\ulcorner\beta\urcorner)$; and by *Tr*-E, $\alpha_1, ..., \alpha_n \vdash \beta$. (Field 2008, p. 284).[7]

Two features of this Validity Argument call for comment. First, notice that it is conducted in a metalanguage containing a validity predicate (the turnstile), but no truth predicate. In taking the argument to establish VTP, then, Field is assuming that the object-language sentence $Tr(\ulcorner\alpha_1\urcorner) \wedge ... \wedge Tr(\ulcorner\alpha_n\urcorner) \rightarrow Tr(\ulcorner\beta\urcorner)$ expresses the claim *that if* $\alpha_1, ..., \alpha_n$ *are all true, so is* β. In §3, we will see that once giving up structural contraction is an option, it becomes controversial whether the claim that "all premises are true" should be expressed using a connective for which the inferences Field justifies using $\wedge$-I and $\wedge$-E are valid. Second, one might worry that the Validity Argument presupposes its own conclusion. The argument establishes that if an argument is valid, then the conditional claiming that the argument preserves truth will likewise be *valid*. But we couldn't take this as establishing VTP itself unless we took for granted that *valid sentences are true*—a claim that is a special case of VTP. Still, even if the Validity Argument doesn't suffice to establish VTP, it does undermine the objections that have been offered against VTP. That is because these objections (which all involve multi-premise arguments) don't purport to challenge the claim that *valid sentences are true*. Thus the Validity Argument should count as a defense of VTP.[8]

---

[7] It may help to make Field's reasoning for the 'only if' direction explicit in natural deduction format, for the special case where we are considering an argument from the single premise α to the conclusion β. Complications raised by the multiple-premise case will be discussed in §3.

$$\dfrac{\dfrac{\alpha \vdash \beta \qquad \dfrac{Tr(\ulcorner\alpha\urcorner) \vdash Tr(\ulcorner\alpha\urcorner)}{Tr(\ulcorner\alpha\urcorner) \vdash \alpha}\,Tr\text{-E}}{Tr(\ulcorner\alpha\urcorner) \vdash \beta}\,\text{Cut}}{\dfrac{Tr(\ulcorner\alpha\urcorner) \vdash Tr(\ulcorner\beta\urcorner)}{\vdash Tr(\ulcorner\alpha\urcorner) \rightarrow Tr(\ulcorner\beta\urcorner)}\,\rightarrow\text{-I}}\,Tr\text{-I}$$

[8] In §2.1, we will see that if our object-language contains a validity predicate, it is also possible to derive VTP using an intuitively compelling elimination rule for that predicate. While we will discuss only a predicate that applies to single-premise arguments, a generalized version of that derivation would be subject to all our conclusions about the Validity Argument.

Field suggests that the Validity Argument, though it "looks thoroughly convincing at first sight," can't be accepted, since it relies on Tr-I, Tr-E, $\rightarrow$-I, and $\rightarrow$-E, "which the Curry Paradox shows to be jointly inconsistent" (Field 2008, pp. 43, 284). Let us unpack this a little. The Diagonal Lemma allows us to construct a sentence κ which, up to equivalence, intuitively says that, if it's true, then (say) you will win the lottery. Assuming that our theory of truth T is strong enough to prove the Diagonal Lemma, this means that

$$\vdash_T \kappa \leftrightarrow (Tr(\ulcorner \kappa \urcorner) \rightarrow \bot).$$

Let Π now be the following derivation of the further theorem $Tr(\ulcorner \kappa \urcorner) \rightarrow \bot$:

$$\dfrac{\dfrac{\dfrac{\vdash_T \kappa \leftrightarrow Tr(\ulcorner \kappa \urcorner) \rightarrow \bot \qquad \dfrac{Tr(\ulcorner \kappa \urcorner) \vdash_T Tr(\ulcorner \kappa \urcorner)}{Tr(\ulcorner \kappa \urcorner) \vdash_T \kappa}\, Tr\text{-E}}{Tr(\ulcorner \kappa \urcorner) \vdash_T Tr\ulcorner \kappa \urcorner \rightarrow \bot}\, \rightarrow\text{-E} \qquad Tr(\ulcorner \kappa \urcorner) \vdash_T Tr(\ulcorner \kappa \urcorner)}{Tr(\ulcorner \kappa \urcorner), Tr(\ulcorner \kappa \urcorner) \vdash_T \bot}\, \rightarrow\text{-E}}{\dfrac{Tr(\ulcorner \kappa \urcorner) \vdash_T \bot}{\vdash_T Tr(\ulcorner \kappa \urcorner) \rightarrow \bot}\, \rightarrow\text{-I}}\, \mathsf{SContr}$$

Using Π, we can then 'prove' that you will win the lottery:

$$\dfrac{\dfrac{\vdash_T Tr(\ulcorner \kappa \urcorner) \rightarrow \bot \qquad \dfrac{\dfrac{\vdash_T \kappa \leftrightarrow (Tr(\ulcorner \kappa \urcorner) \leftrightarrow \bot) \qquad \vdash_T Tr(\ulcorner \kappa \urcorner) \rightarrow \bot}{\vdash_T \kappa}\, \rightarrow\text{-E}}{\vdash_T Tr(\ulcorner \kappa \urcorner)}\, Tr\text{-I}}{\vdash_T \bot}\, \rightarrow\text{-E}}{}$$

(with Π above $\vdash_T Tr(\ulcorner \kappa \urcorner) \rightarrow \bot$ and Π above $\vdash_T Tr(\ulcorner \kappa \urcorner) \rightarrow \bot$)

This is the (standard) conditional-involving version of Curry's Paradox, or c-Curry, as we'll call it.[9] The derivation makes use of Tr-I, Tr-E, $\rightarrow$-I and $\rightarrow$-E, just like the Validity Argument. Hence, Field argues, one can't accept the latter without thereby validating the former. Rcf theorists invalidate c-Curry by rejecting $\rightarrow$-I, thus resisting Π's final step (Priest 2006b; Field 2008; Beall 2009; Beall and Murzi 2013). Therefore, Field suggests, they must reject the 'only if' direction of the Validity Argument, too.

However, as Field notes, the above derivation makes use of the rule SContr. Hence if SContr is rejected—as proposed in this context by Brady (2006), Zardini (2011), Shapiro (2011), and Beall and Murzi (2013)—Curry's paradox no longer stands in the way of our embracing the principles used in the Validity Argument for VTP. One complication: we should note in advance that it isn't clear that all types of contraction-free logics we will be considering support theories of arithmetic that prove a Diagonal Lemma. Where this isn't the case, the reader should suppose that some other means of self-reference built into our semantic theory is responsible for the Curry paradoxes

[9] This terminology was introduced in Beall and Murzi (2013).

we will be considering. In what follows, we will ignore this complication, and assume that T has the resources for at least simulating self-reference.

Will rejecting SContr allow us to endorse the Validity Argument, then? As we will see below, matters are not this simple. Field's argument makes crucial use of rules governing the conjunction symbolized by $\wedge$. Once we no longer accept the standard structural rules, however, the rules for conjunction can take non-equivalent forms, and the soundness of the Validity Argument now depends on which of the available rules for $\wedge$ we accept. In §3, we will examine which of the contraction-free logics that have been proposed in response to semantic paradox underwrite the Validity Argument.[10]

22.1.2 *From VTP to Absurdity via the Modus Ponens Axiom*

In addition to criticizing the most obvious *defense* of VTP, Field offers two arguments according to which VTP can't be embraced without absurdity. In the remainder of this section, then, let us examine whether we can at least *affirm* that valid arguments preserve truth. For simplicity's sake, we focus for now on arguments with only one premise. (Issues raised by multiple-premise arguments will be considered in detail in §3 below.) We will try to affirm VTP in the object-language itself, by introducing a predicate *Val(x, y)* which intuitively expresses that the argument from x to y is valid. VTP may now be naturally represented thus (see Beall 2009):

$$(\mathsf{V0})\ Val(\ulcorner\alpha\urcorner, \ulcorner\beta\urcorner) \rightarrow (Tr(\ulcorner\alpha\urcorner) \rightarrow Tr(\ulcorner\beta\urcorner)).^{11}$$

As Field and Beall point out, V0 entails absurdity, based on principles accepted by rcf theorists (Field 2006, 2008; Beall 2007, 2009).

---

[10] Let us briefly consider how the Validity Argument fares on the alternative substructural approach that restricts transitivity. In the version of c-Curry given above, in natural deduction format, SContr is the only structural rule used. By contrast, the parallel Curry derivation in sequent calculus format will conclude with the following use of the structural rule of Cut

$$\frac{\vdash_T Tr(\ulcorner\kappa\urcorner) \quad Tr(\ulcorner\kappa\urcorner) \vdash_T \bot}{\vdash_T \bot}$$

Ripley (2013) proposes a semantic theory that blocks c-Curry reasoning by invalidating Cut. His theory adds rules for Tr to a sequent calculus with entirely classical operational rules and structural rules except for Cut, which is no longer admissible in the presence of the truth rules. We would like to make two observations about Ripley's proposal. On the one hand, since it retains the rule $\rightarrow$-I, it allows a defense of the Validity Argument's "only if" direction (his truth rules replace Cut in note 7 above), and thus of VTP. On the other hand, though Ripley's theory also endorses the *conclusion* of every instance of the Validity Argument's "if" direction, it won't allow the above intuitive *argument*, since it renders the rule $\rightarrow$-E inadmissible. See note 46 below.

[11] Strictly speaking, this should be expressed by a universal generalisation on codes of sentences, but, for the sake of simplicity, we won't bother.

Since rcf theorists do not accept the rule $\rightarrow$-I, we will need two additional ingredients to obtain paradox from V0. First, the rules Tr-I and Tr-E no longer suffice; our semantic theory T needs to underwrite all instances of the T-Scheme. Second, we will use the principle that if $\vdash_T \alpha \leftrightarrow \beta$, then α and β are intersubstitutable within conditionals.[12] Given these presuppositions, V0 entails

$$(\mathsf{V1})\ Val(\ulcorner\alpha\urcorner, \ulcorner\beta\urcorner) \rightarrow (\alpha \rightarrow \beta).$$

Now let us assume, as rcf theorists do, that our theory T implies the validity of a single-premise version of the *Modus Ponens rule*:

$$(\mathsf{VMP})\ Val(\ulcorner(\alpha \rightarrow \beta) \wedge \alpha\urcorner, \ulcorner\beta\urcorner).$$

Hence V1 in turn entails the *Modus Ponens axiom*:

$$(\mathsf{MPA})\ (\alpha \rightarrow \beta) \wedge \alpha \rightarrow \beta.^{13}$$

However, Meyer et al. (1979) show that MPA generates Curry's Paradox. The only additional ingredient we need is the claim that it is a theorem that "conjunction is idempotent," i.e. that $\vdash \alpha \leftrightarrow \alpha \wedge \alpha$.

To see why this is so, recall that we have assumed T is strong enough to ensure $\vdash_T \kappa \leftrightarrow (Tr(\ulcorner\kappa\urcorner) \rightarrow \bot)$. Hence, given the T-Scheme and the above substitutivity principle, $\vdash_T \kappa \leftrightarrow (\kappa \rightarrow \bot)$. We can now derive absurdity starting with the relevant instance of MPA:

$$(\kappa \rightarrow \bot) \wedge \kappa \rightarrow \bot.$$

Substituting κ for the equivalent $\kappa \rightarrow \bot$ gives us $\kappa \wedge \kappa \rightarrow \bot$. In view of our assumption that $\vdash_T \kappa \leftrightarrow \kappa \wedge \kappa$, another substitution of equivalents yields $\kappa \rightarrow \bot$. By substituting κ for $\kappa \rightarrow \bot$ once again, we get κ. Finally, we use $\rightarrow$-E to derive $\bot$ from $\kappa \rightarrow \bot$ together with κ.

Since VTP and VMP jointly entail the paradox-generating MPA, it would thus appear that rcf theorists can't consistently assert that valid arguments preserve truth.[14] Field (2008, p. 377) and Beall (2009, p. 35) accept the foregoing argument, and consequently reject the claim that valid arguments are guaranteed to preserve truth (assuming, again, that truth-preservation is expressed using a detaching conditional that underwrites the T-Scheme). The need to reject VTP is a perhaps surprising, although ultimately unavoidable, corollary of the revisionary approach to paradox, or so they argue.[15]

---

[12] This principle is endorsed by Field (2008, p. 253) and Beall (2009, pp. 28, 35).

[13] Following Restall (1994), this is sometimes referred to as *pseudo Modus Ponens*. See also Priest (1980), where it is described as the "counterfeit" *Modus Ponens* axiom.

[14] See Beall (2007); Beall (2009, pp. 34–41], Shapiro (2011, p. 341) and Beall and Murzi (2013).

[15] For Field, who rejects excluded middle, rejecting an instance of VTP doesn't mean accepting its negation. Beall, by contrast, accepts that there are valid arguments, e.g. the argument from κ and $\kappa \rightarrow \bot$ to $\bot$, that fail to preserve truth. However, as Field and Beall both note, Beall's position *doesn't* require accepting that there are valid arguments whose premises are all true and whose conclusion is false. See Field (2006, p. 597) and Beall (2009, p. 36).

22.1.3 From VTP to Inconsistency via the Consistency Argument

A second argument for rejecting VTP (Field 2006, 2008, 2009b) proceeds via Gödel's Second Incompleteness Theorem, which states that no consistent recursively axiomatisable theory containing a modicum of arithmetic can prove its own consistency. Field first argues that *if* an otherwise suitable semantic theory could prove that all its rules of inference preserve truth, it could prove its own consistency. Hence, by Gödel's theorem, no semantic theory that qualifies as a "remotely adequate mathematical theory" can prove that its rules of inference preserve truth. Yet insofar as we endorse the orthodox semantic principle VTP, Field says, we should be able to consistently add to our semantic theory an axiom stating that its rules of inference preserve truth (see Field 2009a, p. 351n10). Hence, he concludes, we should reject VTP.

To establish the first step in this argument against VTP, Field considers what he calls the Consistency Argument (Field 2006, pp. 567–568). This is an argument which, one might think, one should be able to run *within* any theory T containing a truth predicate satisfying the unrestricted T-Scheme. The argument proceeds by "(i) inductively proving within T that all its theorems are true, and (ii) inferring from the truth of all theorems of T that T is consistent." Though intuitively sound, the Consistency Argument must fail if T is to be consistent.

Field's claim is that the failure of the Consistency Argument must be blamed on an illicit appeal to VTP. He observes that (ii) can't be problematic for those who hold that "inconsistencies imply everything." The target theories "certainly imply $\neg Tr(\ulcorner 0 = 1 \urcorner)$, so the soundness of T would imply that '$0 = 1$' isn't a theorem of T; and this implies that T is consistent" (Field 2008, p. 286–287). However, (ii) will be equally unproblematic for any *paraconsistent* theorist who holds that an adequate semantic theory must imply the universal generalization over instances of the schema $\neg Tr(\ulcorner \alpha \wedge \neg\alpha \urcorner)$. In this case as well, if T could prove that all its theorems are true, it would thereby prove that no contradiction is a theorem (Field 2006, pp. 593–595). Field therefore concludes that the problem with the Consistency Argument must lie with (i). The argument by induction alluded to in (i) proceeds as follows: "(1) Each axiom of T is true, (2) Each rule of inference of T preserves truth [in the sense of VTP, whence] (3) All theorems of T are true." Field argues persuasively that "[t]he only place that the argument can conceivably go wrong is ... in (2)" (Field 2008, p. 287). This conclusion is endorsed by Beall (2009, pp. 115–116).

In sum, not only does the seemingly obvious Validity Argument in favor of VTP fail, but there are at least two arguments against accepting VTP—or so contemporary revisionary wisdom goes. As Beall writes: "such a claim ... needs to be rejected, and I reject it" (Beall 2009, p. 35).

22.2 Naïve Validity and Validity Curry

What role, then, if any, is left for the notion of validity, if we can no longer affirm that valid arguments preserve truth? Field (2008, 2009b, 2015) suggests that validity normatively constrains belief: very roughly, one shouldn't fully believe the premises of a valid argument without fully believing its conclusion. We take no position here on whether the role of the notion of validity can be explained without recourse to truth-preservation.[16] Instead, we'll suggest in the remainder of this paper that revisionary theorists *need not* and *should not* reject VTP. Provided they accept certain basic principles that would appear to govern the notion of validity, revisionary theorists are required on pain of paradox to adopt the very kind of logic that allows them to embrace VTP.

22.2.1 Naïve Validity

Still restricting our attention to single-premise arguments, consider the following two principles for the use of the validity predicate: that, if one can derive ψ from ϕ, one can derive on no assumptions that the argument from ϕ to ψ is valid, and that, from ϕ and the claim that the argument from ϕ to ψ is valid, one can infer ψ.[17]

Both rules are highly intuitive. If $Val(x, y)$ expresses *validity*, it seems natural to assume that an adequate semantic theory T must include the following introduction rule for $Val(x, y)$, which, by analogy with $\rightarrow$-I or Conditional Proof, we'll call *Validity Proof*:

$$(\mathsf{VP})\frac{\alpha \vdash_T \beta}{\vdash_T Val(\ulcorner \alpha \urcorner, \ulcorner \beta \urcorner)}$$

If T's rules are valid, and we can derive β from α in T, then T must be able to assert the sentence $Val(\ulcorner \alpha \urcorner, \ulcorner \beta \urcorner)$, expressing that the argument from α to β is valid. But it also seems natural to assume that T contains an elimination rule for $Val(x, y)$, which we'll call *Validity Detachment*:

$$(\mathsf{VD})\frac{\Gamma \vdash_T Val(\ulcorner \alpha \urcorner, \ulcorner \beta \urcorner) \quad \Delta \vdash_T \alpha}{\Gamma, \Delta \vdash_T \beta}$$

[16] For the record, we think that even if VTP holds, an explanation of the role of the notion of validity will have to involve normative considerations such as those Field advances.

[17] To the best of our knowledge, these rules are first discussed in Priest (2010). For further discussion, see Beall and Murzi (2013) and Murzi (2014). Shapiro (2011) proposes introducing a validity predicate governed by the equivalences $Val(\ulcorner \alpha \urcorner, \ulcorner \beta \urcorner) \dashv\vdash_T \alpha \Rightarrow \beta$, where $\Rightarrow$ is an *entailment connective* whose introduction and elimination rules in turn render VP and VD derivable. Such a connective is common in the tradition of relevant and paraconsistent logic: see e.g. Anderson and Belnap (1975, p. 7) and Priest and Routley (1982).

If, from a given context of assumptions, we can derive in T the sentence α and from another context we can derive that the argument from α to β is valid, then it must be possible (from the assumptions taken together) to derive β.[18]

The rules VP and VD can also be viewed as *generalizations* of natural rules for a predicate that expresses logical truth: namely, analogues of the rule of Necessitation and of a rule corresponding to the T axiom. To see this, it is sufficient to instantiate VP and VD using a constant $\top$ expressing logical truth. Instantiating VP yields a notational variant of Necessitation, rewritten using our two place predicate $Val(x, y)$ in place of a necessity operator:

$$(\mathsf{NEC}^*)\frac{\top \vdash_T \beta}{\vdash_T Val(\ulcorner\top\urcorner, \ulcorner\beta\urcorner)}$$

Likewise, instantiating VD thus

$$\frac{\Gamma \vdash_T Val(\ulcorner\top\urcorner, \ulcorner\beta\urcorner) \quad \top \vdash_T \top}{\Gamma, \top \vdash_T \beta}$$

yields a notational variant of a rule corresponding to the T axiom for a necessity operator:

$$(\mathsf{T}^*)\ \ Val(\ulcorner\top\urcorner, \ulcorner\beta\urcorner), \top \vdash_T \beta$$

The intuitiveness of our rules VP and VD is thus underscored by the close connection they underwrite between the behavior of a predicate expressing logical truth and the behavior of an operator expressing logical necessity.

We will therefore call the view that 'valid' satisfies VP and VD the *naïve view of validity* (Murzi 2014). One first point that deserves emphasis is that, on the naïve view of truth we've assumed at the beginning of this paper, such a view entails V0, our object-language statement of VTP for single-premise arguments. This can be shown using what is essentially a version of Field's Validity Argument, except that the validity of the argument from α to β is now expressed using an object-language predicate rather than using a turnstile in the metalanguage:[19]

---

[18] We have written the rule VP without side assumptions. That is because the acceptability of a version including side assumptions

$$(\mathsf{VP}^*)\frac{\Gamma, \alpha \vdash_T \beta}{\Gamma \vdash_T Val(\ulcorner\alpha\urcorner, \ulcorner\beta\urcorner)}$$

depends on the properties of the structural comma. For example, if the comma obeys *weakening* and we get $\beta, \alpha \vdash_T \beta$, then VP* allows us to derive $\beta \vdash_T Val(\ulcorner\alpha\urcorner, \ulcorner\beta\urcorner)$. But where β is contingent, it shouldn't follow from β that it is entailed by any sentence. A similar problem arises if the comma obeys *exchange*. From VD and Cut we get $Val(\ulcorner\alpha\urcorner, \ulcorner\alpha\urcorner), \alpha \vdash_T \alpha$, whence exchange yields $\alpha, Val(\ulcorner\alpha\urcorner, \ulcorner\alpha\urcorner) \vdash_T \alpha$ and VP* allows us to derive $\alpha \vdash_T Val(\ulcorner Val(\ulcorner\alpha\urcorner\urcorner, \ulcorner\alpha\urcorner), \ulcorner\alpha\urcorner)$. But if α is contingent, it shouldn't follow from α that it is entailed by a logical truth. Zardini (2013), whose comma obeys both weakening and exchange, avoids these problems by restricting the side assumptions in VP* to logical compounds of *validity claims*. See also Priest and Routley (1982).

[19] Ripley (2013) offers a similar defense of VTP, using VP and the sequent $\alpha, Val(\ulcorner\alpha\urcorner, \ulcorner\beta\urcorner) \vdash_T \beta$. Shapiro (2011) explains that on the version of the naïve view presented there (see note 17 above), $Val(\ulcorner\alpha\urcorner, \ulcorner\beta\urcorner)$ implies $Tr(\ulcorner\alpha\urcorner) \Rightarrow Tr(\ulcorner\beta\urcorner)$.

$$\dfrac{\dfrac{\dfrac{\dfrac{Val(\ulcorner\alpha\urcorner,\ulcorner\beta\urcorner) \vdash_T Val(\ulcorner\alpha\urcorner,\ulcorner\beta\urcorner) \qquad \dfrac{Tr(\ulcorner\alpha\urcorner) \vdash_T Tr(\ulcorner\alpha\urcorner)}{Tr(\ulcorner\alpha\urcorner) \vdash_T \alpha}\,Tr\text{-E}}{Val(\ulcorner\alpha\urcorner,\ulcorner\beta\urcorner), Tr(\ulcorner\alpha\urcorner) \vdash_T \beta}\,\mathsf{VD}}{Val(\ulcorner\alpha\urcorner,\ulcorner\beta\urcorner), Tr(\ulcorner\alpha\urcorner) \vdash_T Tr(\ulcorner\beta\urcorner)}\,Tr\text{-I}}{Val(\ulcorner\alpha\urcorner,\ulcorner\beta\urcorner) \vdash_T Tr(\ulcorner\alpha\urcorner) \to Tr(\ulcorner\beta\urcorner)}\,{\to}\text{-I}}{\vdash_T Val(\ulcorner\alpha\urcorner,\ulcorner\beta\urcorner) \to (Tr(\ulcorner\alpha\urcorner) \to Tr(\ulcorner\beta\urcorner))}\,{\to}\text{-I}$$

A second point to notice is that, natural though they may seem, VP and VD lead us into trouble—which should of course be expected, since NEC^* and T^* are nothing but the key ingredients of the Myhill–Kaplan-Montague Paradox, or Paradox of the Knower (Myhill 1960; Kaplan and Montague 1960; Murzi 2014).[20]

22.2.2 Validity Curry

The Diagonal Lemma allows us to construct a sentence π, which intuitively says of itself, up to equivalence, that it validly entails that you will win the lottery:

$$\vdash_T \pi \leftrightarrow Val(\ulcorner\pi\urcorner,\ulcorner\bot\urcorner)$$

Let Σ now be the following derivation of the further theorem $Val(\ulcorner\pi\urcorner,\ulcorner\bot\urcorner)$:

$$\dfrac{\dfrac{\dfrac{\pi \vdash_T \pi \qquad \vdash_T \pi \leftrightarrow Val(\ulcorner\pi\urcorner,\ulcorner\bot\urcorner)}{\pi \vdash_T Val(\ulcorner\pi\urcorner,\ulcorner\bot\urcorner)}\,{\to}\text{-E} \qquad \pi \vdash_T \pi}{\dfrac{\dfrac{\pi, \pi \vdash_T \bot}{\pi \vdash_T \bot}\,\mathsf{SContr}}{\vdash_T Val(\ulcorner\pi\urcorner,\ulcorner\bot\urcorner)}\,\mathsf{VP}}\,\mathsf{VD}$$

Using Σ, we can then 'prove' that you will win the lottery

[20] Shapiro (2011) identifies two challenges to the naïve view: a "direct argument" that it leads straight to paradox, and an "indirect argument" that it entails a version of the paradox-producing VTP.

$$\dfrac{\dfrac{\Sigma}{\vdash_T Val(\ulcorner\pi\urcorner,\ulcorner\bot\urcorner)} \qquad \dfrac{\vdash_T \pi \leftrightarrow Val(\ulcorner\pi\urcorner,\ulcorner\bot\urcorner) \qquad \dfrac{\Sigma}{\vdash_T Val(\ulcorner\pi\urcorner,\ulcorner\bot\urcorner)}}{\vdash_T \pi}\,\mathsf{VD}}{\vdash_T \bot}\,{\to}\text{-E}$$

Our revisionary theory of truth and validity, T, proves on no assumptions that you will win the lottery.[21] Call this the Validity Curry, or v-Curry, for short, to contrast it with the standard conditional-involving version of Curry's Paradox, or c-Curry.[22]

As we explained above, rcf theorists invalidate c-Curry by rejecting $\to$-I. Unlike c-Curry, however, the v-Curry Paradox makes no use of $\to$-I, and hence it cannot be invalidated by rejecting such a rule. On the other hand, the above derivation of v-Curry presupposes SContr (Beall and Murzi 2013). Hence if VP and VD hold, there is only one revisionary way out of the v-Curry Paradox, viz. rejecting SContr, thus adopting a *substructural* logic—a logic where some of the standardly accepted structural rules fail (Shapiro 2011; Beall and Murzi 2013; Murzi 2014; Zardini 2011).[23]

Before examining in §3 how rejecting SContr affects VTP and the Validity Argument, we'd first like to offer a partial defence of our claim that v-Curry Paradox is a reason for revisionary logician to adopt a substructural logic. To this end, we'll consider in the next section two natural responses to the claim that the Validity Curry is a genuine paradox of validity, and offer replies on the substructural logician's behalf.

