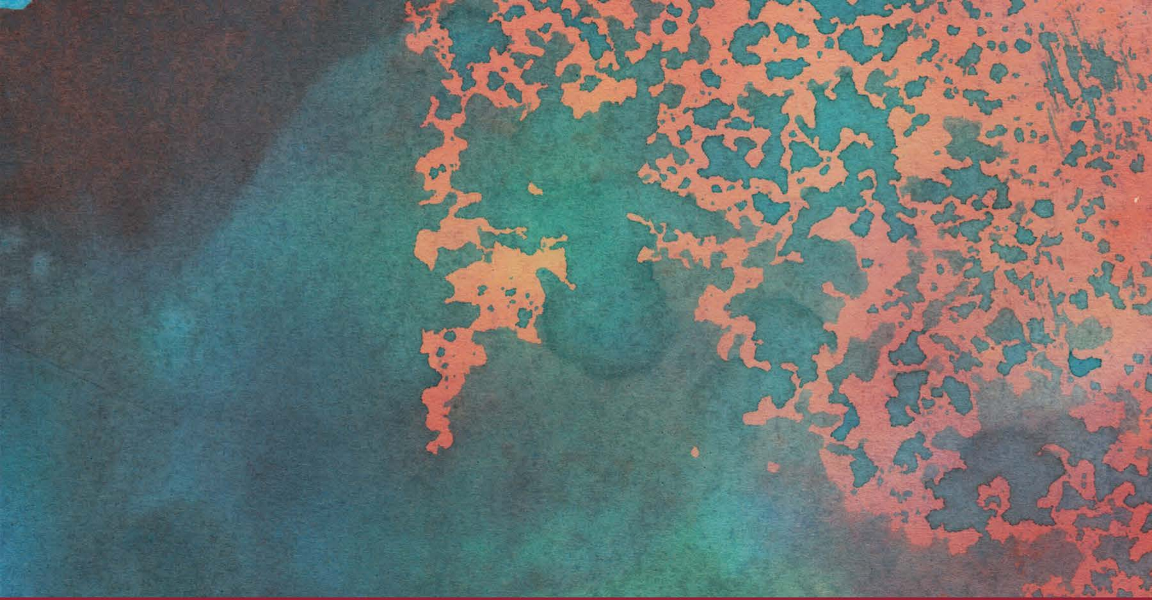
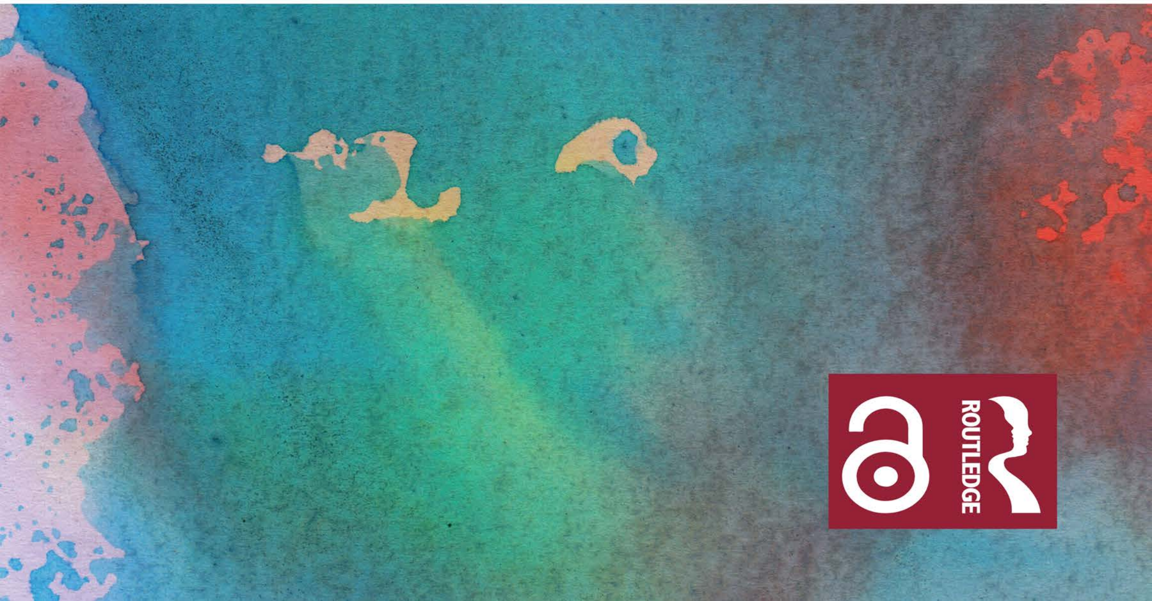