22.2.3 A Genuine Paradox of Validity

If the v-Curry Paradox isn't a genuine paradox of validity, one of VP and VD must not unrestrictedly hold. As it turns out, there are *prima facie* compelling reasons

[21] To the best of our knowledge, the first known occurrence of the Validity Curry is in the 16th-century author Jean de Celaya. See Read (2001, fn. 11–12) and references therein. Albert of Saxony includes a contrapositive version of the paradox among his "insolubles" (Read 2010, p. 165). A more recent version can be found in Priest and Routley (1982), and surfaces again in Whittle (2004, fn. 3), Clark (2007, pp. 234–235) and Shapiro (2011, fn. 29). For a first comprehensive discussion of the Validity Curry, see Beall and Murzi (2013). For a defence of the claim that Validity Curry is a genuine paradox of validity, see §2.3 below and Murzi (2014).

[22] This terminology was first introduced in Beall and Murzi (2013). Ultimately, however, the distinction in terms of predicate versus connective may not be the essential one. Whittle (2004) and Shapiro (2011) discuss a version of Curry's Paradox, involving a "consequence connective" or "entailment connective," which poses much the same challenge to rcf theorists as does v-Curry. [See also Shapiro (2015, p. 82).]

[23] For an early anticipation of the argument from naïve validity to the rejection of SContr (in the form of multiple discharge of assumptions), see Priest and Routley (1982). Priest and Routley, whose entailment connective obeys analogues of VP and VD, discuss several resulting paradoxes which they blame on the "suppression of innocent premises." By contrast, Ripley (2013) blocks v-Curry at the final step using VD, which is inadmissible in his nontransitive theory for the same reason that $\to$-E is inadmissible. See note 46 below.

for restricting both.[24] One argument against VP runs thus. In order to establish $Val(\ulcorner\alpha\urcorner,\ulcorner\beta\urcorner)$ as a theorem using "Validity Proof," it is said, we should be required to produce a logically valid argument from α to β. Yet subderivation Σ above doesn't establish the argument from π to $\bot$ as *logically* valid, for two reasons. First, this subderivation relies on a substitution instance of the *logically invalid* biconditional proved by the Diagonal Lemma, viz. $\pi \leftrightarrow Val(\ulcorner\pi\urcorner,\ulcorner\bot\urcorner)$. Second, it uses VD, and, it might be objected, surely such a rule isn't logical. More precisely, Roy Cook (2012) has argued that the T-Scheme isn't logically valid, if by logical validity one means truth under all uniform interpretations of the non-logical vocabulary. Using the same reasoning, we could conclude that VD doesn't preserve logical validity.[25]

These objections have an important virtue: they help us understand what the v-Curry Paradox really is a paradox of. More precisely, they show that the v-Curry Paradox is not paradox of *purely logical*, or *interpretational*, in John Etchemendy's term, validity (Etchemendy 1990).[26] Indeed, a recent result by Jeff Ketland shows that purely logical validity *cannot* be paradoxical. Ketland (2012) proves that Peano Arithmetic (PA) can be conservatively extended by means of a predicate expressing logical validity, governed by intuitive principles that are themselves derivable in PA. It follows that purely logical validity is a consistent notion if PA is consistent, which should be enough to warrant belief that purely logical validity simply *is* consistent.

However, it seems to us that there are broader notions of validity than purely logical validity.[27] Thus, neither of the above objections applies to versions of the v-Curry Paradox in which 'valid' expresses *representational* validity, whereby (roughly) validity is equated with preservation of truth in all possible circumstances (Read 1988; Etchemendy 1990; McGee 1991). In this sense, at least intuitively, the arithmetic required to prove the Diagonal Lemma is valid and VD is validity-preserving.[28]

---

[24] Thanks to Roy Cook and Jeff Ketland for raising these potential concerns.

[25] Field (2008, §20.4) himself advances versions of this line of argument, while discussing what is in effect a validity-involving version of the Knower Paradox resting on NEC^* and T^*. See especially Field (2008, p. 304 and p. 306). On the question whether his conception of the extension of the validity predicate consistently allows him to do so, see note 27 below.

[26] Here we take the logical vocabulary to be the standard vocabulary of some first-order, perhaps non-classical, logic.

[27] Several semantic theorists, including rcf theorists such as Field and Priest, resort to notions of validity that are not purely logical. For instance, Field (2007, 2008) extensionally identifies validity with, essentially, preservation of truth in all ZFC models of a certain kind, thus taking validity to (wildly) exceed purely logical validity. (Incidentally, it seems to us that this use of 'valid' is in tension with the purely logical sense Field (2008) appeals to at p. 304 and especially p. 306.) Likewise, McGee (1991, p. 43–49) takes logical necessity to extend to arithmetic and truth-theoretic principles.

[28] It might be objected that such a notion of validity presupposes VTP, and hence cannot be appealed to in the present context, where the question whether VTP can be consistently upheld is the very point at issue. Our modest aim here, however, is simply to suggest that someone who *already thinks*, following perhaps logical orthodoxy, that valid arguments preserve truth and that, accordingly, consequence is to be explicated in terms of truth-preservation has a reason—the v-Curry Paradox—to reject SContr. Once SContr is rejected, the standard challenges to VTP no longer stand, as we'll see in §3 below. But, it seems to us, no illicit or question-begging appeal to

Nor does the objection that VP cannot be legitimately applied to non-purely-logical subderivations apply to conceptions of validity which take 'valid' to express the consequence relation of one's semantic theory, provided that the naïve validity rules and enough arithmetic are part of that relation.[29] Insofar as VD preserves validity in one of these broader senses, and insofar as the VP and VD govern the use a predicate expressing validity in that sense, there is at least one—important—reading of 'valid' on which the use of VP in the v-Curry derivation is sound. The v-Curry Paradox is a paradox of *validity*, not *purely logical* validity.

To be sure, one might instead either reject VP on different grounds, or perhaps reject VD. One natural enough argument against the latter rule runs thus. Suppose validity is recursively enumerable. Then, one might argue, T^*, and hence VD, must fail. For, if validity is recursively enumerable, an argument is valid if and only if its conclusion can be derived from its premises in some recursively axiomatisable theory T. That is, the validity predicate $Val(x, y)$ is just a notational variant of $Prov_T(x, y)$, where this expresses that there is a T-derivation of y from x. Yet, the argument continues, we know from Löb's Theorem that, if T contains enough arithmetic (if it proves the so-called derivability conditions), T cannot contain, on pain of triviality, all instances of the provability-in-T analogue of T^*, $Prov_T(\ulcorner \top \urcorner, \ulcorner \alpha \urcorner) \rightarrow \alpha$. Hence, one might conclude, T may not contain all instances of T^* either, and hence of VD, *a fortiori*.

We find this conclusion problematic. It seems to us that rejecting VD, or VP, for that matter, isn't really a comfortable option for proponents of the naïve view of truth. In a nutshell, together with the naïve view of truth, the naïve view of validity is but an instance of the general thought underpinning the revisionary approach to paradox—what we may call the *naïve view of semantic properties*.[30] This is the view that one cannot revise naïve semantic principles without thereby also revising naïve semantic properties, and that, on pain of triviality, semantic properties should be held fixed, and *logic* must change. Arguably, the naïve view of semantic properties has it that validity *is* factive, and that we, and hence our semantic theory, must be able to say so, on pain of not being able to consistently assert what we know to be true. If T does indeed meet the conditions for Löb's Theorem, we would like to

VTP has been made in the course of the foregoing reasoning. We thank an anonymous referee for raising this potential concern.

[29] In fact, Cook (2014) shows how this response can be strengthened: it is possible to formulate a modified Validity Curry paradox in such a way that the arithmetic necessary to prove the Diagonal Lemma need not be included in the scope of the validity relation.

[30] This view is implicitly assumed in the work of contemporary revisionary theorists—see e.g. Priest (2006b); Field (2007, 2008), Beall (2009), Beall and Murzi (2013). In particular, it is implicit in their assumption that the paradoxes of validity are (in an interesting sense) of the same kind as the Liar and c-Curry. One defence of that possibly controversial assumption would involve arguing that the Paradox of the Knower is nothing but a weakened Liar, and that, as we've observed in §2.1, v-Curry is nothing but a generalised Knower, so that whatever the nature of the first paradox, it is inherited by the other two. See also Read (2001) and Beall and Murzi (2013). We should finally stress that in calling validity a semantic property, we merely intend to point to these parallels, without relying on any particular conception of what makes a property semantic.

suggest, then the correct reaction to the objection is instead to concede that $Val(x, y)$ can't be replaced with $Prov_T(x, y)$, and hence that naïve validity is not recursively enumerable.[31]

It might be objected that we could revise, or refine, our naïve conception of validity, which is after all naïve (McGee 1991, p. 45). But, then, a parallel argument would show that, when faced with the Liar Paradox, the c-Curry Paradox, and other paradoxes of truth, we should similarly revise our conception of *truth*, which is precisely what proponents of the naïve view of semantic properties take to be the *wrong* response to semantic paradox. For the time being, we'll assume that the Validity Curry is a genuine paradox of validity, and that giving up SContr, as suggested in Shapiro (2011) and Zardini (2011), is a legitimate revisionary response to it, and to semantic paradoxes more generally. We shall now argue that, on this admittedly controversial assumption, of which we've only offered a partial defence, all three arguments for rejecting VTP break down.

22.3 Validity and Truth-Preservation

All three challenges to VTP turn out to rest crucially on how our object-language expresses validity and truth-preservation for arguments with *multiple premises*. First, recall that Field argues that the most obvious defense of VTP, the Validity Argument, rests on principles that yield paradox. As we have pointed out, the Validity Argument presupposes that the truth-preservingness of an inference from $\alpha_1, ..., \alpha_n$ to β can be expressed using the object-language sentence $Tr(\ulcorner \alpha_1 \urcorner) \wedge ... \wedge Tr(\ulcorner \alpha_n \urcorner) \rightarrow Tr(\ulcorner \beta \urcorner)$. Second, the argument from VTP to PMP and absurdity used the simplifying assumption that the validity of the two-premise *Modus Ponens* rule can be expressed using a single-premise validity predicate as $Val(\ulcorner \alpha \wedge (\alpha \rightarrow \beta) \urcorner, \ulcorner \beta \urcorner)$. Finally, spelling out the Consistency Argument requires expressing in the object-language the claim that each of our semantic theory T's rules of inference preserves truth, where these will include multi-premise rules such as *Modus Ponens*.

[31] We don't have space to expand on this point here. Priest (2006b, §3.2) argues at length that the "naïve notion of proof" is recursive, whence naïve provability, a species of naïve validity, is recursively enumerable. Here we simply notice that his arguments are consistent with the view that naïve *validity* isn't. Finally, we'd like to point out that some SContr-free semantic theories extending contraction-free arithmetics may not be strong enough to satisfy Löb's Theorem's applicability conditions, in which case the objection from Löb's Theorem we are considering would not apply in the first place.

22.3.1 Premise-Aggregating Connectives

We will therefore assume that truth-preservation and validity for arguments with a finite number of premises can be expressed using some "premise-aggregating connective" $\odot$:[32]

(a) The claim that the argument from premises $\alpha_1, ..., \alpha_n$, taken together, to conclusion β preserves truth can be expressed in the object-language as $Tr(\ulcorner\alpha_1\urcorner) \odot ... \odot Tr(\ulcorner\alpha_n\urcorner) \rightarrow Tr(\ulcorner\beta\urcorner)$.
(b) The claim that the argument from premises $\alpha_1, ..., \alpha_n$, taken together, to conclusion β is valid can be expressed using the object-language's binary validity predicate as $Val(\ulcorner\alpha_1 \odot ... \odot \alpha_n\urcorner, \ulcorner\beta\urcorner)$.

Is there an understanding of the logical behavior of $\odot$ on which (a) and (b) are true, but each of our three challenges to VTP is blocked?

Before examining the three challenges in turn, we now consider the chief options for the rules governing $\odot$ in the context of a substructural natural deduction system. For the time being, we will work within a structural framework in which the "taking together" of assumptions—which we have indicated with commas to the left of the turnstile—can be represented using "multisets." These are structures that behave like sets except for the fact that they keep track of the number of occurrences of each member (Meyer and McRobbie 1982a, 1982b). The philosophical significance of multiset structure in natural deduction has been explained in many ways, and the same is the case for the more complex structure we will consider later. This isn't the place to compare various interpretations or defend one of them.[33] Our aim, rather, is to explain how moving to a deduction system in which the structure referred to on the left of the turnstile is finer-grained than a set affects the standard objections to VTP.

Using multisets rather than (e.g.) sequences renders redundant Gentzen's structural rule of *exchange*:

$$(\mathsf{SExch})\frac{\Gamma, \alpha, \beta \vdash \gamma}{\Gamma, \beta, \alpha \vdash \gamma}$$

By contrast, SContr isn't redundant, nor is the structural rule of *weakening*:

$$(\mathsf{SWeak})\frac{\Gamma, \alpha \vdash \gamma}{\Gamma, \beta, \alpha \vdash \gamma}$$

Indeed, once one or more of SContr and SWeak is rejected, one can formulate operational rules for two different connectives, rules that become equivalent only in the presence of both SContr and SWeak. These are the rules that govern, respectively,

[32] For arguments with an infinite number of premises, we will need universal quantification to express truth-preservation. None of the objections to VTP we will consider, however, depend on consideration of infinite-premise arguments.

[33] We have each made different suggestions in previous work: Shapiro (2011) and Beall and Murzi (2013). [The interpretation of structure sketched in Shapiro (2011) is elaborated in Shapiro (2015).] See also Read (1988), Slaney (1990), Restall (2000), and Paoli (2002).

the "multiplicative" and "additive" conjunctions of linear logic, a multiset-based logic in which both SWeak and SContr are rejected (Girard 1987):[34]

$$(\otimes\text{-I})\frac{\Gamma \vdash \alpha \quad \Delta \vdash \beta}{\Gamma, \Delta \vdash \alpha \otimes \beta} \quad (\otimes\text{-E})\frac{\Gamma, \alpha, \beta \vdash \gamma \quad \Delta \vdash \alpha \otimes \beta}{\Gamma, \Delta \vdash \gamma}$$

$$(\&\text{-I})\frac{\Gamma \vdash \alpha \quad \Gamma \vdash \beta}{\Gamma \vdash \alpha \,\&\, \beta} \quad (\&\text{-E1})\frac{\Gamma \vdash \alpha \,\&\, \beta}{\Gamma \vdash \alpha} \quad (\&\text{-E2})\frac{\Gamma \vdash \alpha \,\&\, \beta}{\Gamma \vdash \beta}$$

Since it will be important later, we note that the structural comma appears in the rules for the multiplicative $\otimes$, whereas it does not appear in the rules for the additive &. In the terminology of Belnap (1982, 1993), the additive rules are "structure-free" while the multiplicative rules are "structure-dependent." Finally, in this structural setting, our assumption of the transitivity of validity can be codified using the following version of the cut rule:

$$(\text{Cut})\frac{\Gamma \vdash \alpha \quad \Delta, \alpha \vdash \beta}{\Delta, \Gamma \vdash \beta}$$

22.3.2 The Validity Argument

The first point we would like to make is that, in the absence of SContr, the 'only if' direction of the Validity Argument (the direction that would establish VTP) fails when the premise-aggregating connective $\odot$ is construed as the additive & in a multiset-based logic.

To see why, note that when rewritten using &, this direction of the Validity Argument requires deriving $Tr(\ulcorner \alpha_1 \urcorner)$ & ... & $Tr(\ulcorner \alpha_n \urcorner) \vdash Tr(\ulcorner \beta \urcorner)$ from $Tr(\alpha_1), ..., Tr(\alpha_n) \vdash Tr(\beta)$. That in turn requires $n - 1$ uses of the inference pattern

$$(\&\text{-L})\frac{\Gamma, \alpha_1, \alpha_2 \vdash \beta}{\Gamma, \alpha_1 \,\&\, \alpha_2 \vdash \beta}$$

Field himself justifies this inference by appeal to the rule &-E. Indeed, in the presence of SContr, either of our twin elimination rules &-E1 and &-E2 yields & -L. Here is a derivation using &-E2, SContr, Cut, and the reflexivity of validity:

$$\dfrac{\dfrac{\dfrac{\Gamma, \alpha_1, \alpha_2 \vdash \beta \qquad \dfrac{\alpha_1 \,\&\, \alpha_2 \vdash \alpha_1 \,\&\, \alpha_2}{\alpha_1 \,\&\, \alpha_2 \vdash \alpha_2}\,\&\text{-E2}}{\Gamma, \alpha_1, \alpha_1 \,\&\, \alpha_2 \vdash \beta}\,\text{Cut} \qquad \alpha_1 \,\&\, \alpha_2 \vdash \alpha_1}{\Gamma, \alpha_1 \,\&\, \alpha_2, \alpha_1 \,\&\, \alpha_2 \vdash \beta}\,\text{Cut}}{\Gamma, \alpha_1 \,\&\, \alpha_2 \vdash \beta}\,\text{SContr}$$

[34] While linear logic is standardly presented in sequent calculus format, the above natural deduction rules appear in Avron (1988, p. 165), Troelstra (1992, p. 57) and O'Hearn and Pym (1999).

In a logic without SContr, on the other hand, & -L fails. Moreover, this remains the case if we accept SWeak, thus strengthening linear logic into what is known as an "affine" logic.[35]

Hence, insofar as we wish to preserve the Validity Argument while rejecting SCont (and thus avoiding c-Curry and v-Curry), we ought not interpret the premise-aggregating $\odot$ as the additive conjunction & of a multiset-based logic. On the other hand, both directions of the Validity Argument go through, even in the absence of SContr, provided that $\odot$ is construed as the multiplicative $\otimes$. Given $\alpha_1 \otimes \alpha_2 \vdash \alpha_1 \otimes \alpha_2$, the rule $\otimes$-E immediately yields the inference required for the argument's "only if" direction:[36]

$$(\otimes\text{-L})\frac{\Gamma, \alpha_1, \alpha_2 \vdash \beta}{\Gamma, \alpha_1 \otimes \alpha_2 \vdash \beta}$$

Indeed, with $\otimes$ as premise-aggregating connective, Elia Zardini (2011) has recently proved a generalization of the Validity Argument's "only if" conclusion.[37] And the "if" direction is no harder to establish.

Summarizing, we can say that Field's objection to the "only if" direction of the Validity Argument fails when our semantic theory is based on an underlying logic that lacks SContr, as long as this logic is multiset-based and we state the argument's conclusion using multiplicative conjunction. Admittedly, this method of vindicating the Validity Argument carries a cost. Multiset-based logics can contain no connectives that behave like the conjunction or disjunction of classical logic (see e.g. Belnap 1993). In the case of the additive connectives, for example, we lose *Distribution*: $\alpha \,\&\, (\beta \vee \gamma) \vdash (\alpha \,\&\, \beta) \vee (\alpha \,\&\, \gamma)$. On the multiplicative side, besides losing distribution of $\otimes$ over a corresponding multiplicative disjunction, we lose *Simplification*: $\alpha \otimes \beta \vdash \alpha$. Adding the rule SWeak, as Zardini proposes, restores the latter. But, as we will see below, we still lose *Square-increasingness*: $\alpha \vdash \alpha \otimes \alpha$.

However, adopting a multiset-based logic isn't the only way to vindicate the Validity Argument by rejecting a structural contraction rule. A second way is to use one of the many substructural logics in which assumptions are regarded as "taken together" in two different ways. In such "dual-bunching" logics, the structures referred to on the left of the turnstile are not multisets, but rather finer-grained "bunches" specified using two different punctuation marks (Read 1988; Slaney 1990; Restall 2000). This alternative is of interest for two reasons. First, unlike multiset-based logics, dual-bunching logics do feature connectives whose behavior is classical to the extent

[35] In that case, however, the "if" direction of the Validity Argument *will* go through for & as premise-aggregating connective. Deriving $Tr(\ulcorner\alpha_1\urcorner), ..., Tr(\ulcorner\alpha_n\urcorner) \vdash Tr(\ulcorner\beta\urcorner)$ from $Tr(\ulcorner\alpha_1\urcorner)$ & ... & $Tr(\ulcorner\alpha_n\urcorner) \vdash Tr(\ulcorner\beta\urcorner)$ requires the inverse of &-L, which obtains in the presence of SWeak.

[36] In single-conclusion sequent calculus formulations (which suffice for our purposes, as our derivations all involve the language's negation-free fragment), the connective $\otimes$ is governed by the twin rules $\otimes$-I and $\otimes$-L.

[37] Field's own reasoning, as sketched in §1.3, amounts to a special case of Zardini's proof: the case in which we are considering the truth-preservingness of a single-conclusion argument and employ no side assumptions. Zardini's proof does not depend on his acceptance of SWeak.

that they satisfy Distribution, Simplification, and Square-increasingness. Secondly, as we will see in §3.3, multiset-based and dual-bunching logics underwrite different interpretations of the way in which rejecting structural contraction blocks the argument against VTP via the *Modus Ponens* axiom.

The first kind of bunching is used to formulate all the *structure-dependent operational rules*. For this reason, it will be convenient to indicate this kind of bunching using the comma (though the semicolon is more standard). That way, we can retain our rules $\rightarrow$-I, $\rightarrow$-E, VD and $\otimes$-I, as long as Γ and Δ are now understood as bunches rather than multisets. On the other hand, we need a generalized version of $\otimes$-E, where $\Delta(\alpha, \beta)$ stands for any bunch of which α, β is a subbunch:[38]

$$(\otimes\text{-E}_{db})\frac{\Delta(\alpha, \beta) \vdash \gamma \quad \Gamma \vdash \alpha \otimes \beta}{\Delta(\Gamma) \vdash \gamma}$$

In dual-bunching logics, one or more of the standard structural rules SContr, SWeak or SExch is rejected for the comma.[39] Just as for multiset-based logics, rejecting SContr suffices to block the above derivations of c-Curry and v-Curry.

What is distinctive about dual-bunching logics is the introduction of a second kind of bunching of assumptions, which we will indicate using the colon. This "extensional" bunching *obeys all the standard structural rules*:

$$(\text{eSContr})\frac{\Gamma(\Delta : \Delta) \vdash \beta}{\Gamma(\Delta) \vdash \beta} \quad (\text{eSWeak})\frac{\Gamma(\Delta) \vdash \gamma}{\Gamma(\Delta' : \Delta) \vdash \gamma} \quad (\text{eSExch})\frac{\Gamma(\Delta : \Delta') \vdash \gamma}{\Gamma(\Delta' : \Delta) \vdash \gamma}$$

Unlike the "intensional" comma, the colon need not get mentioned in operational rules for any connective.[40]

We are now ready to consider how the Validity Argument fares for dual-bunching logics. First, the reasoning challenged by Field goes through provided the conclusion is formulated using the structural comma together with the multiplicative $\otimes$ as the premise-aggregating connective. That is because we retain $\otimes$-L, now generalizable to

$$(\otimes\text{-L}_{db})\frac{\Gamma(\alpha_1, \alpha_2) \vdash \beta}{\Gamma(\alpha_1 \otimes \alpha_2) \vdash \beta}$$

Construed this way, the Validity Argument's "only if" direction establishes that $\alpha_1, ..., \alpha_n \vdash_T \beta$ only if $\vdash_T Tr(\ulcorner \alpha_1 \urcorner) \otimes ... \otimes Tr(\ulcorner \alpha_n \urcorner) \rightarrow Tr(\ulcorner \beta \urcorner)$. Moreover,

---

[38] For definitions, see Read (1988, §4.1) and Restall (2000, pp. 19–20). In sequent calculus formulations, $\otimes$-E$_{db}$ is replaced by $\otimes$-L$_{db}$ below. Sequent calculi of this type were developed independently for fragments of relevant logics by Minc (1976) and by Dunn, whose version appears in Anderson and Belnap (1975, §28.5). For natural deduction formulations, see Read (1988), Slaney (1990) and O'Hearn and Pym (1999), whose use of the comma we follow.

[39] Rather than rejecting the structural rule of *associativity*, we are avoiding the need for such a rule by allowing our comma to retain its variable polyadicity.

[40] Since assumptions can now be embedded in bunches specified using both comma and colon, we also need to generalize our statement of the cut rule:

$$(\text{Cut}_{db})\frac{\Gamma \vdash \alpha \quad \Delta(\alpha) \vdash \beta}{\Delta(\Gamma) \vdash \beta}$$

a parallel result now holds for the connective &, known in this structural context as "extensional" conjunction.[41] This is because the fact that the colon obeys eSContr allows us to replicate the above derivation of &-L, yielding

$$(\ \&\text{-L}_{db})\frac{\Gamma(\alpha_1 : \alpha_2) \vdash \beta}{\Gamma(\alpha_1\ \&\ \alpha_2) \vdash \beta}$$

Accordingly, the Validity Argument also goes through when the conclusion is formulated using the structural colon together with & as the premise-aggregating connective. Construed this way, it establishes that $\alpha_1 : \ldots : \alpha_n \vdash_T \beta$ only if $\vdash_T Tr(\ulcorner\alpha_1\urcorner)\ \&\ \ldots\ \&\ Tr(\ulcorner\alpha_n\urcorner) \rightarrow Tr(\ulcorner\beta\urcorner)$.[42] According to dual-bunching logics, then, there are *different kinds of multi-premise arguments*, represented using different antecedent structure, and the validity of each kind of argument entails a different kind of truth-preservation, expressed in the object-language using different premise-aggregating connectives.

There are thus at least two general ways to vindicate the Validity Argument by rejecting SContr: one can use a multiset-based logic with multiplicative conjunction as premise-aggregating connective, or a dual-bunching logic. Versions of both approaches are known to make possible a naïve theory of truth (either a consistent paracomplete theory or a nontrivial paraconsistent theory).[43] We will return to the difference between the two approaches in the next section. For now, we merely note that they yield logics that conflict for the fragment of the language whose only connectives are & and the corresponding disjunction $\vee$. Recall that the rules for these connectives *don't even mention* the nonstandard comma structure. It follows that on the dual-bunching approach, the *single-premise* validities of this fragment will be exactly those of the corresponding fragment of classical logic. As explained above, this stands in contrast to the conjunctive/disjunctive fragment of additive or

[41] But see (Paoli 2007, pp. 569–571) for opposition to the standard claim that the extensional conjunction of such logics is "truth functional."