Routledge Studies in Contemporary Philosophy

MODES OF TRUTH

THE UNIFIED APPROACH TO TRUTH, MODALITY, AND PARADOX

Edited by
Carlo Nicolai and Johannes Stern



Modes of Truth

The aim of this volume is to open up new perspectives and to raise new research questions about a unified approach to truth, modalities, and propositional attitudes. The volume's essays are grouped thematically around different research questions. The first theme concerns the tension between the theoretical role of the truth predicate in semantics and its expressive function in language. The second theme of the volume concerns the interaction of truth with modal and doxastic notions. The third theme covers higher-order solutions to the semantic and modal paradoxes, providing an alternative to first-order solutions embraced in the first two themes. This book will be of interest to researchers working in epistemology, logic, philosophy of logic, philosophy of language, philosophy of mathematics, and semantics.

Carlo Nicolai is Lecturer in Philosophy at King's College London, UK. He was previously a VENI (NWO) Research Fellow at the University of Utrecht, The Netherlands.

Johannes Stern is Research Fellow and permanent member of staff at the Department of Philosophy of the University of Bristol, UK. He directs the ERC Starting Grant *Truth and Semantics*.

Routledge Studies in Contemporary Philosophy

Caring for Liberalism
Dependency and Liberal Political Theory
Edited by Asha Bhandary and Amy R. Baehr

Language and Phenomenology
Edited by Chad Engelland

The Philosophy and Psychology of Ambivalence
Being of Two Minds
Edited by Berit Brogaard and Dimitria Electra Gatzia

Concepts in Thought, Action, and Emotion
New Essays
Edited by Christoph Demmerling and Dirk Schröder

Towards a Philosophical Anthropology of Culture
Naturalism, Reflectivism, and Skepticism
Kevin M. Cahill

Examples and Their Role in Our Thinking
Ondřej Beran

Extimate Technology
Self-Formation in a Technological World
Ciano Aydin

Modes of Truth
The Unified Approach to Truth, Modality, and Pardox
Edited by Carlo Nicolai and Johannes Stern

For more information about this series, please visit: https://www.routledge.com/Routledge-Studies-in-Contemporary-Philosophy/book-series/SE0720

Modes of Truth

The Unified Approach to Truth, Modality, and Paradox

Edited by Carlo Nicolai and Johannes Stern

Routledge
Taylor & Francis Group
NEW YORK AND LONDON

First published 2021
by Routledge
52 Vanderbilt Avenue, New York, NY 10017

and by Routledge
2 Park Square, Milton Park, Abingdon, Oxon, OX14 4RN

Routledge is an imprint of the Taylor & Francis Group, an informa business

Trademark notice: Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

Library of Congress Cataloging-in-Publication Data
Names: Nicolai, Carlo, 1984- editor. | Stern, Johannes, editor.
Title: Modes of truth : the unified approach to truth, modality, and paradox / edited by Carlo Nicolai and Johannes Stern.
Description: New York, NY : Routledge, 2021. | Series: Routledge studies in contemporary philosophy | Includes bibliographical references and index.
Identifiers: LCCN 2020058223 (print) | LCCN 2020058224 (ebook) | ISBN 9780367141097 (hbk) | ISBN 9780429030208 (ebk)
Subjects: LCSH: Truth.
Classification: LCC BD171 .M54 2021 (print) | LCC BD171 (ebook) | DDC 121—dc23
LC record available at https://lccn.loc.gov/2020058223
LC ebook record available at https://lccn.loc.gov/2020058224

ISBN: 978-0-367-14109-7 (hbk)
ISBN: 978-0-367-68867-7 (pbk)
ISBN: 978-0-429-03020-8 (ebk)

Typeset in Sabon
by Apex CoVantage, LLC

Contents

1 A Guide to the Unified Approach to Truth, Modality, and Paradox

Carlo Nicolai and Johannes Stern

The notion of truth, modal notions, and doxastic notions such as belief and knowledge play a crucial role in contemporary philosophy. From its very beginnings philosophy has sought to fully understand these notions. Following the *linguistic turn* and the rise of analytic philosophy, the study of the uses of the aforementioned notions in (natural) language has become an important part of the data against which a theory has to be tested and sharpened. Indeed research on truth within this tradition has focused on the role and uses of the *truth predicate* in language—be it in natural language or some more regimented theoretical language. Research on belief has been importantly influenced by how *belief reports* are to be best understood and similar remarks apply to modal notions and knowledge.