[42] The point extends naturally to cases in which the assumptions are aggregated using both kinds of structure. For instance, $\alpha_1 : (\alpha_2,\ \alpha_3) \vdash_T \beta$ only if $\vdash_T Tr(\ulcorner\alpha_1\urcorner)\ \&\ (Tr(\ulcorner\alpha_1\urcorner) \otimes Tr(\ulcorner\alpha_n\urcorner)) \rightarrow Tr(\ulcorner\beta\urcorner)$.

[43] Most work on this issue has concerned the closely parallel case of a naïve set theory featuring an unrestricted axiom of comprehension. For proofs of the consistency or nontriviality of unrestricted comprehension in some "weak relevant logics" that can be specified via dual-bunching natural deduction, see Brady (1983, 1989, 2006). For applications of Brady's techniques to naïve truth-theory, see Priest (1991) and Beall (2009), which do not however consider natural deduction systems. As for multiset-based logics, the consistency of unrestricted comprehension in an affine logic was shown by V. Grishin in 1974: see Došen (1993). For the consistency of a naïve truth theory based on an affine logic, see Zardini (2011).

multiplicative linear logic.[44] The philosophical interpretation of nonstandard antecedent structure—whether dual-bunching or multiset-based—remains a controversial and important issue. However, it isn't one we can address in this paper, which has the more limited aim of exploring how such logics allow a defense of VTP against the various challenges that have been raised against that thesis.[45]

22.3.3 *From VTP to Absurdity via the Modus Ponens Axiom*

We now turn to the objection that VTP entails the *Modus Ponens* axiom, and thus absurdity via c-Curry reasoning. Using a generic premise-aggregating connective, we can state, respectively, the validity of *Modus Ponens* and the *Modus Ponens* axiom as follows:

$$(\mathsf{VMP}_{\odot})\ Val(\ulcorner(\alpha \to \beta) \odot \alpha\urcorner, \ulcorner\beta\urcorner)$$

$$(\mathsf{MPA}_{\odot})\ (\alpha \to \beta) \odot \alpha \to \beta.$$

In §1.2 we saw that VTP, when expressed in the object-language, implies

$$(\mathsf{V1})\ Val(\ulcorner\alpha\urcorner, \ulcorner\beta\urcorner) \to (\alpha \to \beta).$$

It follows that if our naïve semantic theory implies $\mathsf{VMP}_{\odot}$, it also implies the absurdity-threatening $\mathsf{MPA}_{\odot}$. Thus, in order to evaluate the objection, we need to answer two questions:

(1) If we reject SContr, will our semantic theory still imply $\mathsf{VMP}_{\odot}$? Equivalently, in view of VP and VD, will our underlying contraction-free logic still give us $(\alpha \to \beta) \odot \alpha \vdash \beta$?
(2) If we reject SContr, will $\mathsf{MPA}_{\odot}$ still yield absurdity?

---

[44] As Dave Ripley pointed out to us, a dual-bunching logic could also retain a connective $\&_A$ that behaves like the "additive" conjunction and disjunction of a multiset-based logic, for instance in failing to validate Distribution over the corresponding $\vee_A$. To achieve this, replace &-E1 and &-E2 with

$$(\&_A\text{-E1})\frac{\Gamma, \alpha \vdash \gamma \quad \Delta \vdash \alpha \ \&_A\ \beta}{\Gamma, \Delta \vdash \gamma} \quad (\&_A\text{-E2})\frac{\Gamma, \beta \vdash \gamma \quad \Delta \vdash \alpha \ \&_A\ \beta}{\Gamma, \Delta \vdash \gamma}$$

By contrast, in the presence of Cut_{db}, our original &-E1 and &-E2 have the same "extensional" effect as the rules

$$(\&\text{-E1}_{db})\frac{\Gamma(\alpha) \vdash \gamma \quad \Delta \vdash \alpha \ \&\ \beta}{\Gamma(\Delta) \vdash \gamma} \quad (\&\text{-E2}_{db})\frac{\Gamma(\beta) \vdash \gamma \quad \Delta \vdash \alpha \ \&\ \beta}{\Gamma(\Delta) \vdash \gamma}$$

[45] For relevant work on the interpretation of dual-bunching systems, see Read (1988) and Slaney (1990). For a recent and novel suggestion toward an interpretation of multiset-based systems, see Zardini (2011). For a sketch of a more deflationary approach to antecedent structure, see Shapiro (2011). [This sketch has since been elaborated in Shapiro (2015).]

A negative answer to (1) or (2) will show that the objection against VTP fails.[46]

The answers to these questions vary depending on which connective we employ as our $\odot$. For the additive & of a contraction-free logic, the answer to (1) is negative (Restall 1994, pp. 35–36). It should help to display how SContr is involved in the usual derivation:

$$\dfrac{\dfrac{\dfrac{(\alpha \to \beta)\ \&\ \alpha \vdash (\alpha \to \beta)\ \&\ \alpha}{(\alpha \to \beta)\ \&\ \alpha \vdash \alpha \to \beta}\ \&\text{-E} \qquad \dfrac{(\alpha \to \beta)\ \&\ \alpha \vdash (\alpha \to \beta)\ \&\ \alpha}{(\alpha \to \beta)\ \&\ \alpha \vdash \alpha}\ \&\text{-E}}{(\alpha \to \beta)\ \&\ \alpha, (\alpha \to \beta)\ \&\ \alpha \vdash \beta}\ \to\text{-E}}{(\alpha \to \beta)\ \&\ \alpha \vdash \beta}\ \text{SContr}$$

But the objection to VTP fails as well when we use the the multiplicative $\otimes$. This time, the answer to (1) is affirmative:

$$\dfrac{\dfrac{\alpha \to \beta \vdash \alpha \to \beta \qquad \alpha \vdash \alpha}{\alpha \to \alpha, \alpha \vdash \beta}\ \to\text{-E} \qquad (\alpha \to \beta) \otimes \alpha \vdash (\alpha \to \beta) \otimes \alpha}{(\alpha \to \beta) \otimes \alpha \vdash \beta}\ \otimes\text{-E}$$

However, now the answer to (2) is negative. That is because, as already noted in Meyer et al. (1979), the argument from $\text{MPA}_{\odot}$ to absurdity depends essentially on the left-to-right direction of the *Idempotence* law $\vdash \alpha \leftrightarrow \alpha \odot \alpha$. But when we use multiplicative conjunction in a contraction-free logic, we lose this law (Zardini 2011). Again, notice how SContr is involved in its usual derivation:

$$\dfrac{\dfrac{\dfrac{\alpha \vdash \alpha \qquad \alpha \vdash \alpha}{\alpha, \alpha \vdash \alpha \otimes \alpha}\ \otimes\text{-I}}{\alpha \vdash \alpha \otimes \alpha}\ \text{SContr}}{\vdash \alpha \to \alpha \otimes \alpha}\ \to\text{-I}$$

In summary, to derive absurdity from VTP, the objector presupposes that there is some connective $\odot$ that meets two conditions:

---

[46] According to the theory proposed by Ripley (2013) based on Cobreros et al. (2012), which is "substructural" only in rejecting Cut, the objection to VTP we are considering in this section fails because MPA fails to yield absurdity. This is because the argument's final step from $\vdash_T \kappa \to \bot$ and $\vdash_T \kappa$ to $\vdash_T \bot$ fails. In Ripley's sequent calculus, the rule $\to$-E is inadmissible in the absence of Cut. Indeed, Ripley holds (p.c) that $\to$-E shouldn't be regarded as fundamental to the logic of a detaching conditional, as it covertly builds in extraneous transitivity in comparison with the sequent calculus rule

$$(\to\text{-L})\dfrac{\Gamma \vdash \alpha \quad \Delta, \beta \vdash \gamma}{\Delta, \alpha \to \beta, \Gamma \vdash \gamma}$$

To this, defenders of $\to$-E may reply that *each of* $\to$-E and $\to$-L builds in transitivity in comparison with $\alpha \to \beta, \alpha \vdash \beta$. It is true, as Ripley shows, that the transitivity built in by $\to$-E (which, given $\to$-I, yields Cut) can be blamed for paradox. But in view of the option of blaming paradox on SContr instead, this won't suffice to show that $\to$-L is a more fundamental rule than $\to$-E.

(a) it serves as premise-aggregator for the valid argument $\alpha \to \beta, \alpha \vdash \beta$, so that we have the single-premise rule $(\alpha \to \beta) \odot \alpha \vdash \beta$ and $\mathsf{VMP}_{\odot}$, and
(b) it satisfies the left-to-right direction of Idempotence, $\vdash \alpha \to \alpha \odot \alpha$.

Yet we have now seen that one or the other of these conditions fails for each of our candidate connectives. [47]

At this point, a critic of VTP might object that the response just given is at best incomplete. We have shown that the argument from VTP to absurdity fails, in the absence of SContr, when either & or $\otimes$ is used to state the premise $\mathsf{VMP}_{\odot}$. Still, the critic insists, our task remains that of explaining why the argument fails when $\odot$ expresses our *ordinary notion of conjunction*. After all, ordinary conjunction appears to satisfy both conditions (a) and (b): both single-premise *Modus Ponens* and Idempotence. If we are to avoid absurdity in the presence of a naïve theory of truth, we have argued, at least one of these appearances must be mistaken. The challenge is to explain which.

Zardini (2011, 2013) argues that condition (a) clearly holds for our "informal notion of conjunction." Accordingly, he maintains that ordinary conjunction is best captured by the multiplicative connective $\otimes$ of an affine logic—where the presence of SWeak guarantees such ordinary features as Simplification. Yet, as he recognizes, someone else might argue that condition (b) clearly holds for ordinary conjunction. More generally, we would add, one might maintain that the usual lattice properties are essential to our ordinary conjunction $\wedge$, whence from $\alpha \vdash \beta$ and $\alpha \vdash \gamma$ it must follow that $\alpha \vdash (\beta \wedge \gamma)$, even in the case where $\alpha = \beta = \gamma$.

We don't propose to settle this dispute about our informal notion of conjunction, or examine whether there is a univocal such notion.[48] Instead, we will now explain how the dispute is affected by the availability of dual-bunching logics. The chief reason Zardini insists that ordinary conjunction meets condition (a) is that he takes conjunction to be an all-purpose premise-aggregating connective. As he writes, conjunction is the connective we use to make explicit "how premises are combined in a multi-premise argument" (Zardini 2013). In order for $\odot$ to be conjunction, he holds, it is non-negotiable that it satisfy the rule

$$(\odot\text{-}L)\frac{\Gamma, \alpha_1, \alpha_2 \vdash \beta}{\Gamma, \alpha_1 \odot \alpha_2 \vdash \beta}$$

In a multiset-based logic without SContr, we have seen, the additive connective & violates $\odot$-L. We have a counterexample in the failure of $\alpha \to \beta, \alpha \vdash \beta$ to yield $(\alpha \to \beta)\ \&\ \alpha \vdash \beta$. This is the chief reason why he concludes that $\otimes$ has a stronger claim than & to represent our informal notion of conjunction.[49]

---

[47] It makes no difference whether these connectives are those of a multiset-based or dual-branching logic. Nor, in the latter case, would it make a difference if we considered $\&_A$ of note 44 in place of &.
[48] For arguments to the contrary, see Paoli (2007), Mares and Paoli (2014).
[49] Hjortland (2012) has recently proposed using an affine logic with additive conjunction and disjunction in a revisionary approach to semantic paradox. We take no position here on whether the consideration just rehearsed poses a serious problem for that approach.

But once dual-bunching logics are an option, matters get more complicated. In such logics we have both &-L_{db} and $\otimes$-L_{db}. The additive & corresponds to one mode in which premises may be combined, marked by our colon, while the multiplicative $\otimes$ corresponds to another mode, marked by our comma (Read 1988). According to dual-bunching logics, & doesn't serve as premise-aggregating connective for *Modus Ponens*, since we don't have $\alpha \rightarrow \beta : \alpha \vdash \beta$. Yet & serves as premise-aggregating connective for other arguments, e.g. $\alpha : \beta \vee \gamma \vdash (\alpha \,\&\, \beta) \vee \gamma$. Hence it is no longer clear that Zardini's view, on which ordinary conjunction is multiplicative and obeys single-premise *Modus Ponens* but not Idempotence, holds an advantage over the alternative view on which ordinary conjunction is additive and satisfies Idempotence but not single-premise *Modus Ponens*. Giving up single-premise *Modus Ponens*, understood in terms of ordinary conjunction, needn't amount to giving up conjunction's role as a premise-aggregating connective in a natural deduction system. Of course, as we noted above, the philosophical significance of the twofold bunching of premises needs to be elucidated. But that is also the case for the simpler premise structure in multiset-based deduction systems.

In this section, we have shown that the standard argument from VTP to absurdity breaks down in substructural theories which do not validate SContr, and have explained how the details of *where* it breaks down depend on which connective of the contraction-free logic we use to represent the conjunction appealed to in the standard argument.

22.3.4 The Consistency Argument

Let us finally turn to the Consistency Argument, and the resulting challenge to VTP from Gödel's Second Incompleteness Theorem. There are two ways one might respond: argue that Gödel's limitative results don't obtain for theories of arithmetic based on contraction-free logics, or argue that the Consistency Argument fails for such logics. Since there are contraction-free theories of arithmetic for which the results hold, we won't rely exclusively on the former strategy.[50]

The Consistency Argument requires one to prove, within one's semantic theory T, the following induction step: if *all* conclusions of derivations of length $\leq n$ are true, then *all* conclusions of derivations of length $n + 1$ are true. To prove this, it suffices to prove, for each rule R, that

[50] Restall (1994, Chap. 11) shows that an arithmetic based on the dual-bunching contraction-free logic RWK (which he calls CK) is classical Peano arithmetic, but it isn't known whether RWK supports a nontrivial naïve semantic theory in which $Tr(\ulcorner\alpha\urcorner)$ is everywhere intersubstitutable with α (see Hjortland 2012).

(TP_R) If *all* the premises of an instance of R are true, then the corresponding instance of the conclusion will be true.[51]

Now consider a rule R such that the theory proves that R has precisely two premises. To establish TP_R we will then need to prove

($\mathsf{TP2}_R$) For all x, y, z such that x and y are the two premises of an instance of R and z its corresponding conclusion: if x is true *and* y is true, then z is true.

But how are we to understand the 'and' in $\mathsf{TP2}_R$?

If 'all' in TP_R is understood as the standard "lattice-theoretical" or additive quantifier (Paoli 2005), then $\mathsf{TP2}_R$ will only help establish TP_R provided 'and' is likewise construed as additive.[52] But when R is the two-premise *Modus Ponens*, we won't be able to prove TP_R on this construal. That is because we have already seen that we don't have any instance of $\vdash_T (\alpha \rightarrow \beta)\ \&\ \alpha \rightarrow \beta$. This should mean that we don't have any instance of $\vdash_T Tr(\ulcorner \alpha \rightarrow \beta \urcorner)\ \&\ Tr(\ulcorner \alpha \urcorner) \rightarrow Tr(\ulcorner \beta \urcorner)$ either, whence we can't prove the generalization TP_R. In fact, that is Field's own explanation of how the Consistency Argument breaks down for paracomplete and paraconsistent theories (Field 2008, pp. 377–378). Unlike Field, we don't attribute this breakdown to the argument's illicit appeal to VTP. In our view, rather, the breakdown of the Consistency Argument (on the standard interpretation of the quantifier) results from the argument's illicit use of & as premise-aggregator for the two-premise *Modus Ponens* rule.[53]

Perhaps, then, we could rescue the Consistency Argument by interpreting the 'all' in TP_R as some kind of multiplicative quantifier, one that stands to $\otimes$ the way the standard universal quantifier stands to &. Where R is *Modus Ponens*, we should indeed be able to prove $\mathsf{TP2}_R$ with 'and' interpreted as $\otimes$, since $\otimes$ does serve as premise-aggregator for *Modus Ponens*. If this is to help establish TP_R, however, we would need to know more about the envisioned multiplicative quantifier. Paoli (2005) and Mares and Paoli (2014) note that there is no accepted theory of how such a quantifier should behave. One option is presented by Zardini (2011) in the context of a multiset-based logic. But Zardini's multiplicative quantifier won't serve the purposes of anyone who wishes to use the Consistency Argument to criticize VTP.

---

[51] Here we are no longer thinking of natural deduction rules, but rather of the rules of a Hilbert system, rules for generating theorems.

[52] Here is a rough explanation. In the course of deriving TP_R in our object-language, we will need to establish, under the assumption that three arbitrary sentences (denoted by a_1, a_2 and b) are the respective premises and conclusion of an instance of R, the claim $\forall x(x = a_1 \vee x = a_2 \rightarrow Tr(x)) \vdash Tr(b)$. Assuming $\forall$ is lattice-theoretical, this claim will follow from $Tr(a_1)\ \&\ Tr(a_2) \vdash Tr(b)$, whereas it won't follow from $Tr(a_1) \otimes Tr(a_2) \vdash Tr(b)$. For we have $\forall x \phi(x) \vdash \phi(a_1)\ \&\ \phi(a_2)...\ \&\ ...\phi(a_n)$, but not $\forall x \phi(x) \vdash \phi(a_1) \otimes \phi(a_2)... \otimes ...\phi(a_n)$. See Běhounek et al. (2007).

[53] Field himself claims that $\mathsf{TP2}_R$ will "obviously" fail to establish TP_R when the former is understood using what is, in effect, multiplicative conjunction. See Field (2006, p. 597) and Field (2008, p. 379). In his discussion, $Tr(\ulcorner \alpha \rightarrow \beta \urcorner) \rightarrow (Tr(\ulcorner \alpha \urcorner) \rightarrow Tr(\ulcorner \beta \urcorner))$ takes the place of $Tr(\ulcorner \alpha \rightarrow \beta \urcorner) \otimes Tr(\ulcorner \alpha \urcorner) \rightarrow Tr(\ulcorner \beta \urcorner)$, which is equivalent to the former in the logics we are considering. See also Priest (2010).

For he characterizes the behavior of the multiplicative quantifier using an ω-rule as (right-)introduction rule. Hence, the semantic theory based on this logic won't be recursively axiomatisable, and won't satisfy the conditions for Gödel's theorem.

22.4 Concluding Remarks

In this paper, we've argued for two main claims. First, the v-Curry Paradox shows that SContr is in tension with natural principles governing some (intuitive enough) notions of validity. Hence, if, as we've assumed, the validity relation is transitive, revisionary theorists have strong reason to give up SContr. Second, the standard challenges to VTP presented in §1 all break down once SContr is dropped. Rejecting SContr opens up non-classical ways of aggregating together premises—ways which no longer underwrite the objections to VTP. To be sure, it may be argued instead that the notion of validity that is shown to be paradoxical by the v-Curry Paradox should be rejected as incoherent. Validity, one might think, is interpretational, or purely logical, validity: truth on all uniform interpretations of the non-logical vocabulary. This, however, does not seem in line with the seemingly compelling thought, championed by rcf theorists such as Field (2007, 2008) and Priest (2006a, 2006b), that logical validity is a *species* of a more general notion of validity. Alternatively, it may be contended that paradox-prone notions of validity must be refined, and made less naïve (McGee 1991). But this, too, we've argued, doesn't seem like a viable option for proponents of the revisionary approach to paradox, who rather recommend revising our theory of logic, while preserving the naïve semantic principles. If neither of these foregoing options is viable, then SContr must be restricted on pain of triviality, and we can continue to maintain that valid arguments preserve truth.

References

Anderson, A., & Belnap, N. (1975). *Entailment: The logic of relevance and necessity* (Vol. 1). Princeton: Princeton University Press.

Asenjo, F. G.. (1966). A calculus of antinomies. *Notre Dame Journal of Formal Logic, 16,* 103–105.

Asenjo, F., & Tamburino, J. (1975). Logic of antinomies. *Notre Dame Journal of Formal Logic, 16,* 17–44.

Avron, A. (1988). The semantics and proof theory of linear logic. *Theoretical Computer Science, 57,* 161–184.

Beall, J. (2007). Truth and paradox: A philosophical sketch. In D. Jacquette (ed.), *Philosophy of logic* (pp. 325–410). Amsterdam: Elsevier.

Beall, J. (2009) *Spandrels of truth*. Oxford: Oxford University Press.

Beall, J. (2011). Multiple-conclusion LP and default classicality. *Review of Symbolic Logic, 4*, 326–336.

Beall, J., & Murzi, J. (2013). Two flavors of Curry's paradox. *Journal of Philosophy, 110,* 143–165.

Běhounek, L., Cintula, P., & Horčík, R. (2007). Multiplicative quantifiers in fuzzy and substructural logics. Presented at Logic Colloquium 2007, Wrocław.

Belnap, N. (1982) Display logic. *Journal of Philosophical Logic, 11,* 375–417.

Belnap, N. (1993). Life in the undistributed middle. In P. Schroeder-Heister & K. Došen (eds), *Substructural logics* (pp. 31–41). Oxford: Oxford University Press.
Brady, R. (1983). The simple consistency of set theory based on the logic *CSQ*. *Notre Dame Journal of Formal Logic, 24,* 431–39.
Brady, R. (1989). The non-triviality of dialectical set theory. In Priest, G., R. Routley & J. Norman (eds), *Paraconsistent logic: Essays on the inconsistent* (pp. 437–71). Munich: Philosophia.
Brady, R. (2006). *Universal logic*. Stanford: CSLI Publications.
Clark, M. (2007). *Paradoxes from A to Z* (2nd ed.). London: Routledge.
Cobreros, P., Egré, P., Ripley, D., van Rooij, R. (2012). Tolerant, classical, strict. *Journal of Philosophical Logic, 41,* pp. 347–385.
Cook, R. (2012). The T-schema is not a logical truth. *Analysis, 72,* pp. 231–239.
Cook, R. (2014). There is no paradox of logical validity! *Logica Universalis, 8,* pp. 447–467.
Došen, K. (1993). A historical introduction to substructural logics. In P. Schroeder-Heister & K. Došen (eds), *Substructural logics* (pp. 1–30). Oxford: Oxford University Press.
Etchemendy, J. (1990). *The concept of logical consequence*. Cambridge (Mass.): Harvard University Press.
Field, H. (2003). A revenge-immune solution to the semantic paradoxes. *Journal of Philosophical Logic, 32,* 139–177.
Field, H. (2006). Truth and the unprovability of consistency. *Mind, 115*, 567–606.
Field, H. (2007). Solving the paradoxes, escaping revenge. In J. Beall (ed.), *Revenge of the liar* (pp. 53–144). Oxford: Oxford University Press.
Field, H. (2008). *Saving truth from paradox*. Oxford: Oxford University Press.
Field, H. (2009a). Pluralism in logic. *Review of Symbolic Logic, 2*(2), 342–359.
Field, H. (2009b). What is the normative role of logic? *Proceedings of the Aristotelian Society, 83,* 251–268.
Field, H. (2015). Validity: Model-theoretic, proof-theoretic and genuine. Forthcoming in C. Caret & O. Hjortland (eds), *Foundations of Logical Consequence.* Oxford: Oxford University Press.
Girard, J. Y. (1987). Linear logic. *Theoretical Computer Science, 50,* 1–102.
Harman, G. (1986). *Change in view: Principles of reasoning*. Cambridge (Mass.): MIT Press.
Harman, G. (2009). Field on the normative role of logic. *Proceedings of the Aristotelian Society, 109*, 333–335.
Hjortland, O. T. (2012). Truth, paracompleteness and substructural logic. Unpublished manuscript.
Horsten, L. (2009). Levity. Mind, *118*, 555–581.
Kaplan, D., & Montague, R. (1960). A paradox regained. *Notre Dame Journal of Formal Logic, 1,* 79–90.
Ketland, J. (2012). Validity as a primitive. *Analysis, 72*, 421–430.
Kripke, S. (1975). Outline of a theory of truth. *Journal of Philosophy, 72,* 690–716.
Mares, E., & Paoli, F. (2014). Logical consequence and the paradoxes. *Journal of Philosophical Logic, 43,* pp. 439–469.
Martin, R. L. (ed.) (1984). *Recent essays on truth and the liar paradox*. Oxford: Oxford University Press.
Martin, R. L., & Woodruff, P. W. (1975). On representing 'true-in-L' in L. *Philosophia, 5,* 217–221. Reprinted in Martin 1984.
McGee, V. (1991). *Truth, vagueness, and paradox*. Indianapolis: Hackett Publishing Company.
Meyer, R. K., & McRobbie, M. A. (1982a). Multisets and relevant implication I. *Australasian Journal of Philosophy, 60,* 107–139.
Meyer, R. K., & McRobbie, M. A. (1982b). Multisets and relevant implication II. *Australasian Journal of Philosophy, 60,* 265–181.
Meyer, R. K., Routley, R., & Dunn, J. M. (1979). Curry's paradox. *Analysis, 39,* 124–128.
Minc, G. (1976). Cut-elimination theorem in relevant logics. *Journal of Soviet Mathematics, 6,* 422–428.
Murzi, J. (2014). The inexpressibility of validity. *Analysis, 74,* 65–81.
Myhill, J. (1960). Some remarks on the notion of proof. *Journal of Philosophy, 57*, 461–471.

O'Hearn, P., & Pym, D. (1999). The logic of bunched implications. *Bulletin of Symbolic Logic, 5,* 215–244.
Paoli, F. (2002). *Substructural logics: A primer*. Dordrecht: Kluwer.
Paoli, F. (2005). The ambiguity of quantifiers. *Philosophical Studies, 124,* 313–330.
Paoli, F. (2007). Implicational paradoxes and the meaning of logical constants. *Australasian Journal of Philosophy, 85,* 533–579.
Priest, G. (1979). The logic of paradox. *Journal of Philosophical Logic, 8,* 219–241.
Priest, G. (1980). Sense, entailment, and *modus ponens*. *Journal of Philosophical Logic, 9,* 415–435.
Priest, G. (1991). Intensional paradoxes. *Notre Dame Journal of Formal Logic, 32,* 193–211.
Priest, G. (2006a). *Doubt truth to be a liar*. Oxford: Oxford University Press.
Priest, G. (2006b). *In contradiction*. Oxford: Oxford University Press. Expanded edition (first published 1987).
Priest, G. (2010). Hopes fade for saving truth. *Philosophy, 85,* 109–140. Critical notice of Field 2008.
Priest, G., & Routley, R. (1982). Lessons from Pseudo-Scotus. *Philosophical Studies, 42,* 189–199.
Read, S. (1988). *Relevant logic: A philosophical examination of inference*. London: Basil Blackwell.
Read, S. (2001). Self-reference and validity revisited. In M. Yrjönsuuri (ed.), *Medieval formal logic* (pp. 183–196). Dordrecht: Kluwer.
Read, S. (2010). Field's paradox and its medieval solution. *History and Philosophy of Logic, 31*, 161–176.
Restall, G. (1993). How to be *really* contraction free. *Studia Logica, 52,* 381–391.
Restall, G. (1994). *On logics without contraction*. PhD thesis, The University of Queensland.
Restall, G. (2000). *An introduction to substructural logics*. London: Routledge.
Ripley, D. (2013). Paradoxes and failures of cut. *Australasian Journal of Philosophy, 91,* pp. 139–164.
Shapiro, L. (2011). Deflating logical consequence. *The Philosophical Quarterly, 61,* 320–342.
Shapiro, L. (2015). Naive structure, contraction and paradox. *Topoi*, *34*, 75–87.
Slaney, J. (1990). A general logic. *The Australasian Journal of Philosophy, 68*, 74–88.
Troelstra, A. (1992). *Lectures on linear logic*. Stanford: CSLI.
Weir, A. (2005). Naïve truth and sophisticated logic. In J. Beall & B. Armour-Garb (eds), *Deflationism and paradox* (pp. 218–249). Oxford: Oxford University Press.
Whittle, B. (2004). Dialetheism, logical consequence and hierarchy. *Analysis, 64*, 318–326.
Zardini, E. (2011). Truth without contra(di)ction. *Review of Symbolic Logic, 4,* 498–535.
Zardini, E. (2013). Naive *modus ponens*. *Journal of Philosophical Logic, 42,* pp. 575–592.

Chapter 23
Getting One for Two, or the Contractors' Bad Deal. Towards a Unified Solution to the Semantic Paradoxes

Elia Zardini

Abstract The paper concerns transparent theories of truth, i.e. theories treating ' 'φ' is true' as fully intersubstitutable with φ, and examines what the prospects are of maintaining a suitably refined version of transparency in view of the problem posed by the semantic paradoxes. In particular, three kinds of transparent theories—theories denying the law of excluded middle, theories denying the law of non-contradiction and theories denying the metarule of contraction—are compared with respect to the two most prominent semantic paradoxes: the Liar and Curry's. It is argued that there are versions of the Liar paradox that do not rely on the law of excluded middle or the law of non-contradiction, and that such versions are blocked by the first two kinds

Earlier versions of the material in this paper have been presented in 2012 at the MCMP Conference on *Paradox and Logical Revision* (Ludwig Maximilian University), at the 5th SPFA Meeting in Braga (University of Minho), at the 10th SIFA Conference in Alghero (University of Sassari), at the LOGOS *Semantic Paradoxes and Vagueness* Seminar (University of Barcelona) and at an NIP workshop in honour of Crispin Wright (University of Aberdeen). I'd like to thank all these audiences for very stimulating comments and discussions. Special thanks go to Massimiliano Carrara, Roy Cook, Philip Ebert, Hartry Field, Branden Fitelson, Luke Fraser, Patrick Greenough, Michiel van Lambalgen, José Martínez, Sebastiano Moruzzi, Julien Murzi, Sergi Oms, Marco Panza, Bryan Pickel, Graham Priest, Stephen Read, David Ripley, Sven Rosenkranz, Gonçalo Santos, Ricardo Santos, Daniele Sgaravatti, Vladimir Stepanov, Alan Weir, Robert Williams, Timothy Williamson, Crispin Wright and an anonymous referee. I'm also very grateful to the editors Theodora Achourioti, Kentaro Fujimoto, Henri Galinon and José Martínez for inviting me to contribute to this volume and for their support and patience throughout the process. In writing the paper, I've benefitted, at different stages, from an AHRC Postdoctoral Research Fellowship and the FP7 Marie Curie Intra-European Fellowship 301493 with project on *A Non-Contractive Theory of Naive Semantic Properties: Logical Developments and Metaphysical Foundations* (NTNSP), as well as from partial funds from the project FFI2008-06153 of the Spanish Ministry of Science and Innovation on *Vagueness and Physics, Metaphysics, and Metametaphysics*, from the project FFI2011-25626 of the Spanish Ministry of Science and Innovation on *Reference, Self-Reference and Empirical Data*, from the project CONSOLIDER-INGENIO 2010 CSD2009-00056 of the Spanish Ministry of Science and Innovation on *Philosophy of Perspectival Thoughts and Facts* (PERSP) and from the FP7 Marie Curie Initial Training Network 238128 on *Perspectival Thoughts and Facts* (PETAF).

E. Zardini
LanCog, University of Lisbon, Lisbon, Portugal
e-mail: elia.zardini@campus.ul.pt

T. Achourioti et al. (Eds.), *Unifying the Philosophy of Truth,* Logic, Epistemology, and the Unity of Science 36, DOI 10.1007/978-94-017-9673-6_23

of theories only by (implausibly) severing important connections between logical consequence and negation. Similarly, it is argued that Curry's paradox does not rely on the law of excluded middle or the law of non-contradiction, and that it is blocked by the first two kinds of theories only by (implausibly) severing important connections between logical consequence and the conditional. All the paradoxes discussed are shown however to rely on the metarule of contraction, and so the third kind of theory is revealed to have the advantage of offering a unified solution to such paradoxes.

23.1 Three Kinds of Transparent Theories of Truth

Say that a theory of truth is *transparent* iff the theory treats ' 'φ' is true' as *fully intersubstitutable* with φ. Transparency is an appealing *formal principle* about truth, whose force should be equally recognised by very different theories about the *nature* of truth, such as, for example, a broadly *correspondentist* theory and a broadly *deflationist* theory. As for correspondentism, transparency is naturally understood as saying that φ's being true is in a very strong sense both necessary and sufficient for what φ says being the case, and that is in turn something strongly suggested by the claim that φ's being true consists in φ's *corresponding to the facts*. As for deflationism, something like transparency is directly required if the notion of truth is adequately to serve the *expressive needs* that, according to the deflationist, constitute its *raison d'être*. However, plausible as it may seem from a wide variety of perspectives, transparency is not without its problems. I've discussed some of these at great length in (Zardini 2008, pp. 545–561, 2012, pp. 260–266, 2013e), arguing that they do place substantial limits on transparency. Here, I'd like to focus on a different and, in some respect, even more fundamental problem for transparency—the *semantic paradoxes*—arguing in favour of a view according to which those paradoxes actually do not place any further limit on transparency.

Consider a standard first-order interpreted language $\mathcal{T}$ that is expressive enough as to contain, for every sentence φ of $\mathcal{T}$, a singular term $\ulcorner\varphi\urcorner$ referring (by some means or other) to φ. Suppose also that T is a predicate of $\mathcal{T}$ expressing the notion of truth. Suppose finally that, in $\mathcal{T}$, t expresses the strongest proposition that logic enjoins to accept (the conjunction of all logical truths) and f the weakest proposition that logic enjoins to reject[1] (the disjunction of all logical falsehoods). The semantic paradoxes historically emerge with the *Liar* paradox in some of its versions (see (Bocheński 1970, p. 131)), so it just seems fit to begin with that paradox. Let's examine first a particularly canonical version of the paradox, considering a sentence λ identical to $\neg T(\ulcorner\lambda\urcorner)$. We start with:

[1] Against an influential tendency in much of the recent literature, I should make clear that by 'reject' and its likes I naturally mean something simply along the lines of *refusal to accept* the relevant proposition (in particular, I mean something that is compatible with thinking that the relevant proposition might be true). I'll use 'deny' and its likes to mean *acceptance of the negation* of the relevant proposition.

$$\dfrac{\dfrac{\dfrac{}{T(\ulcorner\lambda\urcorner) \vdash T(\ulcorner\lambda\urcorner)}\,\text{reflexivity} \quad \dfrac{\dfrac{}{T(\ulcorner\lambda\urcorner) \vdash T(\ulcorner\lambda\urcorner)}\,\text{reflexivity}}{T(\ulcorner\lambda\urcorner) \vdash \neg T(\ulcorner\lambda\urcorner)}\,\text{transparency}}{T(\ulcorner\lambda\urcorner), T(\ulcorner\lambda\urcorner) \vdash T(\ulcorner\lambda\urcorner) \,\&\, \neg T(\ulcorner\lambda\urcorner)}\,\text{adjunction}}{T(\ulcorner\lambda\urcorner) \vdash T(\ulcorner\lambda\urcorner) \,\&\, \neg T(\ulcorner\lambda\urcorner)}\,\text{contraction} \qquad \dfrac{}{T(\ulcorner\lambda\urcorner) \,\&\, \neg T(\ulcorner\lambda\urcorner) \vdash \mathsf{f}}\,\text{law of non-contradiction}$$
$$T(\ulcorner\lambda\urcorner) \vdash \mathsf{f}\ \text{transitivity}$$

where $\vdash$ expresses the relation of *following from*[2] and, given the prominent feature of the theory I'd like to defend in this paper, premises as well as conclusions are combined into *multisets* (call this reasoning 'A_0'). We continue with:

$$\dfrac{\dfrac{\dfrac{}{\neg T(\ulcorner\lambda\urcorner) \vdash \neg T(\ulcorner\lambda\urcorner)}\,\text{reflexivity} \quad \dfrac{\dfrac{}{\neg T(\ulcorner\lambda\urcorner) \vdash \neg T(\ulcorner\lambda\urcorner)}\,\text{reflexivity}}{\neg T(\ulcorner\lambda\urcorner) \vdash T(\ulcorner\lambda\urcorner)}\,\text{transparency}}{\neg T(\ulcorner\lambda\urcorner), \neg T(\ulcorner\lambda\urcorner) \vdash T(\ulcorner\lambda\urcorner) \,\&\, \neg T(\ulcorner\lambda\urcorner)}\,\text{adjunction}}{\neg T(\ulcorner\lambda\urcorner) \vdash T(\ulcorner\lambda\urcorner) \,\&\, \neg T(\ulcorner\lambda\urcorner)}\,\text{contraction} \qquad \dfrac{}{T(\ulcorner\lambda\urcorner) \,\&\, \neg T(\ulcorner\lambda\urcorner) \vdash \mathsf{f}}\,\text{law of non-contradiction}$$
$$\neg T(\ulcorner\lambda\urcorner) \vdash \mathsf{f}\ \text{transitivity}$$

(call this reasoning 'A_1'). We close with:

$$\dfrac{\dfrac{}{\mathsf{t} \vdash T(\ulcorner\lambda\urcorner) \vee \neg T(\ulcorner\lambda\urcorner)}\,\text{law of excluded middle} \qquad \dfrac{\dfrac{}{T(\ulcorner\lambda\urcorner) \vdash \mathsf{f}}\,A_0 \quad \dfrac{}{\neg T(\ulcorner\lambda\urcorner) \vdash \mathsf{f}}\,A_1}{T(\ulcorner\lambda\urcorner) \vee \neg T(\ulcorner\lambda\urcorner) \vdash \mathsf{f}}\,\text{reasoning by cases}}{\mathsf{t} \vdash \mathsf{f}}\,\text{transitivity}$$

(call this reasoning 'A_2', and the whole argument 'paradox A'), where I assume throughout that $\mathsf{t} \vdash \mathsf{f}$ is a catastrophe that every theory needs to avoid.

Paradox A is valuable because it makes very clear what are the main moves that a transparent theorist can make, thus serving as a privileged point of entry to a categorisation of the theories I'll discuss in the remainder of this paper: theories that deny the *law of excluded middle* (henceforth, 'LEM'), theories that deny the *law of non-contradiction* (henceforth, 'LNC') and theories that deny the *metarule of contraction* (henceforth, 'contraction'). The focus on the first two is motivated by their intrinsic plausibility and salience in the contemporary debate on the semantic paradoxes; the focus on the third is motivated by the fact that I believe that approach to be the correct one and have elsewhere proposed and developed a specific non-contractive

[2] Throughout, I use 'follow from' and its relatives to denote the relation of logical consequence (broadly understood so as to encompass also the "logic of truth") while I use 'entail' and its relatives to denote the converse relation. I use 'equivalence' and its relatives to denote two-way entailment. Moreover, I use 'implication' and its relatives to denote the status of the semantic values of two sentences that is necessary and sufficient for the conditional from the one sentence to the other sentence to be true (just like falsehood is the semantic value of a sentence that is necessary and sufficient for the negation of the sentence to be true).

theory (henceforth, for reasons that will become obvious shortly, '**IKT**').[3,4] In the remainder of this paper, I'll undertake a *comparative analysis* of how these three kinds of transparent theories fare with respect to the two most prominent kinds of semantic paradoxes. I'll argue that, at least with respect to those paradoxes, **IKT** has the great advantage of offering a *unified solution*.

Theories denying LEM (henceforth, 'non-LEM theories') have a reasonable intrinsic plausibility when considering a semantic paradox like paradox A, and they have long occupied a leading role in the contemporary debate on the semantic paradoxes (see e.g. (Kripke 1975; Brady 2006; Field 2008)). Such theories fully accept A_0 and A_1: they thus accept that both $T(\ulcorner\lambda\urcorner)$ and its negation are logical falsehoods.[5] Apart from LEM, such theories also accept all the other steps of A_2: in particular, they accept that, if both disjuncts are logical falsehoods, then the disjunction itself is a logical falsehood. Therefore, they deny that the conclusion of the relevant instance of LEM ($T(\ulcorner\lambda\urcorner) \vee \neg T(\ulcorner\lambda\urcorner)$) is a logical truth, and in fact flat-out reject that sentence. They thus deny the unrestricted validity of LEM.[6]

---

[3] Paradox A also suggests other interesting options that unfortunately there won't be space to discuss in this paper. Let me however remark that merely denying reasoning by cases and accepting all the other principles employed in paradox A does not suffice to uphold transparency. By LEM, $\mathsf{t} \vdash T(\ulcorner\lambda\urcorner) \vee \neg T(\ulcorner\lambda\urcorner)$ holds, and so, by transparency, $\mathsf{t} \vdash T(\ulcorner\lambda\urcorner) \vee T(\ulcorner\lambda\urcorner)$ holds, and hence, by a version of contraction, $\mathsf{t} \vdash T(\ulcorner\lambda\urcorner)$ holds. By similar reasoning, $\mathsf{t} \vdash \neg T(\ulcorner\lambda\urcorner)$ holds, and so, by adjunction, $\mathsf{t} \vdash T(\ulcorner\lambda\urcorner) \,\&\, \neg T(\ulcorner\lambda\urcorner)$ holds. Since, by LNC, $T(\ulcorner\lambda\urcorner) \,\&\, \neg T(\ulcorner\lambda\urcorner) \vdash \mathsf{f}$ holds, it follows, by transitivity, that $\mathsf{t} \vdash \mathsf{f}$ holds. Denial of reasoning by cases is in fact one of the main features of *supervaluationist* and *revision* theories (see e.g. (McGee 1991) and (Gupta and Belnap 1993) respectively), which are *not* transparent theories. All the theories considered in this paper accept the *specific* version of reasoning by cases employed in paradox A, although I've argued in (Zardini 2011) that theories denying contraction should also deny a more *general*—and very frequently mentioned—version of that metarule (the issue will briefly crop up in fn 16 and at the end of Sect. 23.3).

[4] Transparent theories as a whole can in turn be seen as forming a *logical* and *philosophical natural kind* along the correlated dimensions of *how deeply one deviates from classical logic* and, consequently, of *how tightly one can connect truth with reality*. Theories that retain *full classical logic* must give up even the *equivalence* between $T(\ulcorner\varphi\urcorner)$ and φ, and so must allow for the possibility that truth and reality *straightforwardly* come apart. Theories that at least *are closed under classical laws/rules and structural* (but not necessarily operational) *metarules* must still give up the *intersubstitutability* of $T(\ulcorner\varphi\urcorner)$ with φ (i.e. transparency), and so must allow for the possibility that truth and reality come apart *at least in certain contexts* (for example, in the suppositional contexts created by antecedents of conditionals). Transparent theories are characteristic in that, forcing truth and reality to go together *in every context*, they require a *deeper deviation from classical logic* consisting in denying some of its laws/rules or structural metarules. Thanks to Gonçalo Santos for raising this issue.

[5] Throughout, I assume—plausibly enough in view of our definition of f—that [$\varphi \vdash \mathsf{f}$ holds iff φ itself is a logical falsehood] and—plausibly enough in view of our definition of t—that [$\mathsf{t} \vdash \varphi$ holds iff φ itself is a logical truth]. (Throughout, I use square brackets to disambiguate constituent structure in English.) I'll offer some justification for the first assumption in fn 19.

[6] As the text already suggests, what I typically have in mind when I talk about "denial of an instance of LEM" is denial because of rejection of its conclusion, rather than denial because the conclusion may fail to have properties over and above truth that are deemed necessary for being a logical truth.

Theories denying LNC (henceforth, 'non-LNC theories') have traditionally been thought to have less intrinsic plausibility even when considering a semantic paradox like paradox A, but they have withstood well the initial incredulous stares and thus have managed to occupy a primary role in the contemporary debate on the semantic paradoxes (see e.g. (Priest 2006; Beall 2009)). Such theories fully accept something even more general than the main idea behind A_2: they accept that, if both a sentence and its negation entail a sentence, the latter sentence is a logical truth. Apart from LNC, such theories also accept all the other steps of A_0 and A_1: in particular, they accept that both $T(\ulcorner\lambda\urcorner)$ and its negation entail a contradiction. Therefore, they deny that the premise of the relevant instance of LNC ($T(\ulcorner\lambda\urcorner) \,\&\, \neg T(\ulcorner\lambda\urcorner)$) is a logical falsehood, and in fact flat-out accept that sentence. They thus deny the unrestricted validity of LNC.

As for **IKT**, that non-contractive theory technically corresponds to the multiplicative fragment of sentential *affine* logic (i.e. linear logic plus monotonicity) and can very naturally be presented in sequent-calculus style.[7] The *background logic* is defined as the smallest logic containing as axiom the structural rule:

$$\frac{}{\varphi \vdash_{\mathrm{IKT}} \varphi}\mathrm{I}$$

and closed under the structural metarules:

$$\frac{\Gamma_0 \vdash_{\mathrm{IKT}} \Delta}{\Gamma_0, \Gamma_1 \vdash_{\mathrm{IKT}} \Delta}\mathrm{K}-\mathrm{L} \qquad \frac{\Gamma \vdash_{\mathrm{IKT}} \Delta_0}{\Gamma \vdash_{\mathrm{IKT}} \Delta_0, \Delta_1}\mathrm{K}-\mathrm{R}$$

$$\frac{\Gamma_0 \vdash_{\mathrm{IKT}} \Delta_0, \varphi \quad \Gamma_1, \varphi \vdash_{\mathrm{IKT}} \Delta_1}{\Gamma_0, \Gamma_1 \vdash_{\mathrm{IKT}} \Delta_0, \Delta_1}\mathrm{S}$$

and under the operational metarules:

$$\frac{\Gamma \vdash_{\mathrm{IKT}} \Delta, \varphi}{\Gamma, \neg\varphi \vdash_{\mathrm{IKT}} \Delta}\neg-\mathrm{L} \qquad \frac{\Gamma, \varphi \vdash_{\mathrm{IKT}} \Delta}{\Gamma \vdash_{\mathrm{IKT}} \Delta, \neg\varphi}\neg-\mathrm{R}$$

$$\frac{\Gamma, \varphi, \psi \vdash_{\mathrm{IKT}} \Delta}{\Gamma, \varphi \,\&\, \psi \vdash_{\mathrm{IKT}} \Delta}\&-\mathrm{L} \qquad \frac{\Gamma_0 \vdash_{\mathrm{IKT}} \Delta_0, \varphi \quad \Gamma_1 \vdash_{\mathrm{IKT}} \Delta_1, \psi}{\Gamma_0, \Gamma_1 \vdash_{\mathrm{IKT}} \Delta_0, \Delta_1, \varphi \,\&\, \psi}\&-\mathrm{R}$$

$$\frac{\Gamma_0, \varphi \vdash_{\mathrm{IKT}} \Delta_0 \quad \Gamma_1, \psi \vdash_{\mathrm{IKT}} \Delta_1}{\Gamma_0, \Gamma_1, \varphi \vee \psi \vdash_{\mathrm{IKT}} \Delta_0, \Delta_1}\vee-\mathrm{L} \qquad \frac{\Gamma \vdash_{\mathrm{IKT}} \Delta, \varphi, \psi}{\Gamma \vdash_{\mathrm{IKT}} \Delta, \varphi \vee \psi}\vee-\mathrm{R}$$

$$\frac{\Gamma_0 \vdash_{\mathrm{IKT}} \Delta_0, \varphi \quad \Gamma_1, \psi \vdash_{\mathrm{IKT}} \Delta_1}{\Gamma_0, \Gamma_1, \varphi \to \psi \vdash_{\mathrm{IKT}} \Delta_0, \Delta_1}\to-\mathrm{L} \qquad \frac{\Gamma, \varphi \vdash_{\mathrm{IKT}} \Delta, \psi}{\Gamma \vdash_{\mathrm{IKT}} \Delta, \varphi \to \psi}\to-\mathrm{R}$$

Analogous comments apply for my talk below of "denial of an instance of LNC" and "denial of an instance of contraction".

[7] Given that non-LEM and non-LNC theories are already well-known in the literature, a similar presentation of their details would be superfluous. Moreover, given that there are important disagreements of detail among different non-LEM theories and among different non-LNC theories, a single such presentation would be impossible. Finally, given that the arguments of this paper are robust with respect to possible disagreements of detail among different non-LEM theories and among different non-LNC theories, multiple such presentations would be irrelevant.

IKT can be extended to a *theory of truth* by adding the following metarules for the truth predicate:

$$\frac{\Gamma, \varphi \vdash_{\mathrm{IKT}} \Delta}{\Gamma, T(\ulcorner \varphi \urcorner) \vdash_{\mathrm{IKT}} \Delta} T-\mathrm{L} \qquad \frac{\Gamma \vdash_{\mathrm{IKT}} \Delta, \varphi}{\Gamma \vdash_{\mathrm{IKT}} \Delta, T(\ulcorner \varphi \urcorner)} T-\mathrm{R}$$

In (Zardini 2011), I investigate in some detail the background-logical and truth-theoretic strength of what is essentially **IKT** (while proving its consistency), especially in the surprisingly many respects of philosophically interesting strength in which it outperforms non-LEM and non-LNC theories (and other theories too; indeed, I do all this for my favoured **IKT**'s extension to the first order). On the background-logical side, it's easy to see that **IKT** validates LEM, LNC, certain specific versions of the metarules of *reductio* (see Sect. 23.3) and reasoning by cases (see fn 3), the De Morgan equivalences, double-negation introduction and elimination (so that disjunction is definable by conjunction and negation, and conjunction definable by disjunction and negation), the rules of simplification and adjunction for conjunction and of addition and—so I like to call it—"abjunction" (the dual of adjunction) for disjunction,[8] the rule of *modus ponens* and the deduction theorem for the conditional (plus other principles that are characteristic of the material conditional, like the so-called 'paradoxes of material implication', so that the conditional is in effect definable by conjunction and negation or by disjunction and negation). Moreover, it's also easy to see that contraction can be locally recovered for a set of sentences X by adding, for every $\varphi \in X$, $\varphi \rightarrow \varphi \,\&\, \varphi$[9] as a further axiom to the theory. On the truth-theoretic side, it's easy to see that **IKT** validates the equivalence between $T(\ulcorner \varphi \urcorner)$ and φ, the Tarskian biconditionals, transparency, the standard truth-functional laws and the traditional constraint of truth preservation on logical

[8] The metarules given for & and $\vee$ are essentially the metarules for the "multiplicative" operators tensor and par of linear logics and the "intensional" operators fusion and fission of relevant logics. In both these kinds of logics, such operators are opposed to the "additive" or "extensional" operators, which are typically supposed to express *our informal notions of conjunction and disjunction* (as they occur for example in informal presentations of the semantic paradoxes). Against a likely misunderstanding, I can't emphasise enough that, with **IKT**'s & and $\vee$, I intend to give a *theory of precisely our informal notions of conjunction and disjunction* (hence my use of the standard symbolism), and that I take this interpretation to be warranted by the fact the last four rules just mentioned in the text are valid: together, those rules amount to saying that a conjunction is true iff both of its conjuncts are true, and that a disjunction is true iff either of its disjuncts is, which I take to capture the core of our informal notions of conjunction and disjunction. The divergence of my interpretation with the interpretation typically given for linear and relevant logics is explainable by the fact that those logics lack K-L and K-R, which determines that at least some of the rules just mentioned in the text are not valid for their "multiplicative" or "intensional" operators (hence these logics' search for other operators that can represent more adequately our informal notions of conjunction and disjunction).

[9] To save on brackets, I'll assume the usual scope hierarchy among the operators (with $\neg$ binding more strongly than & and $\vee$, and with these in turn binding more strongly than $\rightarrow$) and right associativity for each 2ary operator (so that $\varphi_0 \star \varphi_1 \star \varphi_2 \ldots \star \varphi_i$ reads $\varphi_0 \star (\varphi_1 \star (\varphi_2 \ldots \star \varphi_i)))\ldots$), with $\star$ being a 2ary operator).

consequence. I refer the reader to (Zardini 2011) for a sustained philosophical and technical study of **IKT** (see also (Zardini 2013a, 2013b, 2013c, 2013d) for studies of various issues of detail), and in this paper focus instead on another important aspect in which I think **IKT** enjoys a great advantage over non-LEM and non-LNC theories—that is, in promising a unified solution to the semantic paradoxes.

IKT fully accepts the main idea behind A_2: it thus accepts that, if both a sentence and its negation were logical falsehoods, a catastrophe would ensue. Apart from contraction, **IKT** also accepts all the other steps of A_0 and A_1: in particular, it accepts that both $T(\ulcorner\lambda\urcorner)$, $T(\ulcorner\lambda\urcorner)$ and $\neg T(\ulcorner\lambda\urcorner)$, $\neg T(\ulcorner\lambda\urcorner)$ entail a contradiction and so that they entail f. Therefore, it denies that $T(\ulcorner\lambda\urcorner)$ entailing f ($\neg T(\ulcorner\lambda\urcorner)$ entailing f) can be validly inferred from $T(\ulcorner\lambda\urcorner)$, $T(\ulcorner\lambda\urcorner)$ entailing f (from $\neg T(\ulcorner\lambda\urcorner)$, $\neg T(\ulcorner\lambda\urcorner)$ entailing f), and in fact flat-out denies the former entailments and flat-out accepts the latter entailments. It thus denies the unrestricted validity of contraction.