In natural language we say

(1) This apple is red.
(2) Goldbach's conjecture is true.
(3) Goldbach's conjecture is necessary.
(4) Mary believes Goldbach's conjecture.

In light of these examples it seems, at least prima facie, that saying of Goldbach's conjecture that it is true, possible, or believed is not fundamentally different from the claim made in (1): in both cases we seem to ascribe a particular characteristic, feature, or property to an object, that is, a specific apple in (1) but the denotatum of Goldbach's conjecture in (2), (3), and (4). In other words 'Goldbach's conjecture' occupies a subject position in (2), (3), and (4) that typically can be occupied by first-order singular terms and can be bound by first-order quantification.[1] The objects of truth are not apples, however, but commonly thought to be sentences (types), utterances, or propositions.[2] While sentences or propositions are, of course, objects of a very different kind from apples, they remain objects of the same semantic type on this view. In the context of Montague Grammar, they will all be objects of type *e*. As a consequence, the truth predicate enables generalization over sentences or propositions by means of standard first-order quantification. Indeed this feature is widely acknowledged as one of the principal uses of the truth predicate in language (see, e.g., Quine, 1970).

However, according to an alternative view generalization over sentences or propositions does not arise via quantification into the argument position of a sentential predicate, but rather via some sort of higher-order quantification over sentence position. On this view, in sentences like (2), despite appearances to the contrary, we do not ascribe truth to some object of type *e*. Rather, on this alternative semantic picture the truth predicate will not occur in the logical form of the sentence. Instead the truth predicate will be eliminated and replaced by propositional quantification, which is standardly analysed as quantification over objects of type $\langle s, t\rangle$. Arguably this line of research originates with Ramsey's Redundancy Theory of Truth (Ramsey, 1927, 1929) and finds more contemporary proponents in the form of the Prosentential Theory of Truth (Grover et al., 1975; Grover, 1992) and Mulligan (2010).

In the literature on truth the idea that the quantifier in sentences such as

(5) Every axiom of ZFC is either true or false

is a propositional, i.e., a higher-order quantifier ranging over objects of type $\langle s, t\rangle$ remains a minority position and truth is usually formalized (and interpreted) by a first-order predicate constant. In contrast, and somewhat surprisingly, in philosophical logic it is customary to formalize attitude verbs such as 'believe', as well as, knowledge and various modal notions by means of sentential operators that apply to object of type $\langle s, t\rangle$. Quantification into the argument position of attitude verbs, knowledge, and the modal notions then need to take the form of propositional quantification, that is, quantification over objects of type $\langle s, t\rangle$ (Bull, 1969; Fine, 1970).[3]

The discrepancy in the formal treatment of truth, modality, and doxastic notions leads to the problem of cross-quantification: how is the quantifier in sentences like

(6) Everything Mary believes is true

best understood? Is it a first-order quantifier as required by the standard view on truth or is it a propositional quantifier that binds the argument position of the attitude verb if the latter is formalized as a doxastic operator?

Unfortunately, sentences like (6) are abundant in natural language and, specifically, philosophical discourse, so the problem of quantification cannot be put aside. To resolve the problem, some level of uniformization between the formal treatment of truth, doxastic notions, and modal notions seems required. In other words, the notion of truth, doxastic notions, and modal notions as they appear in sentences like (6) need to be treated as expressions of the same logical category to allow for the type of cross-quantification displayed in (6).[4]