23.2 Some Historical Background on Non-Contractive Approaches to Paradox

Before examining how the three kinds of theories presented in Sect. 23.1 deal with another version of the Liar paradox and with the other most prominent kind of semantic paradox, I'd like to provide some brief historical background on the most significant historical antecedents of the non-contractive solution to the semantic paradoxes that I'm defending. I actually don't know of anywhere in the literature where restriction of contraction is the key component in a *philosophically motivated approach specifically focussed on the semantic paradoxes*. There are though a couple of *logical* and *computer-science* (rather than *philosophical*) traditions that have worked on the technical details of certain non-contractive logics, and have typically applied these to the *set-theoretic* (rather than *semantic*) paradoxes.[10]

The most relevant tradition is probably that represented by BCK *set theories* (whose study has been initiated by (Grišin 1974)). Very regrettably, it would appear that that paper is only in Russian with no translation available into English, with even the Russian version not being easily accessible. Because of these circumstances, a more informed study of the technical points of similarity and dissimilarity between the two theories will have to wait. However, so much can already be said. Grišin's theory includes both *multiplicative* connectives and *additive* ones (see fn 8), without specifying which (if any) are supposed to express our informal notions of conjunction and disjunction (this is probably due to the more general lack of philosophical underpinnings of the paper; it is precisely the focus on the multiplicative operators that gives **IKT** much of its logical strength—for example, the validity of LEM and LNC—and so much of its philosophical interest). Relatedly, touching on an issue

[10] A fascinating earlier reference envisaging failure of contraction (but not in relation to the paradoxes) may be (Pastore 1936), a work that however requires further historical and exegetical investigation. Thanks to Luke Fraser for alerting me to the existence of Pastore's work.

that I'll have to leave in the background in this paper, Grišin's theory only includes additive *quantifiers* (as defined for example by the standard metarules adopted in the sequent calculus for classical logic and many other logics), which, as I've argued in (Zardini 2011, pp. 510–511), yield an objectionable theory of quantification.[11]

As I've said, both antecedent non-contractive traditions tend to be concerned with sets rather than truth, and Grišin's theory is no exception (see however (Stepanov 2007) for a purely technical, non-philosophical non-contractive approach, which unfortunately includes no theory of conjunction, disjunction or quantification). That tendency is something this paper aims to correct: a non-contractive logic fits transparent truth in a surprisingly *optimal* fashion, getting us what we want and getting us rid of what we don't want (as I'll argue for important aspects in this paper and as I've argued for other aspects in my works referenced in Sect. 23.1), and promises a unified solution to its paradoxes (as I'll argue in this paper). Such features do not seem to be replicated in the case of the application of a non-contractive logic to the naive theory of sets, where restriction of contraction does not seem to yield a satisfactory theory. Grišin's own theory serves as an exemplification of the problem, since, in merely restricting contraction, while it manages to uphold the principle of full *comprehension* it does not seem to manage to retain consistency with the principle of full *extensionality* (as is realised in (Grišin 1981)), which is arguably essential to sets.[12]

Similar points apply to the couple of theorists that have followed Grišin on all these questionable choices ((White 1987, 1993; Petersen 2000, 2002), all of which actually develop theories that are even weaker than Grišin's in some philosophically crucial respect). That said, I should like to add that, contrary to all other authors mentioned in this section, Petersen's theory has rich philosophical foundations that I hope in future work to be able to discuss and compare with the picture I'll sketch at the end of Sect. 23.4.

The second, more loosely connected tradition I'd like to mention is that represented by *linear logics* (whose study has been initiated by (Girard 1987)). These logics have all the drawbacks I've mentioned about the background logic of BCK set theories. Additionally, they are substantially weaker in that they do not validate the metarule of *monotonicity*, which, as I've mentioned in fn 8, plays a crucial role in securing the strength of my theory. Much of the work done in linear logics is actually focussed on certain 1ary modal connectives, called '*exponentials*', which, while unexceptionable (and indeed very interesting and useful) from a purely logical and technical point of view, have no space in my theory, as I hope to be able to discuss in future work. Some authors have studied whether and to what extent linear logics can be used as background logics for set theories based on the principle of full

[11] Although I can't go into the details in this paper, the latter difference leads to Grišin's theory remaining *finitary*, while mine goes *infinitary*, since it introduces quantifiers by means of metarules akin to the ω-rule; the latter difference also leads to major divergences in the *deductive systems* codifying the respective theories and in the consistency proofs based on those (see (Zardini 2011, pp. 511–512, 524–532)).

[12] Ultimately, I don't think that Grišin's result forecloses a non-contractive naive theory of sets, but the issue lies beyond the scope of this paper.

comprehension (see e.g. (Girard 1998)). However, exactly because of the presence of exponentials, while shedding light on important *computational issues* the results have been rather disappointing for those in seek of a viable *theory of sets*.

23.3 A Unified Solution to the Liar Paradox

The Liar paradox can be introduced in an interestingly different—and equally common—way. We start with:

$$\dfrac{\dfrac{}{T(\ulcorner\lambda\urcorner) \vdash \mathsf{f}}\,\mathrm{A}_0}{\mathsf{t} \vdash \neg T(\ulcorner\lambda\urcorner)}\,\text{single-premise reduction theorem}$$

(call this reasoning 'B_0'). We continue with:

$$\dfrac{\dfrac{\dfrac{\dfrac{}{\mathsf{t} \vdash \neg T(\ulcorner\lambda\urcorner)}\,\mathrm{B}_0}{\mathsf{t} \vdash T(\ulcorner\lambda\urcorner)}\,\text{transparency} \qquad \dfrac{}{\mathsf{t} \vdash \neg T(\ulcorner\lambda\urcorner)}\,\mathrm{B}_0}{\mathsf{t}, \mathsf{t} \vdash T(\ulcorner\lambda\urcorner) \,\&\, \neg T(\ulcorner\lambda\urcorner)}\,\text{adjunction}}{\mathsf{t} \vdash T(\ulcorner\lambda\urcorner) \,\&\, \neg T(\ulcorner\lambda\urcorner)}\,\text{contraction}$$

(call this reasoning 'B_1').[13] We close with:

$$\dfrac{\dfrac{}{\mathsf{t} \vdash T(\ulcorner\lambda\urcorner) \,\&\, \neg T(\ulcorner\lambda\urcorner)}\,\mathrm{B}_1 \qquad \dfrac{}{T(\ulcorner\lambda\urcorner) \,\&\, \neg T(\ulcorner\lambda\urcorner) \vdash \mathsf{f}}\,\text{LNC}}{\mathsf{t} \vdash \mathsf{f}}\,\text{transitivity}$$

(call this reasoning 'B_2', and the whole argument 'paradox B').[14]

A major feature of paradox B is that it does not employ LEM. This is not to say that paradox B has typically been taken to tell against non-LEM theories, for these theories, while accepting all the other principles employed in paradox B, would typically deny the metarule of the *single-premise reduction theorem* (if $\varphi \vdash \mathsf{f}$ holds, $\mathsf{t} \vdash \neg\varphi$ holds). A discussion of the single-premise reduction theorem (and of its relationship to *reductio*) is thus in order. Consider first *reductio*. This is well-known to come at least in two versions: a *distinctively classical* version (if $\Gamma, \neg\varphi \vdash \Delta, \varphi$

[13] Contraction on t is admissible even in **IKT** and so I won't henceforth bother to make it explicit.

[14] While paradox A corresponds to the kind of informal presentation of the Liar paradox that proves that the Liar sentence cannot be true and proves that the Liar sentence cannot be untrue (while observing that it must be either true or untrue), paradox B corresponds to the kind of informal presentation of the Liar paradox that first proves that the Liar sentence is untrue and then on that basis proves that it is also true (while observing that it cannot be both true and untrue).

holds, $\Gamma \vdash \Delta, \varphi$ holds) and an *intuitionistically acceptable* version (if $\Gamma, \varphi \vdash \Delta, \neg\varphi$ holds, $\Gamma \vdash \Delta, \neg\varphi$ holds).[15] Now, any transparent theory generates so much *impredicativity* as to yield *ungrounded sentences that are strongly equivalent with their own negation*—sentences that in some sense consist in their self-negation (λ being a standard example of this). It is for this reason that, in the theoretic context of transparent theories, classical and intuitionist *reductio* immediately look as rather silly principles: in that context, it becomes one of the most distorting features of classical and intuitionist logic that *they rule out the very existence of such sentences*, and, at least barring the possibility of true contradictions, classical or intuitionist *reductio* are already sufficient to yield that most distorting result. More generally, in the theoretic context of transparent theories, even allowing for the possibility of true contradictions classical or intuitionist *reductio* still yield that such sentences can only exist if some contradictions are true—a result which is, without further justification, very dubious at best. Indeed, in that context, even allowing for the possibility of true contradictions the inference from the *intensional* fact that an ungrounded sentence is strongly equivalent with its own negation to the *categorical* fact that that sentence (and its negation, and the negation of its negation etc.) holds comes across, without further justification, as an egregious *non sequitur*: in that context, it seems perfectly coherent to accept the intensional fact—which simply determines what the ungrounded content of the sentence is—while rejecting the categorical facts—which would determine whether such content holds.[16,17]

---

[15] Having officially recorded the versions with side-premises and side-conclusions, for simplicity I'll ignore these in my treatment of *reductio*.

[16] Even adding the further disjunctive assumption that either the ungrounded sentence or its negation holds (an instance of LEM) does not substantially improve the appeal of the inference in the theoretic context of transparent theories. For, in that context, to suppose that the sentence holds is to suppose something intuitively equivalent with its failing to hold (since the sentence actually denies of itself that it holds), and so the fact that the suppositions represented by the two disjuncts can both be so developed as to reach the supposition that the sentence holds is after all not great evidence in favour of the sentence holding (since the supposition so reached is an unstable one). Thanks to Sven Rosenkranz for pressing me on this issue.

[17] Classical *reductio* is traditionally known as '*consequentia mirabilis*' (or 'Clavius' law', from the (Latin) name of the Counter-Reformation Jesuit priest who, actually in the steps of Girolamo Cardano, did much for promoting this method of proof in mathematics). The two names betray two completely different understandings of the principle. The *modern* name signals an understanding of the principle under which it *implicitly involves deriving the contradiction* $\varphi \,\&\, \neg\varphi$ from $\neg\varphi$, and from this fact inferring φ (presumably with the implicit thought that the first derivation shows that $\neg\varphi$ is false, which, in a bivalent spirit, is then taken to suffice for φ's being true). The traditional name signals an understanding of the principle under which it only involves deriving φ from $\neg\varphi$, and from this fact inferring φ (presumably with the implicit thought that the first derivation shows that, even if $\neg\varphi$ is true, φ still is, which, in a bivalent spirit, is then taken to suffice for φ's being true), *without any derivation of a contradiction justifying the reasoning* (one can see the two understandings nicely opposed in an epistolary debate between Christiaan Huygens and André Tacquet; see (Nuchelmans 1992) for a historical reconstruction of the dispute). Neither understanding helps in mitigating the objections against the principle levelled in the text. I'll discuss how both understandings covertly assume contraction at the end of this section.

Let's now move on to consider the reduction theorem. Notice first that non-LEM and non-LNC theories unfortunately turn out to be banned from accepting close kins of the single-premise reduction theorem (and turn out to be so banned independently of the semantic paradoxes). As for non-LEM theories, by reflexivity $\varphi \vdash \varphi$ holds, and so, by a particularly uncontroversial application of monotonicity, $\varphi \vdash \varphi, \mathsf{f}$ holds, and hence, by the *multiple-conclusion reduction theorem* (if $\varphi \vdash \Delta, \mathsf{f}$ holds, $\mathsf{t} \vdash \Delta, \neg\varphi$ holds), $\mathsf{t} \vdash \varphi, \neg\varphi$ holds. By disjunction in the conclusions, LEM follows. Since non-LEM theories typically accept all the other principles employed in this reasoning, they must deny the multiple-conclusion reduction theorem. A dual point can be made for non-LNC theories. By reflexivity, $\varphi \vdash \varphi$ holds, and so, by a particularly uncontroversial application of monotonicity, $\varphi, \mathsf{t} \vdash \varphi$ holds, and hence, by the *multiple-premise demonstration theorem* (if $\Gamma, \mathsf{t} \vdash \varphi$ holds, $\Gamma, \neg\varphi \vdash \mathsf{f}$ holds), $\varphi, \neg\varphi \vdash \mathsf{f}$ holds. By conjunction in the premises, LNC follows. Since non-LNC theories typically accept all the other principles employed in this reasoning, they must deny the multiple-premise demonstration theorem.

Let's grant for the sake of argument that non-LEM and non-LNC theories can make good sense of their denial of close kins of the single-premise reduction theorem; after all, as I've already mentioned, the problems raised in the previous paragraph do not rely on any truth-theoretic principle, and so they are much more general problems affecting virtually all theories that deny LEM or LNC (such as intuitionist logic and dual intuitionist logic respectively). The problem becomes much more specific with the single-premise reduction theorem. Contrary to classical and intuitionist *reductio* and to the multiple-conclusion reduction theorem, the input of the single-premise reduction theorem is not that a sentence entails its own contradictory or that it entails f as a side-conclusion—it is the intuitively much stronger one that a sentence entails f, and so that it is a logical falsehood.[18] Even on a transparent theory, the single-premise reduction theorem would seem irresistible: if a sentence is a logical falsehood, certainly it is false as a matter of logic and so its negation is true as a matter of logic and hence a logical truth—in other words, just like we have the principle of *truth functionality* that [the negation of a sentence is true if the sentence is false], we also have the principle of *logical-truth functionality* that [the negation of a sentence is logically true if the sentence is logically false]. The bad news for non-LEM theories is that, compelling as these considerations may seem, they ought to be rejected by a non-LEM theory on pain of paradox B.

There is an apparently terminological issue with the way the argument in the previous paragraph in favour of the single-premise reduction theorem has been presented which may be worth addressing. The argument plausibly assumes (as per fn 5) that *sentences entailing* f are *logical falsehoods*. The apparently terminological issue

[18] Because of this, in (Zardini 2011, p. 514, fn 38) I've argued also for the terminological point that the traditional name '*reductio ad absurdum*' for what I'm in this paper calling '*reductio*' is an egregious misnomer. For a metarule properly called '*reductio ad absurdum*' should have an input saying that certain premises lead to an absurdity, which is clearly the case for the single-premise reduction theorem and clearly not the case for either classical or intuitionist *reductio*, which would more properly be called '*reductio ad ipsius contradictoriam*'.

concerns the likely protest of a non-LEM theorist to the effect that the epithet 'logical falsehood' should be reserved for *sentences whose negation is a logical truth*. In fact, while the two properties are co-extensional in classical logic and in **IKT**, they come apart in both non-LEM theories (in particular, in these theories, a sentence can entail f while its negation is not a logical truth) and non-LNC theories (in particular, in these theories, the negation of a sentence can be a logical truth while the sentence does not entail f). Let's grant (only for the rest of this paragraph) this apparently terminological point about 'logical falsehood' and let's label instead with 'inconsistent' sentences entailing f.[19] That does not improve much the situation for non-LEM

[19] As might have been intimated by the qualification 'apparently', I don't think that the issue is purely terminological. I think that we can quite convincingly argue in a variety of ways that logical falsehood just is inconsistency. A first argument for that conclusion mimics the standard argument against the definition of falsehood of a sentence in terms of truth of its negation, which consists in the simple observation that that definition gets things dramatically wrong *for sentences belonging to languages which don't have negation*. In a completely analogous fashion, we can argue against the definition of logical falsehood of a sentence in terms of logical truth of its negation, by simply observing that that definition gets things dramatically wrong *again for sentences belonging to languages which don't have negation*. And, if the logical falsehood of a sentence is not to be identified with the logical truth of its negation, it's hard to see *what else* is left for it to be other than the inconsistency of the sentence. A second argument for the conclusion starts from the observation that inconsistent sentences behave dually with respect to logical truths, in the sense that, just like logical truths correspond to valid *no-premise, single-conclusion* arguments, inconsistent sentences correspond to valid *single-premise, no-conclusion* arguments. Assuming very plausibly that *logical falsehood behaves dually with respect to logical truth*, that forces the identification of logical falsehood with inconsistency. Or, similarly, just as the *best judgement that logic can pass on a sentence in itself* (i.e. not *qua* component of more complex sentences) is that it is the conclusion of a valid no-premise, single-conclusion argument, so the *worst judgement that logic can pass on a sentence in itself* is that it is the premise of a valid single-premise, no-conclusion argument. Assuming very plausibly that, just as the best judgement that logic can pass on a sentence in itself is equal to a judgement of logical truth, so *the worst judgement that logic can pass on a sentence in itself is equal to a judgement of logical falsehood*, that again forces the identification of logical falsehood with inconsistency. A third argument for the conclusion (and indeed, as we'll see, for something even stronger) assumes very plausibly that just as *(logical) truth is closed under entailment* (at least in the sense that, if φ is a (logical) truth and φ entails ψ, it follows that ψ is a (logical) truth), so *(logical) falsehood is closed under logical consequence* (at least in the sense that, if ψ is a (logical) falsehood and ψ follows from φ, it follows that φ is a (logical) falsehood). The argument then bifurcates. As for non-LEM theories, we only need to establish that inconsistency implies logical falsehood. We do so by observing that, at least in the systems of interest for this paper, being inconsistent implies *entailing sentences that are (logical) falsehoods by anyone's lights* (for example, 'For every P, P'), from which the desired implication follows by closure of (logical) falsehood under logical consequence (and, as I'll observe in the text, the weaker result just implicitly established that inconsistent sentences are merely false is already incompatible with non-LEM theories). As for non-LNC theories, we only need to establish that logical falsehood implies inconsistency. We do so by reducing to absurdity the claim that (it doesn't because) some logical falsehoods are also logical truths. In turn, we do so by observing that, at least in the systems of interest for this paper, being a logical truth implies *being entailed by sentences that are not (logical) falsehoods by anyone's lights* (for example, 'For some P, P'), from which the desired absurdity follows by closure of (logical) falsehood under logical consequence (and, as I'll observe in fn 33, the stronger result just implicitly established that logical truths lack mere falsehood is also incompatible with non-LNC theories). (As a matter of fact, non-LEM and non-LNC theorists

theories, for the argument in the previous paragraph in favour of the single-premise reduction theorem can be recast by saying that, if a sentence is inconsistent, certainly it cannot$_L$[20] be true and so must$_L$ be untrue and hence, by the full power of transparency, its negation must$_L$ be true and thus is a logical truth—the force of the argument remains strong and independent of our getting to attach the label 'logical falsehood' to inconsistent sentences. I suppose that the non-LEM theorist will have to deny the inference from 'φ is inconsistent' to 'φ cannot$_L$ be true' just as she had to deny the inference from 'φ is inconsistent' to 'φ is a logical falsehood'.[21] The non-LEM theorist has thus to deny that *even the worst judgement* that logic can pass on a sentence in itself suffices for the necessary$_L$ untruth of that sentence—indeed, it's easy to see that the non-LEM theorist even has to deny that it suffices for the mere untruth of that sentence (in this sense, although logic would still be powerful enough to establish many truths, somehow it would no longer be powerful enough to establish any untruths). Such a *lack of connection* between *logical consequence* on the one hand and *negation* and *logical modality* (and even *mere* negation) on the other hand strikes me as a great cost of these theories.[22]

---

are wont to reject closure of falsehood under logical consequence. The high implausibility of rejecting such a fundamental principle connecting logical consequence with falsehood is typically masked by recommending an apparently similar principle of closure of *rejectability* under logical consequence (if ψ ought to be rejected and ψ follows from φ, it follows that φ ought to be rejected). The recommendation is both disappointing and wrong. The recommendation is disappointing as it offers a *superficial, merely normative Ersatz* talking about *what people ought to do* in substitution for a *deep, fully descriptive* principle connecting logical consequence with *semantics* and so with *the ways the objective world can be* independently of people and of what they ought to do (in fact, even if it were true, the normative principle would cry out for a deeper explanation appealing, among other things, to something along the lines of the descriptive principle). The recommendation is also wrong as virtually every *counterexample to closure of knowledge and justification* can be turned into a counterexample to closure of acceptability and rejectability—those normative notions, as opposed to the semantic notions of truth and falsehood, are not suitable for the formulations of appropriate closure principles.)

[20] Throughout, modals and their likes subscripted with 'L' express logical modality.

[21] Might she not deny instead the *duality* of necessity$_L$ and possibility$_L$? In our dialectical context, the move would be extremely problematic in several respects. Firstly, short of transparency failing in contexts of logical modality, the move in effect now accepts the metarule from $\varphi \vdash \mathsf{f}$ to $\mathsf{t} \vdash$ 'It is not possible$_L$ that φ'. If in addition we still have the extremely plausible rule φ, 'It is not possible$_L$ that φ' $\vdash \mathsf{f}$, the resulting non-LEM theory is trivial (see the final version of paradox B mentioned in Sect. 23.4). Secondly, the duality of necessity$_L$ and possibility$_L$ is entailed by the duality of universal quantification and particular quantification plus the standard modality-worlds-linking principles 'It is necessary$_L$ that φ iff, for every possible$_L$ world w, 'φ' is true at w' and 'It is possible$_L$ that φ iff, for some possible$_L$ world w, 'φ' is true at w' (together with an appropriate version of transparency for 'true at w' like 'In w, 'φ' is true at w iff φ' and with the auxiliary assumption ' 'φ' is true at w_0 iff, for every possible$_L$ world w_1, in w_1, 'φ' is true at w_0'). Thirdly, the contrapositive of 'If φ, then it is possible$_L$ that φ' suffices to license the inference from ' 'φ' cannot$_L$ be true' to ' 'φ' is not true', which, as I'll observe in the text, is already incompatible with non-LEM theories. (See also the similar argument at the end of fn 33 that does not employ the duality of necessity$_L$ and possibility$_L$.) Thanks to Robert Williams for urging me to consider these issues.

[22] I should note that the argument in the text does not fall afoul of the worry I've raised above regarding the inference from intensional facts to categorical ones. For that worry concerned the

Obviously, the foregoing dialectic concerning the single-premise reduction theorem could profitably be continued. Let's stop here however since, for the purposes of this paper, the point is not so much to argue against the non-LEM theorist's denial of the single-premise reduction theorem, but only to establish the weaker point that the non-LEM theorist's solution to paradox B consisting in the denial of the single-premise reduction theorem *does not flow* from her solution to paradox A consisting in the denial of LEM. The latter point can easily be established by noticing that denial of LEM is quite compatible with acceptance of the single-premise reduction theorem: *intuitionist* logic provides a prime example of a coherent logical system lacking the former but having the latter. Thus, whatever rationale the non-LEM theorist might eventually come up with in order to justify her denial of the single-premise reduction theorem, it cannot merely consist in her denial of LEM—in fact, our discussion of the single-premise reduction theorem already amply shows that such a rationale would have to appeal to some rather unobvious considerations concerning the *lack of connection* between *logical consequence* on the one hand and *negation* and the accompanying notions of *(necessary$_L$) untruth* and *(logical) falsehood* on the other hand that are quite foreign to the issue as to whether LEM is valid.

At this point, the non-LEM theorist might grant that denial of LEM does not offer a unified solution to the semantic paradoxes, but suggest that her unified solution comes rather from whatever turns out to be the fundamental thought behind her solution to paradox B consisting in the denial of the single-premise reduction theorem. Let's assume, plausibly, that such fundamental thought is the idea that *some sentences are such that both they and their negation are logical falsehoods*. Let's set aside that, when so bluntly presented, the new view enjoys considerably less intrinsic plausibility than the original view that LEM is not valid. If that fundamental thought is correct, it provides a rationale for denying the single-premise reduction theorem (and thus solving paradox B): if both φ and $\neg\varphi$ are logical falsehoods, the single-premise reduction theorem implies the multiply repugnant conclusion that $\neg\varphi$ and $\neg\neg\varphi$ are logical truths. The fundamental thought also provides a rationale for denying LEM (and thus solving paradox A): if both φ and $\neg\varphi$ are logical falsehoods, reasoning by cases $\varphi \vee \neg\varphi$ is also a logical falsehood, and so presumably not a logical truth.

Have we thus hit on a more appropriate formulation of non-LEM theories, one that allows for a unified solution to the semantic paradoxes? Emphatically no. Firstly, and less conclusively, the new fundamental thought is *not at all distinctive* of non-LEM theories: for example, *supervaluationist* and *revision* theories that do accept LEM and deny transparency (see fn 3) would agree that some sentences are such that both they and their negation are logical falsehoods.[23] On the conception in question,

specific case in which the intensional fact is that an ungrounded sentence is strongly equivalent with its own negation, while the different specific case in which the intensional fact is that a sentence entails f is a case in which the intensional fact itself already intuitively involves, if not even is constituted by, the categorical fact that the sentence is a logical falsehood (or by the categorical fact that the sentence is inconsistent). Thanks to Sven Rosenkranz and an anonymous referee for raising this issue.

[23] Independently of the issue of which formulation of non-LEM theories is most appropriate, the observation in the text shows that a non-LEM theory becomes *completely indistinguishable* from

the *distinctiveness* of non-LEM theories against all the other main alternative theories would not consist in what would be the *fundamental thought* affording them a unified solution to the semantic paradoxes (i.e. the thought that some sentences are such that both they and their negation are logical falsehoods); it would rather consist in (that thought plus) an *ancillary thought* about disjunction (i.e. the thought that LEM fails for all sentences such that both they and their negation are logical falsehoods). The two thoughts need not of course be completely unrelated—I've mentioned in the previous paragraph how the fundamental thought can be used to yield the ancillary one given plausible additional assumptions about disjunction.[24] But the sore point remains that, on the conception in question, the fundamental thought affording to a non-LEM theory a unified solution to the semantic paradoxes is divorced from the ancillary thought characterising it against the all the other main alternative theories. Notice that, in general, the fact that a theory shares its fundamental thought with a theory differing from it in some other respects is of course just to be expected. What was not expected was that non-LEM theories and, say, supervaluationist theories—one of their traditional arch-rivals—would stand in such a relationship of being merely different variations on exactly the same theme, and so that non-LEM theories' take on the essence of the semantic paradoxes would be something wholeheartedly endorsed by some of the other main alternative theories.[25] Moreover, as we'll appreciate over the next few paragraphs, this problem has the tendency to degenerate badly—especially with respect to the connection between the fundamental thought and the ancillary thought—once it interacts with another somewhat converse problem to which we now turn.[26]

Secondly, and more conclusively, not every semantic paradox has truck with even apparent logical falsehoods: the *truth-teller* paradox ('This sentence is true', see (Mortensen and Priest 1981)), the *no-no* paradox ('The next sentence is not true',

supervaluationist and revision theories in a language lacking disjunction (and the resources to define it) but expressive enough as to generate semantic paradoxes (the language needed to generate the essence of paradox B is a good example of such a language).

[24] Such assumptions start to look less plausible *vis-à-vis* the alternative assumptions made by supervaluationist and revision theories once it is realised that, keeping fixed the duality of conjunction and disjunction and the idea that some sentences are such that both they and their negation are logical falsehoods, the latter assumptions allow supervaluationist and revision theories to uphold the very plausible claim that *the negation of any contradiction is a logical truth*, while the former assumptions force non-LEM theories to endorse the very implausible claim that *the negations of certain contradictions are logical falsehoods*.

[25] Contrast with **IKT**, whose fundamental thought that some sentences are such that they fail to contract is strong enough to characterise it against all the other main alternative theories. Of course, **IKT** too has its own variations. In (Zardini 2013a), I study in some detail a particularly natural one that replaces the *multiplicative* operators with the *additive* ones, and argue that such variation suffers from a *lack of connection* between *logical consequence* on the one hand and *conjunction* and *disjunction* on the other hand similar to the *lack of connection* between *logical consequence* on the one hand and *negation* on the other hand suffered by non-LEM theories.

[26] Thanks to Branden Fitelson, David Ripley, Sven Rosenkranz and an anonymous referee for criticisms of an earlier draft of this paragraph.

'The previous sentence is not true', see (Sorensen 2001, pp. 165–170)), certain versions of *Epimenides'* paradox ('This sentence is not true and the number of stars in the universe is not even', see (Goldstein 1986)) are paradigmatic examples of *semantic paradoxes without even apparent logical falsehoods*.[27] The natural solution given to these paradoxes by non-LEM theories consists in rejecting the relevant excluded middles[28] (and hence denying LEM), rather than in claiming that the relevant sentences are such that both they and their negation are logical falsehoods.[29]

In response, one could, I suppose, stretch the notion of "logic" at play in a non-LEM theory so that every sentence rejected by the theory counts as a "logical falsehood". But such a proposal would be unsatisfactory for various related reasons. Firstly, in the new stretched sense of 'logical falsehood', the apparently *objective* claim that some sentences are such that both they and their negation are "logical falsehoods" boils down to the *autobiographical* claim that some sentences are such that both they and their negation are rejected by the theory (the theorist really), which can hardly be taken to be a fundamental thought behind a solution to the semantic paradoxes. In fact, such a claim immediately cries out for a deeper explanation: why does the theory reject both a sentence and its negation? Secondly, the proposal does nothing but *exacerbate* the problem discussed two paragraphs back of distinguishing the fundamental thought of non-LEM theories from what the other main alternative theories endorse: *virtually all theories*—including my favoured interpretation of **IKT**—that accept the equivalence between $T(\ulcorner\varphi\urcorner)$ and φ but do not accept both a sentence and its negation agree in rejecting the relevant paradoxical sentences and their negation. Thirdly, the ancillary thought about disjunction concerning the failure of LEM *does no longer even remotely follow from the fundamental thought as so weakly understood*: usually, rejecting both disjuncts is no sufficient ground for rejecting a disjunction (as evidenced by ordinary cases of *agnosticism*).

One may try to address all these points by modifying slightly the proposal in question, saying that the fact that the theory rejects a sentence is actually a reflection of a more fundamental fact to the effect that the sentence is false *or somehow indeterminate*. A first problem with this modification is that, pending an *explanation* of the notion of indeterminacy at play, it is far from clear that it can effectively address the last two points of the previous paragraph. A second problem with the modification is that the relevant notion of indeterminacy supposed to apply to all paradoxical

[27] Obvious as this point may seem, its import has frequently been overlooked in certain debates involving the semantic paradoxes (see López de Sa and Zardini 2006, 2007, 2011).

[28] Since we've reserved 'law of excluded middle' (i.e. 'LEM') for the logical claim expressed by '$\mathfrak{t} \vdash \varphi \vee \neg\varphi$', let's use 'excluded middles' to refer to the typically non-logical claims expressed by instances of $\varphi \vee \neg\varphi$.

[29] I hasten to add that I don't mean to imply that a theory *must* solve all the paradoxes just mentioned in the text by deploying its fundamental thought, since, at least for the first two kinds of paradoxes, a theory might also appeal to *independently plausible truth-theoretic principles*, and, at least in the case of the truth-teller paradox, such appeal might *suffice to yield already a solution to the paradox* (for example, one might argue that the truth-teller simply lacks truth on the strength of general considerations concerning truth and grounding in reality, see (Priest 2006, p. 66)). Thanks to Patrick Greenough for urging this clarification.

sentences is then typically explained in non-LEM theories *in terms of* LEM *failing for those sentences*, so that the alleged fundamental thought would finally collapse on the thought that LEM fails, which we've already seen does not in itself offer the materials for a solution to paradox B. A third problem with the modification is that it actually *still does not apply to all paradoxical sentences*: for example, the Epimenides sentence 'This sentence is not true and the number of stars in the universe is not even' is extremely plausibly false if the number of stars in the universe is even, and so there is no warrant to regard it as indeterminate in any reasonable sense. Concerning such a sentence, a non-LEM theorist could and should say that she rejects it and its negation because, *given what she knows*, it *might* be indeterminate. But, since, given what she knows, that sentence also might be not indeterminate, to say so is in effect to concede that *there are paradoxical sentences which might be not indeterminate*. And, since either our original Epimenides sentence or the opposite Epimenides sentence 'This sentence is not true and the number of stars in the universe is even' is false (for either the number of stars in the universe is even or it is not), to say so is in effect to concede that *there are paradoxical sentences which are not indeterminate*.[30] Moreover, rejection on such *epistemic* grounds no longer serves the purpose of addressing any of the points made in the previous paragraph which the modification discussed in this paragraph was supposed to address.[31]

[30] I've argued that there are paradoxical sentences which are not indeterminate, and that is enough to undermine the modification discussed in the text. It then becomes a secondary question whether the non-LEM theorist should say that LEM *fails* for whichever is the false sentence in the pair of opposite Epimenides sentences considered in the text. For what it's worth, to me it would be very plausible to say that it does, and so concede that, in some cases, LEM fails even if a sentence is not indeterminate. But I suppose that one could instead say that it does not. To me, that would be very implausible. Obviously, since the sentence is false, the relevant excluded middle is true. But, since the sentence might have been indeterminate and cannot be known *a priori* not to be such (for the number of stars in the universe might have had the other parity and cannot be known *a priori* not to have it), to conclude from the truth of the relevant excluded middle that the relevant instance of LEM is valid is in effect to concede that, in some cases, *a sentence is a logical truth even if it is not necessarily such and even if it cannot be known a priori*, as well as to concede that, at least given transparency and in the broad sense explained in fn 2, *there is a logical proof* of the parity of the number of stars in the universe—in fact, a logical proof *of every actual truth*. Notice that exactly the same points apply to an Epimenides sentence *which we do know to be false*, such as 'This sentence is not true and there are no stars' (given which it would then become extremely natural to give the same treatment also to sentences which are necessarily false and known *a priori* (but not logically) to be false, such as 'This sentence is not true and something is not part of itself'). Given this, the non-LEM theorist would have to concede for the Epimenides sentences she knows to be false (not only that they are nevertheless paradoxical but also) that *LEM nevertheless fails for them* (as she would have to for other paradoxical sentences), even if she believes them to be false, and so *even if she does not reject their negation*. Thanks to Timothy Williamson for discussion of this question.

[31] In view of this dialectic, one may try to take a rather different, more concrete approach focussing for example on a specific *model-theoretic construction* and saying that it is from the fundamental thought behind the construction that *both* denial of LEM *and* denial of the single-premise reduction theorem flow (although it should be mentioned that a non-LEM theorist in the spirit of (Field 2008) would be reluctant to assign such a *fundamental explanatory role* to the model theory). A natural candidate for such a proposal is the strong-Kleene construction of (Kripke 1975). There is no doubt

I've focussed so far on how well non-LEM theories fare with respect to paradox B. Non-LNC theories would block the paradox as presented already at subargument A_0, thus offering a unified solution to both paradox A and paradox B. If we try instead to run a *modified* version of paradox B using subargument A_0 only to get $T(\ulcorner\lambda\urcorner) \vdash \neg T(\ulcorner\lambda\urcorner)$, and then employ *reductio* to infer $\mathsf{t} \vdash \neg T(\ulcorner\lambda\urcorner)$ from that, we do obtain an argument that is valid by the lights of non-LNC theories, although, as I've noted at the beginning of this section, in the theoretic context of transparent theories the inference in question is, without further justification, very dubious at best (so much seems to be acknowledged by non-LNC theorists, since they typically feel the need to justify *reductio* and the immediately following claim that there are true contradictions by appealing to LEM along the lines of the argument I'll present at the end of this section; see e.g. (Priest 2006, pp. 12–16, 64–66)). But non-LNC theories would only accept the modified version of paradox B up to subargument B_1, and would deny the final subargument B_2 *qua* employing LNC, thus offering a unified solution to both paradox A and paradox B (original or modified).

However, we can also run a *dual modified* version of paradox B. In particular, given $\mathsf{t} \vdash \neg T(\ulcorner\lambda\urcorner)$, we can employ the metarule of the *single-conclusion demonstration theorem* (if $\mathsf{t} \vdash \varphi$ holds, $\neg\varphi \vdash \mathsf{f}$ holds) and infer $T(\ulcorner\lambda\urcorner) \vdash \mathsf{f}$ (typically, non-LNC

---

of course about the *technical* fact that the overall construction invalidates both LEM and the single-premise reduction theorem. What is open to doubt, however, is whether there is *a single fundamental thought* behind the overall construction. Let me explain. The *basic thought* behind the construction seems to be the "gaps-and-grounds" picture supporting the package constituted by the strong-Kleene valuation scheme together with the definition of truth at a later stage in terms of how things are at the earlier stage(s). Such basic thought thus includes the thought that *sentences might have a gappy status* (not to be identified with lack of truth and falsehood) which is a fixed point for negation and that is inherited by a disjunction from both of its disjuncts, a thought which, given plausible additional assumptions, is indeed sufficient to invalidate LEM. But such basic thought remains silent about the single-premise reduction theorem. That issue is only addressed by the *additional theoretic decision* of defining (single-premise, single-conclusion) logical consequence as *downwards preservation of the truth-entailing non-gappy status* (at the relevant fixed point(s)). But, precisely in the context of the basic thought in which *downwards preservation of the truth-entailing non-gappy status* does no longer coincide with *upwards preservation of the untruth-entailing non-gappy* status, that decision is arbitrary and indeed questionable in that it gives more importance to the truth-entailing non-gappy status than to the untruth-entailing one. The arbitrariness and indeed questionability of the decision may not be immediately apparent because one can still infer from the *truth of the premise* of a "valid" argument the *truth of the conclusion*, but it does emerge once it is noticed that one can no longer infer from the *untruth of the conclusion* of a "valid" argument the *untruth of the premise* (see also the principle of closure of (logical) falsehood under logical consequence discussed at the end of fn 19). Once the alternative but more natural definition requiring *both* downwards preservation of the truth-entailing non-gappy status *and* upwards preservation of the untruth-entailing non-gappy status is adopted, the single-premise reduction theorem is validated. Moreover, the basic thought by itself already seems strongly to suggest the single-premise reduction theorem, and seems in any case to have consequences incompatible with non-LEM theories. For that thought also involves the thought that *a sentence is true in virtue of its positive grounding in reality*, from which it seems to follow that sentences that are not so grounded are not true (since they lack that in virtue of which a sentence is true). But, in the Kripke construction, if a sentence is a logical falsehood, it is not positively grounded in reality, and so it is not true. This strongly suggests the single-premise reduction theorem, and is in any case incompatible with non-LEM theories. Thanks to José Martínez, Sebastiano Moruzzi and Bryan Pickel for urging me to consider this alternative proposal.

theories, along with virtually all other theories, treat $\neg\neg\varphi$ as fully intersubstitutable with φ). Given that, by transparency, $\mathsf{t} \vdash \neg T(\ulcorner\lambda\urcorner)$ also implies $\mathsf{t} \vdash T(\ulcorner\lambda\urcorner)$, by transitivity we get $\mathsf{t} \vdash \mathsf{f}$.

A major feature of the dual modified version of paradox B is that it does not employ LNC. This is not to say that the dual modified version of paradox B has typically been taken to tell against non-LNC theories, for these theories, while accepting all the other principles employed in the dual modified version of paradox B, would deny the metarule of the single-conclusion demonstration theorem.

However, all the reasons given above in favour of the single-premise reduction theorem have dual reasons speaking in favour of the single-conclusion demonstration theorem (just like the single-premise reduction theorem infers logical truths from logical falsehoods, the single-conclusion demonstration theorem infers logical falsehoods from logical truths). If a sentence is a logical truth, it is true as a matter of logic and so its negation is false as a matter of logic and hence certainly a logical falsehood and thus entails f—in other words, just like we have the principle of *truth functionality* that [the negation of a sentence is false if the sentence is true], we also have the principle of *logical-truth functionality* that [the negation of a sentence is logically false if the sentence is logically true].

Using (only for the rest of this paragraph) 'inconsistent' instead of 'logical falsehood' in the same sense of and for the reasons explained eleven paragraphs back, the argument can be recast by saying that, if a sentence is a logical truth, it must$_L$ be true and so cannot$_L$ be untrue and hence, by the full power of transparency, its negation cannot$_L$ be true and thus certainly is inconsistent—the force of the argument remains strong and independent of our getting to attach the label 'logical falsehood' only to inconsistent sentences. I suppose that the non-LNC theorist will have to deny the inference from 'φ cannot$_L$ be true' to 'φ is inconsistent' just as she had to deny the inference from 'φ is a logical falsehood' to 'φ is inconsistent'.[32] The non-LNC theorist has thus to deny that *even the worst judgement* that logical modality can pass on a sentence in itself suffices for the inconsistency of that sentence—indeed, it's easy to see that the non-LNC theorist even has to deny that it suffices for the mere lack of truth of that sentence (in this sense, although logical modality would still be powerful enough to establish many truths, somehow it would no longer be powerful enough to establish any lack of truth). Such a *lack of connection* between *logical consequence* (and even *mere lack of truth*) on the one hand and *negation* and *logical modality* on the other hand strikes me as a great cost of these theories.[33]

---

[32] *Mutatis mutandis*, all the three points made in fn 21 apply if she denies the duality of necessity$_L$ and possibility$_L$ (see also the similar argument in fn 33 that does not employ the duality of necessity$_L$ and possibility$_L$).

[33] Similarly, if a sentence is a logical truth, certainly it must$_L$ lack falsehood and so its negation must$_L$ lack truth and hence is inconsistent. I suppose that the non-LNC theorist will have to deny the inference from 'φ is a logical truth' to 'φ must$_L$ lack falsehood'. The non-LNC theorist has thus to deny that *even the best judgement* that logic can pass on a sentence in itself suffices for the necessary$_L$ lack of falsehood of that sentence—indeed, it's easy to see that the non-LNC theorist even has to deny that it suffices for the mere lack of falsehood of that sentence (in this sense,

Obviously, the foregoing dialectic concerning the single-conclusion demonstration theorem could profitably be continued. Let's stop here however since, for the purposes of this paper, the point is not so much to argue against the non-LNC theorist's denial of the single-conclusion demonstration theorem, but only to establish the weaker point that the non-LNC theorist's solution to paradox B consisting in the denial of the single-conclusion demonstration theorem *does not flow* from her solution to paradox A consisting in the denial of LNC. The latter point can easily be established by noticing that denial of LNC is quite compatible with acceptance of the single-conclusion demonstration theorem: *dual intuitionist* logic provides a prime example of a coherent logical system lacking the former but having the latter (see e.g. (Urbas 1996)). Thus, whatever rationale the non-LNC theorist might eventually come up with in order to justify her denial of the single-conclusion demonstration theorem, it cannot merely consist in her denial of LNC—in fact, our discussion of the single-conclusion demonstration theorem already amply shows that such a rationale would have to appeal to some rather unobvious considerations concerning the *lack of connection* between *logical consequence* on the one hand and *negation* and the accompanying notions of *(necessary$_L$) untruth* and *(logical) falsehood* on the other hand that are quite foreign to the issue as to whether LNC is valid.[34]

At this point, the non-LNC theorist might grant that denial of LNC does not offer a unified solution to the semantic paradoxes, but suggest that her unified solution comes rather from whatever turns out to be the fundamental thought behind her solution to the dual modified version of paradox B consisting in the denial of the single-conclusion demonstration theorem. Let's assume, plausibly, that such fundamental thought is the idea that *some sentences are such that both they and their*

although logic would still be powerful enough to establish many truths, somehow it would no longer be powerful enough to establish any lack of falsehood). Such a *lack of connection* between *logical consequence* on the one hand and *lack of falsehood* and *logical modality* (and even *mere* lack of falsehood) on the other hand strikes me as a great cost of these theories. (Notice that a dual argument is available against non-LEM theories.)

[34] Another aspect of this lack of connection (strictly related to the principle of closure of (logical) falsehood under logical consequence discussed at the end of fn 19) concerns the traditional idea that *logical consequence consists in the impossibility$_L$ that [the premise is true and the conclusion is not true]*. Non-LEM theories need to reject that a premise entails a conclusion *only if* it is impossible$_L$ that [the premise is true and the conclusion is not true], since they accept that λ entails f, and so would have to accept that it is impossible$_L$ (i.e. it is not possible$_L$) that [λ is true and f is not true]. But, by the duality of necessity$_L$ and possibility$_L$, that implies that it is necessary$_L$ that it is not the case that [λ is true and f is not true], and so, by the relevant De Morgan rule and closure of necessity$_L$ under entailment, it would be necessary$_L$ that either λ is not true or f is true, and hence, reasoning by cases, by the properties of f and closure of necessity$_L$ under entailment, it would be necessary$_L$ that λ is not true, which is however unacceptable for non-LEM theories. Non-LNC theories need to reject that a premise entails a conclusion *if* it is impossible$_L$ that [the premise is true and the conclusion is not true], since they accept that it is necessary$_L$ that λ is not true. By the contrapositive of simplification and closure of necessity$_L$ under entailment, that implies that it is necessary$_L$ that it is not the case that [λ is true and f is not true], and so, by the duality of necessity$_L$ and possibility$_L$, that it is not possible$_L$ (i.e. it is impossible$_L$) that [λ is true and f is not true], and hence non-LNC theories would have to accept that λ entails f, which is however unacceptable for them.

negation are logical truths. If that fundamental thought is correct, it provides a rationale for denying the single-conclusion demonstration theorem (and thus solving the dual modified version of paradox B): if both φ and $\neg\varphi$ are logical truths, the single-conclusion demonstration theorem implies the multiply repugnant conclusion that $\neg\varphi$ and $\neg\neg\varphi$ are logical falsehoods. The fundamental thought also provides a rationale for denying LNC (and thus solving paradox A): if both φ and $\neg\varphi$ are logical truths, by adjunction φ & $\neg\varphi$ is also a logical truth, and so presumably not a logical falsehood.

Have we thus hit on a more appropriate formulation of non-LNC theories, one that allows for a unified solution to the semantic paradoxes? Emphatically no. Firstly, and less conclusively, the new fundamental thought is *not at all distinctive* of non-LNC theories: for example, *subvaluationist* and *non-standard revision* theories that do accept LNC and deny transparency would agree that some sentences are such that both they and their negation are logical truths.[35,36] On the conception in question, the *distinctiveness* of non-LNC theories against all the other main alternative theories would not consist in what would be the *fundamental thought* affording them a unified solution to the semantic paradoxes (i.e. the thought that some sentences are such that both they and their negation are logical truths); it would rather consist in (that thought plus) an *ancillary thought* about conjunction (i.e. the thought that LNC fails for all sentences such that both they and their negation are logical truths). The two thoughts need not of course be completely unrelated—I've mentioned in the previous paragraph how the fundamental thought can be used to yield the ancillary one given plausible additional assumptions about conjunction.[37] But the sore point remains that, on the conception in question, the fundamental thought affording to a non-LNC theory a unified solution to the semantic paradoxes is divorced from the thought characterising it against all the other main alternative theories. Moreover, as we'll appreciate over the next few paragraphs, this problem has the tendency to degenerate badly—especially with respect to the connection between the fundamental thought

[35] To the best of my knowledge, such theories have not been investigated in relation to the semantic paradoxes. I won't go into their details in this paper—suffice it to say that they naturally arise by dualising, respectively, the familiar Kripke construction based on the supervaluationist evaluation scheme and the familiar revision-sequence construction. Such theories are very similar to one another in the respects that are relevant for our discussion: in particular, they accept both a sentence and its negation without accepting any contradiction (more strongly, they hold that the contradiction is a logical falsehood and accept LNC in its full generality). The underlying idea, to put it very roughly, is that conjunction is sensitive to *compatibility relationships* between the conjuncts.

[36] A comment analogous to that in fn 23 applies concerning the *complete indistinguishability* of all these theories in expressively impoverished paradoxical languages.

[37] Such assumptions start to look less plausible *vis-à-vis* the alternative assumptions made by subvaluationist and non-standard revision theories once it is realised that, keeping fixed the duality of conjunction and disjunction and the idea that some sentences are such that both they and their negation are logical truths, the latter assumptions allow subvaluationist and non-standard revision theories to uphold the very plausible claim that the negation of any excluded middle is a logical falsehood, while the former assumptions force non-LNC theories to endorse the very implausible claim that the negations of certain excluded middles are logical truths.

and the ancillary thought—once it interacts with another somewhat converse problem to which we now turn.

Secondly, and more conclusively, not every semantic paradox has truck with even apparent logical truths: the same examples discussed in connection with non-LEM theories are relevant here as well. The natural solution given to these paradoxes by non-LNC theories (with the exception of the relevant versions of Epimenides' paradox) consists in accepting the relevant contradictions (and hence denying LNC), rather than in claiming that the relevant sentences are such that both they and their negation are logical truths.

In response, one could, I suppose, stretch the notion of "logic" at play in a non-LNC theory so that every sentence accepted by the theory counts as a "logical truth". But such a proposal would be unsatisfactory for various related reasons. Firstly, in the new stretched sense of 'logical truth', the apparently *objective* claim that some sentences are such that both they and their negation are "logical truths" boils down to the *autobiographical* claim that some sentences are such that both they and their negation are accepted by the theory (the theorist really), which can hardly be taken to be a fundamental thought behind a solution to the semantic paradoxes. In fact, such a claim immediately cries out for a deeper explanation: why does the theory accept both a sentence and its negation? Secondly, the proposal does nothing but *exacerbate* the problem discussed two paragraphs back of distinguishing the fundamental thought of non-LNC theories from what the other main alternative theories endorse: *virtually all theories* that accept the equivalence between $T(\ulcorner\varphi\urcorner)$ and φ but do not reject both a sentence and its negation agree in accepting the relevant paradoxical sentences and their negation. Thirdly, the ancillary thought about conjunction concerning the failure of LNC *does no longer even remotely follow from the fundamental thought as so weakly understood*: sometimes, accepting both conjuncts is no sufficient ground for accepting a conjunction (as evidenced by the *preface paradox*, see (Makinson 1965)).

One may try to address all these points by modifying slightly the proposal in question, saying that the fact that the theory accepts a sentence is actually a reflection of a more fundamental fact to the effect that the sentence is true-only *or somehow overdeterminate*. A first problem with this modification is that, pending an *explanation* of the notion of overdeterminacy at play, it is far from clear that it can effectively address the last two points of the previous paragraph. A second problem with the modification is that the relevant notion of overdeterminacy supposed to apply to all paradoxical sentences is then typically explained in non-LNC theories *in terms of* LNC *failing for those sentences*, so that the alleged fundamental thought would finally collapse on the thought that LNC fails, which we've already seen does not in itself offer the materials for a solution to the dual modified version of paradox B. A third problem with the modification is that it actually *still does not apply to all paradoxical sentences*: for example, the Epimenides sentence 'This sentence is not true and the number of stars in the universe is not even' is extremely plausibly false-only if the number of stars in the universe is even, and so there is no warrant to regard it as overdeterminate in any reasonable sense. Rather implausibly, concerning such a sentence a non-LNC theorist could say that she *accepts* it because, *given what she*

knows, it *might* be overdeterminate. But, since, given what she knows, that sentence also might be not overdeterminate, to say so is in effect to concede that *there are paradoxical sentences which might be not overdeterminate*. And, since either our original Epimenides sentence or the opposite Epimenides sentence 'This sentence is not true and the number of stars in the universe is even' is false-only (for either the number of stars in the universe is even or it is not), to say so is in effect to concede that *there are paradoxical sentences which are not overdeterminate*.[38] Moreover, acceptance on such *epistemic* grounds no longer serves the purpose of addressing any of the points made in the previous paragraph which the modification discussed in this paragraph was supposed to address. More plausibly, concerning both such sentences a non-LNC theorist could say that she *rejects* them because, *given what she knows*, either *might* be false-only. But, if she does so, there would be paradoxical sentences that are not such that both they and their negation are accepted by the theory, contrary to what the proposal in question requires.[39]

In stark contrast with the severe difficulties for non-LEM and non-LNC theories in giving a unified solution to paradoxes A and B (original or modified or dual modified), **IKT** offers a smooth treatment of both: that theory denies subargument A_0 as it employs contraction, thus blocking paradox A and the original version of paradox B; moreover, the theory does accept subargument A_0 up until $T(\ulcorner\lambda\urcorner) \vdash \neg T(\ulcorner\lambda\urcorner)$, but denies *reductio*, thus blocking the modified and dual modified versions of paradox B. This is in the ball-park for being a unified solution because, contrary to the single-premise reduction theorem, I've argued above that, in the theoretic context of transparent theories, *reductio* is in the "better" case (i.e. if there are true contradictions), without further justification, very dubious at best while it is in the "worse" case (i.e. if there are no true contradictions) straightforwardly incompatible with the impredicative phenomena that go together with transparency, and there is no obligation for a unified transparent solution to deploy its fundamental thought in order to deny an argument that relies, without further justification, on such a problematic principle.[40]

---

[38] Comments analogous to those in fn 30 apply concerning the question whether the non-LNC theorist should say that LNC fails for whichever is the false-only sentence in the pair of opposite Epimenides sentences considered in the text.

[39] Comments analogous to those in fn 31 apply concerning a rather different, more concrete approach focussing for example on a specific *model-theoretic* construction and saying that it is from the fundamental thought behind the construction that *both* denial of LNC and denial of the single-conclusion demonstration theorem flow. In particular, the first comment in fn 31 has an analogue to the effect that the decision of defining (single-premise, single-conclusion) logical consequence as *downwards preservation of truth-entailing non-glutty or glutty status* (at the relevant fixed point(s)), without requiring *upwards preservation of untruth-entailing non-glutty or glutty status*, is arbitrary and indeed questionable. The second comment in fn 31 has an analogue to the effect that the basic thought behind the relevant construction involves the thought that *a sentence is untrue in virtue of its negative grounding in reality*, from which it seems to follow that sentences that are not so grounded lack untruth (since they lack that in virtue of which a sentence is untrue).