If the grammatical form of sentences (2)–(6) is taken as a guide, attitude verbs, knowledge, and modal notions may be best conceived of as predicates. This would facilitate interpreting quantification in these sentences along the lines of first-order quantification. Alternatively, one could take a revisionist stance towards the surface-level grammatical form and conceive of quantification in these sentences as propositional quantification: there is no singular term position that is bound by the quantifier expression in (5) and (6). The quantifier ranges over sentence position rather than their nominalizations. All else equal, taking the first-order route may seem advantageous. First-order quantification is well understood and it comes with a fully developed semantics and proof theory. In contrast, while there exist attractive semantics for higher-order quantification and more specifically propositional quantification, these semantics typically cripple the language's expressive resources: self-reference is usually eliminated in favour of an open-ended hierarchy of types. As will become clear later, Bacon's *Opacity and Paradox* highlights that this need not be the case, i.e., propositional quantification can accommodate self-reference without appealing to an open-ended hierarchy of types. However, the semantics and proof theory of non-hierarchical theories of propositional quantification is underdeveloped in comparison to the first-order approach.

Putting all of this together, there is a strong case to uniformly conceive of the objects of truth, modality, knowledge, and attitudinal relations as first-order objects that can be quantified over via first-order quantification—call this the Unified Approach. However, grammatical form and a well developed proof theory and semantics do not amount to conclusive arguments for adopting the first-order approach. If convincing arguments in favour of a higher-order approach are available, the latter may well be preferable. Indeed, several philosophers and logicians pursue the higher-order approach. The chapters by Bacon and Studd in this volume contribute to the development of the higher-order approach.

Independently of whether one prefers a first-order or higher-order approach, a uniform treatment of truth, modality, and attitudinal relation seems a prerequisite for an adequate semantics and proof theory of sentences which involve cross-quantification in the sense discussed above. However, unless special precautions are taken, such a uniform treatment of the notion of truth, modal notions, and propositional attitudes will encounter paradoxes such as the liar paradox, Montague's paradox (Myhill, 1960; Montague, 1963), and paradoxes of belief (Thomason, 1980; Cross, 2001).[5] A popular reaction to the paradox is to banish self-reference from the framework by introducing either syntactic or semantic restriction (e.g., typing restrictions) that prevent expressing self-reference of any sort. Such approaches are hardly satisfactory, as has been convincingly argued by Kripke (1975) and others: they don't provide an interpretation of fragments of natural language that fit the data. Rather than banishing self-reference from the framework, strategies have to be put in place for handling vicious or infelicitous forms of self-reference.

The paradoxes not only pose a challenge for any adequate formal account of truth and related notions, they also highlight the need of developing such an account in an unified way, that is, in a framework in which the various notions are allowed to freely interact. As has been stressed by Horsten and Leitgeb (2001), Halbach (2006, 2008), and Stern and Fischer (2015), the interaction of the various notions may trigger new and unexpected pathologies. This indicates that proceeding in a piecemeal fashion and tackling truth, modality, or the propositional attitudes individually rather than simultaneously is bound to lead to problems.[6]

In the first-order framework, self-reference is usually achieved via the various forms of nominalizations available, e.g., via Gödel numbering or other means that enable one to talk about expressions of the language (alternatively, component parts of structured propositions). However, in the higher-order framework, that is, a framework that allows for quantification into sentence-position, the liar-like paradoxes arise directly via quantification into the argument position of the truth-like notions (or related means). In this form the paradoxes are known as paradoxes of indirect discourse, and have been discussed, e.g., by Prior (1961, 1971).[7] The paradoxes of indirect discourse are basically versions of the Epimenides paradox, but in contrast to the more standard liar-like paradoxes, require rethinking the logic and semantics of propositional quantification rather than the logic and semantics of the truth-like notions. An example to this effect is given by Bacon's contribution but also Asher (1990), yet research on the semantics of propositional quantification in light of the paradoxes of indirect discourse is somewhat more scarce than the research on adequate semantics for the truth predicate and related notions. This volume aims to continue research on the advantages and limitations of the first-order framework. Addressing such challenges necessitates a closer look at the truth predicate, its role and function in language, and its semantics. And this is where the journey begins.

1.1 Truth: Semantics and Disquotation

Even if truth is best conceived as a predicate applying to first-order entities, this does not settle many other issues concerning the role of truth in language and reasoning. Theorizing about truth comes in different forms.