[40] Compare the similar argument consisting only in the inference from $T(\ulcorner\lambda\urcorner) \vdash \neg T(\ulcorner\lambda\urcorner)$ and $\neg T(\ulcorner\lambda\urcorner) \vdash T(\ulcorner\lambda\urcorner)$ to $\mathsf{t} \vdash$ 'Classical mathematics is inconsistent'. Although that argument is classically valid, there is no obligation for a unified transparent solution to deploy its fundamental thought in order to deny the only inference employed in the argument—in the theoretic context of

Still, albeit in these respects extremely problematic, *reductio* can actually be further justified by a couple of apparently compelling arguments, so that—supplementing the modified or dual modified version of paradox B with some such argument—we do have a genuine paradox to block. The first argument corresponds to the first understanding of *reductio* mentioned in fn 17, and is most naturally applied to its intuitionistically acceptable version. The argument assumes that a formula φ entails its own negation and reasons as follows:

$$\dfrac{\dfrac{\dfrac{\dfrac{}{\varphi \vdash \varphi}\,\text{reflexivity} \qquad \varphi \vdash \neg\varphi}{\varphi, \varphi \vdash \varphi \,\&\, \neg\varphi}\,\text{adjunction}}{\varphi \vdash \varphi \,\&\, \neg\varphi}\,\text{contraction} \qquad \dfrac{}{\varphi \,\&\, \neg\varphi \vdash \mathsf{f}}\,\text{LNC}}{\mathsf{t} \vdash \neg\varphi}\,\text{transitivity}}{\varphi \vdash \mathsf{f}}\,\text{single-premise reduction theorem}$$

As we've seen, non-LEM theories block this argument by denying the apparently compelling single-premise reduction theorem, getting involved in the dialectic examined above concerning the original version of paradox B; non-LNC theories, while not accepting this particular argument, accept *reductio* anyways (as I've already mentioned, typically feeling the need to justify it and the immediately following claim that there are true contradictions by appealing to LEM along the lines of the argument I'll present in the next paragraph), getting rather involved in the dialectic examined above concerning the dual modified version of paradox B. **IKT**, on the contrary, denies this argument as it employs contraction, thus blocking it (and the original version of paradox B) *in exactly the same way* as paradox A.

The second argument corresponds to the second understanding of *reductio* mentioned in fn 17, and is most naturally applied to its distinctively classical version. The argument assumes that a formula φ follows from its own negation and reasons as follows:

$$\dfrac{\dfrac{}{\mathsf{t} \vdash \varphi \vee \neg\varphi}\,\text{LEM} \qquad \dfrac{\dfrac{\dfrac{}{\varphi \vdash \varphi}\,\text{reflexivity} \qquad \neg\varphi \vdash \varphi}{\varphi \vee \neg\varphi \vdash \varphi, \varphi}\,\text{abjunction}}{\varphi \vee \neg\varphi \vdash \varphi}\,\text{contraction}}{\mathsf{t} \vdash \varphi}\,\text{transitivity}$$

transparent theories, that inference is, without further justification, very dubious at best. This is not to deny of course that the inference might be *further justified* by appeal to *more fundamental and, even in the theoretic context of transparent theories, apparently compelling principles*, so as to produce a *genuine paradox* for those theories that should ideally be blocked by deploying their fundamental thought. That is in fact what I've done for the single-premise reduction theorem and the single-conclusion demonstration theorem (with the main aim of showing that non-LEM and non-LNC theories need to appeal to some rather unobvious considerations concerning the lack of connection between logical consequence and negation that are quite foreign to the issue as to whether LEM or LNC are valid), and what I'll proceed to do for *reductio* (with the main aim of showing that the justifying arguments do involve contraction). Thanks to Sven Rosenkranz and an anonymous referee for criticisms of an earlier draft of this paragraph.

Non-LEM theories block this argument by denying LEM; non-LNC theories accept this argument and, as I've already mentioned, typically use it to justify *reductio* and the immediately following claim that there are true contradictions, getting involved in the dialectic examined above concerning the dual modified version of paradox B. **IKT**, on the contrary, denies this argument as it employs contraction, thus blocking it (and the original version of paradox B) *in exactly the same way* as paradox A.

Since **IKT** blocks all of paradox A, the original version, the most compelling modified versions and the most compelling dual modified versions of paradox B in the same way, it does offer a *unified* solution to these paradoxes—what non-LEM and non-LNC theories fail to do. Notice that the solution is also *not overdetermined*: all the other principles employed in paradox A and in the original version, the most compelling modified versions and the most compelling dual modified versions of paradox B are valid according to **IKT**—in particular, LEM, LNC, the single-premise reduction theorem and the single-conclusion demonstration theorem are all valid according to **IKT**.

23.4 A Unified Solution to the Liar and Curry's Paradox

Although the semantic paradoxes historically emerge with the Liar paradox in some of its versions, as a matter of autobiographical remark the original and constantly guiding source of inspiration for **IKT** has been a certain version of *Curry's* paradox ((Curry 1942) is the modern *locus classicus* while (Ashworth 1974, p. 125) mentions some prominent scholastic antecedents). So let's examine in some detail the workings of that version, considering a sentence κ identical to $T(\ulcorner\kappa\urcorner) \rightarrow \bot$ (where $\bot$ is the conjunction of all propositions). We start with:

$$\dfrac{\dfrac{\dfrac{}{T(\ulcorner\kappa\urcorner) \rightarrow \bot, T(\ulcorner\kappa\urcorner) \vdash \bot}\,\textit{modus ponens}}{T(\ulcorner\kappa\urcorner), T(\ulcorner\kappa\urcorner) \vdash \bot}\,\text{transparency}}{T(\ulcorner\kappa\urcorner) \vdash \bot}\,\text{contraction}$$

(call this reasoning 'C_0'). We continue with:

$$\dfrac{\dfrac{\dfrac{}{T(\ulcorner\kappa\urcorner) \vdash \bot}\,C_0}{\mathtt{t} \vdash T(\ulcorner\kappa\urcorner) \rightarrow \bot}\,\text{single-premise deduction theorem}}{\mathtt{t} \vdash T(\ulcorner\kappa\urcorner)}\,\text{transparency}$$

(call this reasoning 'C_1'). We close with:

$$\dfrac{\dfrac{}{\mathsf{t} \vdash T(\ulcorner \kappa \urcorner)}\,\mathrm{C}_1 \qquad \dfrac{}{T(\ulcorner \kappa \urcorner) \vdash \bot}\,\mathrm{C}_0}{\mathsf{t} \vdash \bot}\,\text{transitivity}$$

(call this reasoning 'C_2', and the whole argument 'paradox C').

Like paradox B, a major feature of paradox C is that it employs neither LEM nor LNC. This is not to say that paradox C has typically been taken to tell against non-LEM or non-LNC theories, for these theories, while accepting all the other principles employed in paradox C, would typically deny the metarule of the *single-premise deduction theorem* (if $\varphi \vdash \psi$ holds, $\mathsf{t} \vdash \varphi \to \psi$ holds). A discussion of the single-premise deduction theorem (and of its relationship to *absorption*) is thus in order. Consider first absorption. This comes at least in two versions (both of which are classically valid): a well-known *law* version ($\mathsf{t} \vdash (\varphi \to (\varphi \to \psi)) \to (\varphi \to \psi)$) and a less well-known *metarule* version (if $\Gamma, \varphi \vdash \Delta, \varphi \to \psi$ holds, $\Gamma \vdash \Delta, \varphi \to \psi$ holds).[41] Now, any transparent theory generates so much *impredicativity* as to yield *ungrounded sentences that are strongly equivalent with their own conditional to an unacceptable sentence*—sentences that in some sense consist in their self-conditional to an unacceptable sentence (κ being a standard example of this). It is for this reason that, in the theoretic context of transparent theories, the law and the metarule of absorption immediately look as rather silly principles: in that context, it becomes one of the most distorting features of classical logic that *it rules out the very existence of such sentences*, and, given closure of logical truth under *modus ponens*, the law or the metarule of absorption are already sufficient to yield that most distorting result.

Let's now move on to consider the deduction theorem. Consider first the *multiple-premise deduction theorem* (if $\Gamma, \varphi \vdash \psi$ holds, $\Gamma \vdash \varphi \to \psi$ holds). Now, any theory that both accepts *monotonicity* and aims at preserving *relevance* for $\to$ should deny the multiple-premise deduction theorem. For, by reflexivity, $\varphi \vdash \varphi$ holds, and so, by monotonicity, $\varphi, \psi \vdash \varphi$ holds, and hence, by the multiple-premise deduction theorem, $\varphi \vdash \psi \to \varphi$ holds, thus yielding the "positive paradox of material implication". Interesting as it may be in other contexts, this point against the multiple-premise deduction theorem is far from being conclusive in ours, as many non-LEM and non-LNC theories do not accept relevance constraints on $\to$. Nevertheless, non-LEM and non-LNC theories unfortunately turn out to be banned on other grounds from accepting close kins of the single-premise deduction theorem (and turn out to be so banned independently of the semantic paradoxes). As for non-LEM theories, by reflexivity $\varphi \vdash \varphi$ holds, and so, by monotonicity, $\varphi \vdash \varphi, \bot$ holds, and hence, by the *multiple-conclusion deduction theorem* (if $\varphi \vdash \Delta, \psi$ holds, $\mathsf{t} \vdash \Delta, \varphi \to \psi$ holds), $\mathsf{t} \vdash \varphi, \varphi \to \bot$ holds. Since in non-LEM theories $\varphi \to \bot \vdash \neg\varphi$ typically holds, by transitivity $\mathsf{t} \vdash \varphi, \neg\varphi$ holds, and so, by disjunction in the conclusions, LEM follows. Since non-LEM theories typically accept all the other principles employed in this

[41] Having officially recorded the version with side-premises and side-conclusions, for simplicity I'll ignore these in my treatment of the metarule of absorption.

reasoning, they must deny the multiple-conclusion deduction theorem. A dual point can be made for non-LNC theories by introducing the dual of the conditional, the *unconditional* $\dot{-}$ (read $\varphi \dot{-} \psi$ informally as something like 'Its being the case that φ does not require its being the case that ψ'). By reflexivity, $\varphi \vdash \varphi$ holds, and so, by monotonicity, $\varphi, \top \vdash \varphi$ holds (where $\top$ is the disjunction of all propositions), and hence, by the *multiple-premise deduction theorem for the unconditional* (if $\Gamma, \varphi \vdash \psi$ holds, $\Gamma, \varphi \dot{-} \psi \vdash \mathsf{f}$ holds), $\varphi, \top \dot{-} \varphi \vdash \mathsf{f}$ holds. Since in non-LNC theories $\neg\varphi \vdash \top \dot{-} \varphi$ should presumably hold, by transitivity $\varphi, \neg\varphi \vdash \mathsf{f}$ holds, and so, by conjunction in the premises, LNC follows. Since non-LNC theories typically accept or should presumably accept all the other principles employed in this reasoning, they must deny the multiple-premise deduction theorem for the unconditional.

Let's grant for the sake of argument that non-LEM and non-LNC theories can make good sense of their denial of close kins of the single-premise deduction theorem; after all, as I've already mentioned, the problems raised in the previous paragraph do not rely on any truth-theoretic principle, and so they are much more general problems affecting virtually all theories that deny LEM or LNC (such as intuitionist logic and dual intuitionist logic respectively). The problem becomes much more specific with the single-premise deduction theorem. Contrary to the law and the metarule of absorption and to the multiple-conclusion deduction theorem, the input of the single-premise deduction theorem is not that a sentence implies or entails its own conditional to a sentence or that it entails a sentence as a side-conclusion—it is the intuitively much stronger one that a sentence entails a sentence, and so that the former sentence logically implies the latter sentence. Even on a transparent theory, the single-premise deduction theorem would seem irresistible: if a sentence logically implies a sentence, certainly the former sentence implies the latter sentence as a matter of logic, and so the conditional is true as a matter of logic and hence a logical truth. The bad news for non-LEM and non-LNC theories is that, compelling as these considerations may seem, they ought to be rejected by a non-LEM or non-LNC theory on pain of paradox C.

There is an apparently terminological issue with the way the argument in the previous paragraph in favour of the single-premise deduction theorem has been presented which may be worth addressing. The argument plausibly assumes that *cases of entailment* are also *cases of logical implication*. The apparently terminological issue concerns the likely protest of a non-LEM or non-LNC theorist to the effect that the epithet 'logical implication' should be reserved for *cases of logically true conditionals*. In fact, while the two properties are co-extensional in classical logic and in **IKT**, they come apart in both non-LEM and non-LNC theories (in particular, in these theories, although every case of a logically true conditional is a case of entailment, some cases of entailment are not cases of logically true conditionals). Let's grant (only for the rest of this paragraph) this apparently terminological point about 'logical implication'.[42] That does not improve much the situation for non-LEM

[42] Comments analogous to the first two in fn 19 apply concerning the only apparent terminological character of the point. In particular, the second comment in fn 19 has an analogue to the effect that,

and non-LNC theories, for the argument in the previous paragraph in favour of the single-premise deduction theorem can be recast by saying that, if a sentence entails a sentence, certainly the latter sentence must$_L$ be true if the former sentence is true, and so, by the full power of transparency, the conditional from the former to the latter must$_L$ be true and thus is a logical truth—the force of the argument remains strong and independent of our getting to attach the label 'logical implication' to cases of entailment. I suppose that the non-LEM or non-LNC theorist will have to deny the inference from 'φ entails ψ' to 'ψ must$_L$ be true if φ is true' just as she had to deny the inference from 'φ entails ψ' to 'φ logically implies ψ'. The non-LEM or non-LNC theorist has thus to deny that *even the best judgement* that logic can pass on a sentence in itself with respect to a sentence in itself suffices for the necessary$_L$ truth preservation from the latter sentence to the former sentence—indeed, it's easy to see that the non-LEM or non-LNC theorist even has to deny that it suffices for the mere truth preservation from the latter to the former (in this sense, although logic would still be powerful enough to establish many truths, somehow it would no longer be powerful enough to establish any preservation of truth). Such a *lack of connection* between *logical consequence* on the one hand and the *conditional* and *logical modality* (and even the *mere* conditional) on the other hand strikes me as a great cost of these theories.

Obviously, the foregoing dialectic concerning the single-premise deduction theorem could profitably be continued. Let's stop here however since, for the purposes of this paper, the point is not so much to argue against the non-LEM or non-LNC theorist's denial of the single-premise deduction theorem, but only to establish the weaker point that the non-LEM or non-LNC theorist's solution to paradox C consisting in the denial of the single-premise deduction theorem *does not flow* from her solution to paradox A consisting in the denial of LEM or LNC. The latter point can easily be established by noticing that denial of LEM or LNC is quite compatible with acceptance of the single-premise deduction theorem: the latter principle—and paradox C more generally—concerns the conditional and does not concern negation at all, while the former principles concern negation and do not concern the conditional at all. Thus, whatever rationale the non-LEM or non-LNC theorist might eventually come up with in order to justify her denial of the single-premise deduction theorem, it cannot merely consist in her denial of LEM or LNC—in fact, our discussion of the single-premise deduction theorem already amply shows that such a rationale would

---

just as the best judgement that logic can pass on a sentence in itself—that it is the conclusion of a no-premise, single-conclusion argument—is equal to a judgement of validity for the categorical statement consisting in that sentence, so *the best judgement that logic can pass on a sentence in itself with respect to a sentence in itself*—i.e. that it is the conclusion of a single-premise, single-conclusion argument whose premise is the latter sentence—*should be equal to a judgement of validity for the hypothetical statement from the latter sentence to the former sentence*—i.e. to a logical implication from the latter to the former. A third argument for the conclusion that logical implication just is entailment assumes very plausibly that entailment, usually presented as a *metalinguistic* relation, is ultimately just a kind of *object-linguistic* implication (lying at one extreme of a spectrum at whose other extreme lies material implication), which should then be identified with logical implication.

have to appeal to some rather unobvious considerations concerning the *lack of connection* between *logical consequence* on the one hand and the *conditional* and the accompanying notions of *(necessary$_L$) truth preservation* and *(logical) implication* on the other hand that are quite foreign to the issue as to whether LEM or LNC is valid—and, more generally, quite foreign to the further issues involving negation discussed in Sect. 23.3.[43]

In fact, in a way sharper than the one exemplified by the Epimenides sentences discussed in Sect. 23.3 (especially those discussed at the end of fn 30), a Curry sentence like κ is an example of a full-blooded paradoxical sentence for which the non-LNC theorist *cannot possibly accept the corresponding contradiction* (and so *cannot possibly envisage a failure of LNC*), as she *cannot possibly accept* κ in the first place (on pain of having to accept $\bot$ by contraction, transparency and *modus ponens*). This point is already well-known even if not often stressed (see (Field 2008, pp. 380–381) for a recent restatement). What bears emphasis is that a dual point affects non-LEM theories just as well. Considering a sentence ι identical to $\top \dot{-} T(\ulcorner \iota \urcorner)$ and employing *modus ponens for the unconditional* ($\varphi \vdash \psi, \varphi \dot{-} \psi$) and the *single-premise deduction theorem for the unconditional* (if $\varphi \vdash \psi$ holds, $\varphi \dot{-} \psi \vdash \mathsf{f}$ holds), a dual version of paradox C leads to the conclusions that $\top \vdash T(\ulcorner \iota \urcorner)$, $T(\ulcorner \iota \urcorner) \vdash \mathsf{f}$ and $\top \vdash \mathsf{f}$ hold (I assume that, at least for some unconditional, non-LEM theorists will accept *modus ponens* for it and deny the single-premise deduction theorem for it, thus accepting only the first consequence mentioned). Thus, a dual Curry sentence like ι is an example of a full-blooded paradoxical sentence for which the non-LEM

[43] The point is in effect conceded by a prominent non-LNC theorist who has often emphasised the importance of offering a unified solution to the semantic paradoxes: "[. . .] the curried versions of the paradoxes belong to a quite different family. Such paradoxes do not involve negation and, *a fortiori*, contradiction", "They are paradoxes that involve essentially conditionality [. . .] Genuine Curry paradoxes are therefore ones that depend on a mistaken theory of the conditional, and are perhaps best thought of as more like the 'paradoxes of material implication'" (Priest 2003, pp. 169, 278). In fact, the claim could be made that the fundamental version of Curry's paradox does not satisfy the *inclosure schema* that (Priest 2003) has argued to be at the root of the *semantic paradoxes* (where these are understood as that kind of paradox that is instantiated by the Liar paradox). Although a proper treatment of this issue lies beyond the scope of this paper (see (Zardini 2013b)), I should record that, if that claim were correct, it would seem to me more plausible to take it to reflect badly on the inclosure schema as a diagnosis of the semantic paradoxes rather than on Curry's paradox as a semantic paradox (also given the fact that the rationale behind the claim would equally well establish that Epimenides' paradox and related paradoxes using material implication are not semantic paradoxes!). It may also be worth mentioning that a non-LEM or non-LNC theory might be such that, if the relevant instances of LEM or LNC are added for a certain fragment of the language, that fragment "behaves classically", and so in particular obeys the single-premise deduction theorem (the theory in (Field 2008) is an example of such a system). Even this (possible) technical fact would however be very far from indicating that, in some reasonable sense, the single-premise deduction theorem fails in such a theory *because* LEM or LNC fail in the theory. Compare: the technical fact about intuitionist logic that, if the relevant instances of *Peirce's law* ($\mathsf{t} \vdash ((\varphi \to \psi) \to \varphi) \to \varphi$) are added for a certain fragment of the language, that fragment "behaves classically", and so in particular obeys LEM, is very far from indicating that, in some reasonable sense, LEM fails in intuitionist logic *because* Peirce's law fails in the logic. Thanks to Graham Priest for putting forth to me the claim about the inclosure schema mentioned in this fn.

theorist *cannot possibly reject the corresponding excluded middle* (and so *cannot possibly envisage a failure of LEM*), as she *cannot possibly reject* ι in the first place (on pain of having to reject $\top$ by contraction, transparency and *modus ponens* for the unconditional).

In stark contrast with the severe difficulties for non-LEM and non-LNC theories in giving a unified solution to paradoxes A and C, **IKT** offers a smooth treatment of both: the theory denies subargument C_0 (and its dual version) as it employs contraction, thus blocking paradox C (and its dual version) *in exactly the same way* as paradox A. Since **IKT** blocks all of paradox A, the original version, the most compelling modified versions, the most compelling dual modified versions of paradox B and the original version and the dual version of paradox C in the same way, it offers a *unified* solution to these paradoxes—what non-LEM and non-LNC theories fail to do. Notice that the solution is also *not overdetermined*: all the other principles employed in paradox A, in the original version, the most compelling modified versions, the most compelling dual modified versions of paradox B and in the original version and the dual version of paradox C are valid according to **IKT**—in particular, LEM, LNC, the single-premise reduction theorem, the single-conclusion demonstration theorem, the single-premise deduction theorem and the single-premise deduction theorem for the unconditional (as well as *modus ponens* and *modus ponens* for the unconditional) are all valid according to **IKT**.

The contrast between the unified solution offered by **IKT** and the non-unified solutions offered by non-LEM and non-LNC theories is but amplified if we consider for example the denial of the law of absorption (in terms of which (Curry 1942) originally presented Curry's paradox). I've explained above how, in the theoretic context of transparent theories, that principle is extremely problematic. One could try to justify the principle by saying that, if φ implies $\varphi \rightarrow \psi$, since it also implies φ it should imply ψ, as it implies both premises of *modus ponens*. This justification would not have much force for non-LEM and non-LNC theories, since, for better or worse, on these theories the deduction theorem has to fail in such a dramatic way that, although *modus ponens* is valid, its two premises do not imply its conclusion. But one could also justify the law of absorption by saying that, *under the supposition that* P*, nothing new comes up if it is "further" supposed that* P, so that if, under the supposition that P, if it is "further" supposed that P, it results that Q, then, under the supposition that P, it already results that Q. The law of absorption follows. The argument must presumably be blocked at the assumption that, under the supposition that P, nothing new comes up if it is "further" supposed that P. But, while that assumption is indeed extremely problematic in the absence of contraction, it has nothing to do with LEM, LNC or the single-premise deduction theorem (in fact, sometimes in conversation sympathisers of non-LEM or non-LNC theories have incautiously made fun of **IKT** precisely because of its denial of that and similar assumptions). Not only cannot non-LEM and non-LNC theories offer a unified solution to the Liar and Curry's paradox (as I've argued above in this section), and not only cannot they offer a unified solution to the Liar paradox (as I've argued in Sect. 23.3); they cannot even offer a unified solution to Curry's paradox.

I'd like to close this discussion by making more explicit an underlying theme of this paper. It was relatively easy to make the main point of this section, as it is pretty

clear that paradox C requires non-LEM and non-LNC theories to revise the logic of the *conditional* in (implausible) ways that go beyond what is required by denial of LEM or LNC (in particular, in ways that prevent facts about logical consequence from having the expected effects at the level of facts about the conditional). The ambitious task was rather the one undertaken in Sect. 23.3, to the effect that, in an analogous fashion, already a variation on paradox A like paradox B requires non-LEM and non-LNC theories to revise the logic of *negation* in (implausible) ways that go beyond what is required by denial of LEM or LNC (in particular, in ways that prevent facts about logical consequence from having the expected effects at the level of facts about negation). Thus, it is true that, with respect to paradox A, paradox C reveals new conceptual difficulties for non-LEM and non-LNC theories in the treatment of the conditional, but reflection on paradox B reveals that *analogous difficulties were already present in the treatment of negation*. Indeed, paradox C suggests a *final* version of paradox B which replaces κ with λ, $T(\ulcorner\kappa\urcorner)$ with $T(\ulcorner\lambda\urcorner)$, $\bot$ with f, *modus ponens* with the law of *exclusion* ($\varphi, \neg\varphi \vdash \mathsf{f}$) and the single-premise deduction theorem with the single-premise reduction theorem. I regard this final version of paradox B as one of the most challenging semantic paradoxes for non-LEM, non-LNC and, more generally, non-substructural theories: just as paradox C shows that, very surprisingly, in these theories no conditional that is *weak enough to record logical implication* can still be *strong enough to licence the inference from its antecedent to its consequent* (two jobs that, far for from being in tension, seem to cohere very well with each other), so the final version of paradox B shows that, very surprisingly, in these theories no negation that is *weak enough to record logical falsehood* can still be *strong enough to exclude its negatum* (again, two jobs that, far for from being in tension, seem to cohere very well with each other).

23.5 Getting One for Two

The semantic paradoxes obviously come in many more kinds than the Liar or Curry's paradox, and it is an incumbent task to show how the unified non-contractive solution to the latter two paradoxes presented here can be extended to cover all semantic paradoxes. Such task lies however beyond the scope of this paper—it was here sufficient to show that **IKT** does offer a unified solution to the two most prominent kinds of semantic paradoxes. Such solution is in a sense simple: all the paradoxes reviewed in this paper commit the fallacy of inferring from the fact that a certain premise *taken twice* entails a certain conclusion that the premise *taken once* still entails the conclusion (or the corresponding fallacy concerning contraction in the conclusions). They commit the fallacy of *contracting* two occurrences of a premise or conclusion into one. The fallacy is thus *purely structural*: it does not have anything to do with *specific logical operations* like negation or the conditional. In particular, it does not have anything to do with negation. In fact, in **IKT** negation is in a very good sense *completely classical*, since it obeys the characteristically Boolean principles $\neg$-L and $\neg$-R: $\neg\varphi$ is the sentence that holds in all and only those cases in which φ fails to hold. Non-LEM and non-LNC theories suppose otherwise, and assume

that rejecting either of these features of Boolean negation is the key to the semantic paradoxes. However, we've seen that such rejection is neither here nor there even for some versions of the Liar paradox, let alone for Curry's paradox.

But why is getting one for two a fallacy? Why does contraction fail? My own view—which I have to some extent developed in (Zardini 2011, pp. 503–506) and of which I can only offer the merest sketch here—is that it fails because the relevant sentences express *unstable states-of-affairs*, i.e. states-of-affairs that *lead to consequences without thereby co-obtaining with them*.[44] If φ expresses the state-of-affairs s_0, $\varphi, \varphi \vdash \psi$ may hold, let's suppose, only because s_0 and some state-of-affairs s_1 consequence of s_0 *together* directly lead to the state-of-affairs s_2 expressed by ψ. If s_0 is however unstable, it does not follow that s_0 *by itself* leads to s_2, and so it does not follow that $\varphi \vdash \psi$ holds. For, although s_0 does of course by itself lead to its consequence s_1, by its instability s_0 *need not co-obtain with* s_1, while we're supposing that s_0 can lead to s_2 *only together with* s_1. Failure of contraction is thus the logical symptom of an underlying unstable metaphysical reality.[45] The investigations in this paper invite then the conjecture that it is precisely this unstable reality that is at the root of the semantic paradoxes.

References

Ashworth, E. (1974). *Language and logic in the post-medieval period*. Dordrecht: Reidel.
Beall, J. C. (2009). *Spandrels of truth*. Oxford: Oxford University Press.
Bocheński, J. (1970). *A history of formal logic* (2nd ed.). New York: Chelsea.
Brady, R. (2006). *Universal logic*. Stanford: CSLI Publications.

[44] As usual, states-of-affairs are abstract entities that can either obtain or fail to obtain. A locution like 'State-of-affairs s_0 leads to state-of-affairs s_1' must be understood as 'The *obtaining* of state-of-affairs s_0 leads to the *obtaining* of state-of-affairs s_1'.

[45] On this view, the state-of-affairs t_0 expressed by the Curry sentence τ identical to $T(\ulcorner\tau\urcorner) \rightarrow \top$ is unstable even if $\varphi \vdash \psi \rightarrow \varphi$ does hold in **IKT** (by I, K-L and $\rightarrow$-R), and so even if t_0 unproblematically obtains. For it is still the case that t_0 leads to the state-of-affairs t_1 expressed by $T(\ulcorner\tau\urcorner)$ without *thereby* co-obtaining with it: t_0 does obtain, and does obtain even if t_1 does, and so leads to t_1 and co-obtains with it, but it does all this *only because it is a consequence of the state-of-affairs expressed by* $\top$—the *source* of t_0 co-obtaining with t_1 relies in the state-of-affairs expressed by $\top$ rather than in t_0 itself. In this (intended) sense, t_0 leads to t_1 without *thereby* co-obtaining with it (contrast, for example, with the state-of-affairs that snow is white leading to the state-of-affairs that 'Snow is white' is true and *thereby* co-obtaining with it). Thus, on this view, t_0 *is unstable*, contraction on τ cannot be *built into the relevant system* (although, since $\mathsf{t} \vdash \top$ holds in the system, contraction on τ can be *derived in it*) and Curry's paradox cannot be used to *prove* $\top$ (since the contraction step is only justified by an antecedent proof of $\top$). Instability *only defeasibly* brings about failure of contraction. Notice that, while non-LEM and non-LNC theories can uphold claims analogous to the second and third one in the second last sentence, it is utterly unclear whether they can identify an analogous defeasible cause that would support those claims (for one thing, indeterminacy or overdeterminacy won't do). Analogous points apply of course to the relevant versions of Epimenides' paradox. Thanks to Daniele Sgaravatti for discussion of this issue.

Curry, H. (1942). The inconsistency of certain formal logics. *The Journal of Symbolic Logic, 7,* 115–117.
Field, H. (2008). *Saving truth from paradox*. Oxford: Oxford University Press.
Girard, J.-Y. (1987). Linear logic. *Theoretical Computer Science, 50,* 1–102.
Girard, J.-Y. (1998). Light linear logic. *Information and Computation, 143,* 175–204.
Goldstein, L. (1986). Epimenides and Curry. *Analysis, 46,* 117–121.
Grišin, V. (1974). A nonstandard logic and its applications to set theory. In *Studies in formalized languages and nonclassical logics* (pp. 135–171). Moscow: Nauka.
Grišin, V. (1981). Predicate and set-theoretic calculi based on logics without contractions. *Izvestija Akademij Nauk SSSR Serija Matematičeskaja, 45,* 47–68.
Gupta, A., & Belnap, N. (1993). *The revision theory of truth*. Cambridge MA: MIT Press.
Kripke, S. (1975). Outline of a theory of truth. *The Journal of Philosophy, 72,* 690–716.
López de Sa, D., & Zardini, E. (2006). Does this sentence have no truthmaker? *Analysis, 66,* 154–157.
López de Sa, D., & Zardini, E. (2007). Truthmakers, knowledge and paradox. *Analysis, 67,* 242–250.
López de Sa, D., & Zardini, E. (2011). No-no. Paradox and consistency. *Analysis, 71,* 472–478.
Makinson, D. (1965). The paradox of the preface. *Analysis, 25,* 205–207.
McGee, V. (1991). *Truth, vagueness, and paradox*. Indianapolis: Hackett.
Mortensen, C., & Priest, G. (1981). The truth teller paradox. *Logique et Analyse, 24,* 381–388.
Nuchelmans, G. (1992). A 17th-century debate on the *consequentia mirabilis*. *History and Philosophy of Logic, 13,* 43–58.
Pastore, A. (1936). *La logica del potenziamento.* Naples: Rondinella.
Petersen, U. (2000). Logic without contraction as based on inclusion and unrestricted abstraction. *Studia Logica, 64,* 365–403.
Petersen, U. (2002). *Diagonal method and dialectical logic*. Osnabrück: Der Andere.
Priest, G. (2003). *Beyond the limits of thought* (2nd ed.). Oxford: Oxford University Press.
Priest, G. (2006). *In contradiction* (2nd ed.). Oxford: Oxford University Press.
Sorensen, R. (2001). *Vagueness and contradiction*. Oxford: Oxford University Press .
Stepanov, V. (2007). Propositional logics of reflexive sentences. *Naučno-Techničeskaja Informacija Serija 2*, *5*, 8–14.
Urbas, I. (1996). Dual-intuitionistic logic. *Notre Dame Journal of Formal Logic, 37,* 440–451.
White, R. (1987). A demonstrably consistent type-free extension of the logic BCK. *Mathematica Japonica, 32,* 149–169.
White, R. (1993). A consistent theory of attributes in a logic without contraction. *Studia Logica, 52,* 113–142.
Zardini, E. (2008). Truth and what is said. *Philosophical Perspectives, 22,* 545–574.
Zardini, E. (2011). Truth without contra(di)ction. *The Review of Symbolic Logic, 4,* 498–535.
Zardini, E. (2012). Truth preservation in context and in its place. In C. Dutilh-Novaes & O. Hjortland (Eds.), *Insolubles and consequences* (pp. 249–271). London: College Publications.
Zardini, E. (2013a). Naive *modus ponens*. *Journal of Philosophical Logic* (Forthcoming).
Zardini, E. (2013b). Naive truth and naive logical properties. *The Review of Symbolic Logic* (Forthcoming).
Zardini, E. (ms, 2013c). It is not the case that [*P* and 'It is not the case that *P*' is true].
Zardini, E. (ms, 2013d). Restriction by non-contraction.
Zardini, E. (ms, 2013e). The opacity of truth.

Chapter 24
Kripke's Thought-Paradox and the 5th Antinomy

Graham Priest

Abstract In 'A Puzzle about Time and Thought' Saul Kripke published a new paradox. The paradox is clearly a relative of Russell's paradox; but it deploys, as well as the notion of set, an intentional notion, *thought*. This ensures that it raises significantly different issues from Russell's paradox. Notably, the solution to Russell's paradox provided by ZF does not apply in any obvious way to this paradox.

In this paper I will first explain Kripke's paradox and compare it with another paradox which deploys the notion of thought. I will then show that it fits the Inclosure Schema, and so may be expected to have a solution which is the same as other inclosure paradoxes. Next, the paradox is stripped down to a much more acute form. Finally, in the light of this, some thoughts concerning possibilities for resolving the paradox are offered.

24.1 Introduction

In 'A Puzzle about Time and Thought'[1] Saul Kripke published a new paradox. The paradox is clearly a relative of Russell's paradox; but it deploys, as well as the notion of set, an intentional notion, *thought*. This ensures that it raises significantly different issues from Russell's paradox. Notably, the solution to Russell's paradox provided by ZF does not apply in any obvious way to this paradox.

In this paper I will first explain Kripke's paradox and compare it with another paradox which deploys the notion of thought. I will then show that it fits the Inclosure Schema, and so may be expected to have a solution which is the same as other inclosure paradoxes. Next, the paradox is striped down to a much more acute form. Finally, in the light of this, some thoughts concerning possibilities for resolving the paradox are offered.

[1] Kripke (2011).

G. Priest
Departments of Philosophy, The Graduate Center, City University of New York, New York, USA, and the University of Melbourne, Australia
e-mail: priest.graham@gmail.com

T. Achourioti et al. (Eds.), *Unifying the Philosophy of Truth,* Logic, Epistemology, and the Unity of Science 36, DOI 10.1007/978-94-017-9673-6_24

24.2 Puzzles About Thought

A number of paradoxes of self-reference employ intensional notions. In what follows, it will be helpful to have in mind one particular one of these.[2]

Dedekind proved the existence of an infinite set as follows. Take any object you like—for the sake of illustration, let us say the empty set; then there is a thought of it; a thought of a thought of it; a thought of a thought of a thought of it; and so on. The resulting totality is infinite.[3] But we can iterate the process. Let s be the set of all these thoughts. Then there is a thought of s; a thought of a thought of s; and so on. We can iterate the process indefinitely. Indeed, we can simplify the process, at the same time as making it uniform. Suppose we index the stages by ordinals. Let us write tx for 'the thought of x'. Then we can define a sequence, f, as follows: $f(\alpha) = t\{f(\beta) : \beta < \alpha\}$. Each thought generated in this sequence is distinct—just as the ordinals are. Let Π be the totality of all thoughts generated in this way.[4] Then there is no such thing as $t\Pi$. If there were, it would be the next member of the series, which, *ex hypothesi*, there is not. But *there is* a thought of Π. Indeed, you have just had it.

In *Beyond the Limits of Thought*,[5] and with reference to Kant, I called this paradox the *5th Antinomy*.[6] The paradox is obviously, in some sense, an intentional version of Burali-Forti's paradox. However, it does no good to try to solve it, as one solves the Burali-Forti paradox in ZF, by saying that Π does not exist. Notoriously, there can be thoughts of non-existent objects, such as Pegasus and Sherlock Holmes.

Now to Kripke's paradox, which is as follows. Let k be the set of all times at which I am thinking of a set of times of which that time itself is not a member. There is a time at which I think of k. (There has just been such a time.) Let this be τ. Then by familiar reasoning, τ is both in and not in k.

Let us spell out the paradox more formally. Let $\theta_t x$ be 'I am thinking about x at time t. Let T be the set of all times. Then $k = \{t \in T : \exists s \subseteq T(\theta_t s \wedge t \notin s)\}$. We are given that, at τ, k is the one and only thing I'm thinking about: $\theta_\tau x \leftrightarrow x = k$. Suppose that $\tau \notin k$; then, since $k \subseteq T$, $\exists s \subseteq T(\theta_\tau s \wedge \tau \notin s)$, namely, when s is k; that is, $\tau \in k$. So $\tau \in k$. But then $\exists s \subseteq T(\theta_\tau s \wedge \tau \notin s)$. And since the $\theta_\tau(s)$ entails that $s = k$, $\tau \notin k$.[7]

We can give an equivalent formulation of the paradox using definite descriptions. Let us use ι as a definite description operator, and E as an existence predicate (so that

---

[2] For a general discussion of intensional paradoxes, see Priest (1991).

[3] Dedekind (1888), Theorem 66. 'Thought' here means content, not act. Actual thoughts must give out after some finite time.

[4] Presumably, there are ordinal-many; but it doesn't really matter if the sequence peters out before the ordinals are exhausted.

[5] Priest (2002), hereafter, BLoT.

[6] BLoT, 6.9.

[7] I note that there is also a Curried version of the paradox. Let $k' = \{t \in T : \exists s \subseteq T(\theta_t s \wedge (t \in s \rightarrow \bot))\}$. Then reasoning in a natural way, one establises that $\tau \in k' \rightarrow (\tau \in k' \rightarrow \bot)$, and hence, by contraction, that $\tau \in k' \rightarrow \bot$. It follows that $\exists s \subseteq T(\theta_\tau s \wedge (\tau \in s \rightarrow \bot))$, that is, $\tau \in k'$; hence $\bot$.

$E(\iota x A)$ means $\exists y \forall x (A \leftrightarrow x = y)$. Then $k = \{t \in T : E(\iota x \theta_t x) \wedge t \notin \iota x \theta_t x\}$. Hence, $t \in k \leftrightarrow (E(\iota x \theta_t x) \wedge t \notin \iota x \theta_t x)$. We are now given as a premise that $k = \iota x \theta_\tau x$. *A fortiori*, $E(\iota x \theta_\tau x)$. Then $\tau \in k \leftrightarrow \tau \notin k$, and so we have a contradiction.

And just as with the 5th Antinomy, one cannot deploy the resources of standard set theory to solve it. In particular k is a subset of times (that is, we may suppose, a subset of the real numbers), and so is not an absolutely infinite set. One cannot, therefore, say that k does not exist on this ground, as one says of Russell's set in ZF.

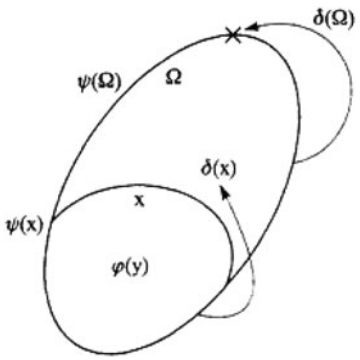
24.3 The Inclosure Schema

All the standard paradoxes of self-reference are *inclosure paradoxes*. Such paradoxes are ones that fit the Inclosure Schema. This it is which provides their underlying structure.[8] The Schema concerns an operator, δ, and conditions, $\varphi(y)$ and $\psi(x)$, such that, where $\Omega = \{y : \varphi(y)\}$, it appears to be the case that:

- Ω exists and $\psi(\Omega)$ (Existence)
- For all x such that $\psi(x)$ and $x \subseteq \Omega$:
  - $\delta(x) \notin x$ (Transcendence)
  - $\delta(x) \in \Omega$ (Closure)

Let is write $A \wedge \neg A$ as $A!$ The paradox arises when we take Ω itself for x. Then Transcendence and Closure give $\delta(\Omega) \in \Omega$!

The Scheme may be depicted as follows:



Applying δ to Ω gives an object that is both within and without Ω (depicted as on its boundary).

Kripke's paradox is an inclosure paradox too. In this:

- $\varphi(t)$ is $\exists s \subseteq T(\theta_t s \wedge t \notin s)$

[8] BLoT, 9.4, 11.0, 17.2.

- $\psi(x)$ is $\exists t \forall y(\theta_t y \leftrightarrow y = x)$
- $\delta(x)$ is $\varepsilon t \forall y(\theta_t y \leftrightarrow y = x)$

Given that $\psi(x)$, $\delta(x)$ is well-defined. That is, we have: $\forall y(\theta_{\delta(x)} y \leftrightarrow y = x)$.[9]

Checking the conditions of the Inclosure Schema: $\Omega = \{t \in T : \exists s \subseteq T(\theta_t s \wedge t \notin s)\}$. Being a subset of T, Ω exists, and we are given that $\psi(\Omega)$. (Existence). Suppose that $\psi(x)$ and $x \subseteq \Omega$. If $\delta(x) \in x$ then $\delta(x) \in \Omega$. That is, $\exists s \subseteq T(\theta_{\delta(x)} s \wedge \delta(x) \notin s)$. But since $\forall y(\theta_{\delta(x)} y \leftrightarrow y = x)$, this s is x. That is, $\delta(x) \notin x$. Hence, $\delta(x) \notin x$. (Transcendence). Thus, $\theta_{\delta(x)} x \wedge \delta(x) \notin x$. And since $x \subseteq \Omega \subseteq T$, we have $\exists s \subseteq T(\theta_{\delta(x)} s \wedge \delta(x) \notin s)$. That is, $\delta(x) \in \Omega$. (Closure.) The paradox is that $\delta(\Omega) \in \Omega$! In the presentation of Kripke's paradox above, $k = \Omega$ and $\tau = \delta(k)$.

The following tabulates the components of the Inclosure Schema for various paradoxes for comparison:

	$\varphi(y)$	$\psi(x)$	$\delta(x)$
Burali-Forti	y is an ordinal	$x = x$	$\mu z \forall y \in x\, z > x$
König	y is a definable ordinal	x is definable	$\mu z \forall y \in x\, z > x$
5th antinomy	$\exists \alpha\, y = f(\alpha)$	$x = x$	tx
Russell	y is a set	$x = x$	$\{z \in x : z \notin z\}$
Kripke	$\exists s \subseteq T(\theta_y s \wedge y \notin s)$	$\exists t \forall y(\theta_t y \leftrightarrow y = x)$	$\varepsilon t \forall y(\theta_t y \leftrightarrow y = x)$

For good measure (and future reference) I have thrown in König's paradox. The number of definable ordinals is countable. There must therefore be a least. By definition, this is indefinable; but it is defined by the 'the least indefinable ordinal'.

24.4 Stripping Down the Paradox

Kripke's paradox is clearly modelled on Russell's. However, what we will now see is that it can be stripped-down and simplified in such a way as to be much more acute.

First, note that the contradiction is generated by a thought that is being had at time τ (now). That thoughts might also be had at other times is irrelevant. Hence, we can drop the subscript from $\theta_t x$, and just understand θx as: x is the one and only set being thought of (now). k then becomes: $\{t \in T : \exists s \subseteq T(\theta s \wedge t \notin s)\}$ (the set of all times which are not members of the set I am thinking about now). Let it be given, again, that I am thinking (now) about just this set: $\theta y \leftrightarrow y = k$. Let t be *any* time. Suppose that $t \notin k$. Then clearly $\exists s \subseteq T(\theta s \wedge t \notin s)$, namely, when s is k itself.

[9] ε is an indefinite description operator. Since times are not well-ordered, one cannot assume that any non-empty set of times has a first member. But thinkers being finite, if there are any times at which I think of just x, there is, presumably, a first. The indefinite description could therefore be traded in for a definite description.

That is, $t \in k$. So $t \in k$. But then $\exists s \subseteq T(\theta s \wedge t \notin s)$. And since $\theta(s)$ entails that $s = k$, $t \notin k$.

Next, note that the fact that it is a set of times which is being thought about, is irrelevant: it could be people (as Kripe indicates), numbers, or any other kind of thing. Indeed, it could be a very small set—one which is either ϕ (0) or $\{\phi\}$ (1). Let us redefine k in this way, as follows: $\{n < 1 : \exists s \subseteq 1(\theta s \wedge n \notin s)\}$. Given, again, that $\theta x \leftrightarrow x = k$, we have:

(*) $0 \in k \leftrightarrow (\theta k \wedge 0 \notin k)$

and then $0 \in k$ and $0 \notin k$. Hence either $0 \in 0$ or $0 \notin 1$.

I note that the stripped-down version of the paradox still fits the Inclosure Schema: $\phi(y)$ is $\exists s \subseteq 1(\theta s \wedge y \notin s)$; $\psi(x)$ is $\forall y(\theta y \leftrightarrow y = x)$; $\delta(x) = 0$. The details are easy to check.

24.5 Solving the Paradox

In this final section, I will make some comments on solving Kripke's paradox, and in particular, its stripped-down form. As we have seen, these paradoxes are inclosure paradoxes. Since all inclosure paradoxes are of the same form, they require the same sort of solution (the Principle of Uniform Solution).[10] This imposes tight constraints on acceptable solutions.

Let us start with the two possible solutions which Kripke himself mentions (though he does not commit himself to either of these). The first is some form of ramification. Specifically, in his version of the paradox (similar comments can be applied to the stripped-down version) the predicate θ_t is ramified into a hierarchy of predicates θ_t^n. The definition of k then becomes $\{t \in T : \exists s \subseteq T(\theta_t^n s \wedge t \notin s)\}$. Since k is specified by a predicate of order n, any thought of it must be of order $n + 1$. Hence we have only $\theta_t^{n+1} x \leftrightarrow x = k$, and the argument is broken. Ramification has many and well known problems, which it is out of place to discuss here.[11] Here I note only that ramification will not solve the 5th Antinomy. The construction in this "blows the top" off all ramification.

The second possible solution which Kripke notes is to use a non-classical logic of the kind described in Kripke (1975). This breaks the paradox by rejecting the Law of Excluded Middle (LEM). Arguably, the most sophisticated extant account of this kind is that given by Field (2008). Again, this is not the place to discuss the problems of this kind of approach in detail.[12] I note here, again, only that the approach does not seem to be able to handle the 5th Antinomy, since this does not employ the LEM.[13]

---

[10] This is argued in BLoT, Chap. 11.

[11] See, e.g, BLoT, Chaps. 9 and 10.

[12] See Priest (2006), Chap. 1, and Priest (2010a).

[13] Or other paradoxes that do not employ the LEM, such as Berry's and König's paradoxes. See Priest (2010a), Sect. 6. Note that the 5th Antinomy does not even deploy the least-number principle, as these do.

Another possible solution is to deny the empirical premise of the paradoxes: that there was a time when I was thinking of the one thing, k. This is the line taken by Prior (1961) in another paradox involving thought. This is a version of the Liar paradox concerning a thought of the form 'Every thought being had in Room 7 at the present moment is false', where that is the only thought being had in Room 7 at that time. Prior avers that, much as one might suppose otherwise, no such unique thought is being had. This strikes me as a move of desperation. When a mathematician says 'Think of the set of prime numbers. I will show you that it is infinite', it is clear that their utterance gets us to think of the set of prime numbers. So it is with k. (k does not even have to exist, note, to be thought of.) Could it be that when I think of k I think of something *else* as well? This can hardly be the case: it is of the nature of intentionality to be directed at a single target. Moreover, the assumption of uniqueness plays no role in the 5th Antinomy, so this move is of no avail.[14] In any case, and again, the whole line obtains no purchase with respect to the set theoretic paradoxes, such as Russell's and Burali-Forti's.

Finally, to dialethic solutions. In BLoT, Part 3, I argued that all inclosure paradoxes require such a solution. There had better, therefore, be such a solution to Kripke's paradox and its stripped-down form. A dialetheic solution to the paradoxes involves accepting a contradiction delivered by the paradoxical argument. In many cases, this may be the explicit paradoxical conclusion, but it need not be. Sometimes the argument may deliver only a disjunction of contradictions.[15] In the case of Kripke's paradox, the most obvious dialetheic thought is that the time τ is both in and not in k. Assuming that we identify times with the real numbers, this shows us that a certain set of real numbers is contradictory. Perhaps that is palatable. But in the stripped-down version of the paradox, the corresponding thought is that a small finite set—one which has either no members of whose only member is the empty set—is contradictory. This would appear to be much harder to swallow.

At this point, we need break the discussion into two cases. One may formulate the axioms of naive paraconsistent set theory using either a detachable conditional or a material non-detachable conditional.[16] Suppose that we adopt the first of these possibilities, and take the comprehension principle in a completely unrestricted form, where the set being specified can occur in the specification itself (building in the possibility of a certain fixed point).[17] Then every non-empty set, a, has an inconsistent subset. Thus we have: $x \in c \leftrightarrow (x \in a \wedge x \notin c)$. Clearly, $c \subseteq a$. And given that

[14] For a more general critique of Prior, see Priest (1991).

[15] Thus, in the case of the sorites paradox, the argument forces the conclusion that at least *one* of the objects in the sorites progression is contradictory. See Priest (2010b). More generally, see BLoT, p. 130, fn. 7.

[16] Both possibilities are considered in Priest (2006), Chap. 18.

[17] This form of the principle is found in Routley (1977) and Weber (2010). It is known to be very powerful, but non-trivial.

$z \in a$, we can infer that $z \in k$ and $z \notin k$. It may not, then, be surprising that k has an inconsistent subset (itself).[18]

However, I think a more plausible solution involves taking the second option in formulating paraconsistent set theory, where the conditional is material. In that case, the biconditional (*) delivers only $0 \in k! \vee \neg\theta k$; and then the empirical premise gives us $0 \in k! \vee \theta k!$. We can put the contradictory blame on the second disjunct: k both is and is not being thought about. This is exactly the case in the 5th Antinomy, of course. There is no thought of the totality, Π; *a fortiori* it cannot be thought about; but, by inspection, one can.

There are other precedents. Consider König's paradox. This concerns definability, not thought; but these are not that different: thinking about something can be taken to be bringing before the mind an appropriate noun-phrase which refers to it.[19] Moreover, the Denotation schema involved in König's paradox, 'n' defines $x \leftrightarrow x = n$, is the analogue of the condition $\theta(x) \leftrightarrow x = k$ involved in Kripke's paradox. And given that confluence, the paradoxes have similar shape. In both paradoxes, that a certain object *cannot* be thought about or named is established by a theoretical argument; and that it *can* be thought about or named is established by "direct inspection".

Kripke's paradox and the second dialetheic solution therefore seem in good company.[20]

References

Dedekind, R. (1888). *Was sind und was sollen die Zahlen?* Reprinted in English translation as Part II of *Essays on the theory of numbers*, 1955. New York: Dover Publications.

Field, H. (2008). *Saving truth from paradox*. Oxford: Oxford University Press.

Kripke, S. (1975). Outline of a theory of truth. *Journal of Philosophy, 72,* 690–716.

Kripke, S. (2011). A puzzle about time and thought (ch. 13 of *Philosophical troubles*). Oxford: Oxford University Press.

Priest, G. (1991). Intensional paradoxes. *Notre Dame Journal of Formal Logic, 32,* 193–211.

Priest, G. (2002). *Beyond the limits of thought* (2nd (extended) ed.). Oxford: Oxford University Press.

Priest, G. (2006). *In contradiction* (2nd (extended) ed.). Oxford: Oxford University Press.

Priest, G. (2010a). Hopes fade for saving truth. *Philosophy, 85,* 109–140.

[18] It is important to note that in this version of paraconsistent set theory, sets that are extensionally equivalent, in a certain sense, may be distinct. Thus, one may be able to show that there is nothing satisfying either $\varphi(x)$ or $\psi(x)$. But this does not suffice to establish that $\forall x(\varphi(x) \leftrightarrow \psi(x))$, and so that $\{x : \varphi(x)\} = \{x : \psi(x)\}$. In particular, then, the set I have labelled '1' need not be the same as the von Neumann ordinal 1.

[19] The connection between thinking and referring is noted in BLoT, 4.8.

[20] Versions of this paper were given at a one day workshop of the Melbourne Logic Group in August 2011, a conference on Kripke's *Philosophical Troubles*, at the Graduate Center, CUNY, in September 2011, and to the Logic Group at the University of Indiana, Bloomington, in October 2011. Thanks go to the people in those audiences for their thoughts, and particularly to Colin Caret, Lloyd Humberstone, and, especially, Zach Weber.

Priest, G. (2010b). Inclosures, vagueness, and self-reference. *Notre Dame Journal of Formal Logic, 51*, 69–84.

Prior, A. (1961). On a family of paradoxes. *Notre Dame Journal of Formal Logic, 2,* 16–32.

Routley, R. (1977). Ultralogic as universal? *Relevance Logic Newsletter*, 2, 51–89. Reprinted as an appendix to Exploring Meinong's jungle and beyond, 1980. Canberra: Australian National University.

Weber, Z. (2010). Transfinite numbers in paraconsistent set theory. *Review of Symbolic Logic, 3,* 1–22.