1.1.1 Truth and Semantics

The standard *semantic* notion of truth lives in a metalanguage that is distinct from the fragment of natural language one wants to analyse, the object-language. This metalanguage comprises a rich ontology of mathematical objects and resources to talk about the semantic values of the linguistic components of the object-language. Semantic truth is typically defined in such a metalanguage, although it is possible to employ a direct axiomatization of one's metatheoretic truth predicate. Semantic truth can be employed to study the object-linguistic truth predicate, whose properties may not coincide with the semantic one.

Several theorists interpret natural language data as supporting the desideratum that the semantic value of a sentence *A* ought to be the same as the semantic value of '"A" is true', where *A* is an arbitrary sentence of the object-language. There are different ways to construct a formal semantics with this property. The liar paradox tells us that a formal semantics in which *A* and '"A" is true' have the same semantic value cannot validate all principles of classical logic. Kripke (1975) is arguably the starting point of modern investigations on the semantics of self-applicable truth: the formal semantics proposed by Kripke does not validate some of the classical logical principles for negation (or, equivalently, the law of excluded middle). The sentences that generate paradox are, in Kripke's semantics, truth-value gaps.

Kripke's approach is the basis for semantic frameworks that are mathematically very close to Kripke's original semantics, but that are conceptually quite apart. By conceiving of the liar paradox as a datum supporting the inconsistency of truth, paraconsistent approaches to the semantics of the truth predicate are based on the idea that there are sentences that are both true and false (Priest, 2006; Beall, 2009), that is, truth-value gluts. Such sentences crucially contain the object-linguistic truth predicate. The paraconsistent semantics developed in Priest (2006) also restrict some principles of negation (or, equivalently, the classical inference *ex-falso quodlibet*).

A paraconsistent version of Kripke's theory of truth can be constructed to yield models of the language with a self-applicable truth predicate that validate all classical inferences—including the ones involving negation—but that invalidate some classical meta-inferences (Ripley, 2012; Cobreros et al., 2013). Let's restrict our attention to models with only three values $\{1, {}^1\!/_2, 0\}$. Fixed-point models inspired by the non-transitive approach do not validate the structural rule of *cut*. This is achieved by modifying the notion of logical consequence associated with Kripke fixed-point models. Assuming a fully structural notion of logical consequence, φ follows from Γ precisely when designated truth-values are preserved. In the non-transitive approach, φ follows from Γ precisely when, if all sentences in Γ receive value 1, then φ cannot receive value 0. This is the notion of *strict-tolerant consequence*. Since both the liar sentence λ and its negation $\neg\lambda$ do not receive value 0 in any model, they are both valid. However, the non-transitive approach to truth does not lead to triviality, as the meta-inference of cut is not validated.

Paul Égré's *Half Truths and Liars* aims to refine the analysis of the semantics of the object-linguistic truth predicate given by the non-transitive approach. According to the standard non-transitive theorist, sentences such as λ and '$2 + 2 = 4$' do not differ in their strict semantic status: they are both valid. However, unlike what happens with $2 + 2 = 4$, the negation of λ is also valid, but there is no sense in which the non-transitivist can deem λ to be *less true* than '$2 + 2 = 4$'. In the standard non-transitive approach, λ and '$2 + 2 = 4$' can be seen to diverge at the *pragmatic level*, in particular in their assertibility conditions. '$2 + 2 = 4$' is strictly and tolerantly assertible, whereas λ is tolerantly but not strictly assertible.

By reflecting on natural language data, Égré proposes a genuinely *semantic* analysis of 'is true' as a vague predicate. He analyses 'is true' as as an absolute gradable adjective. 'Is true' should be compatible with uses such as 'true in some sense', or 'true in some respect', but also faithful to absolute uses of 'is true' as 'true *simpliciter*' or 'perfectly true'. Égré defines a partial and a total meaning of 'is true', roughly corresponding to the semantic values of 'true in some respects' and 'true in all respects'. The liar sentence λ cannot be true in all respects, but can only be *half true*. Crucially, Égré's analysis reconciles the non-transitive approach to truth with its original analysis of vagueness as an essentially semantic phenomenon. This shows that the pragmatic machinery of assertibility conditions applied to the analysis of truth ascriptions may not be intrinsic to the non-transitive approach.

1.1.2 Truth and Logic

Semantics is one theoretical context in which the notion of truth has been employed. Traditional truth-theoretic deflationism denies that it is an important one. Truth is best conceived of as a quasi-logical device that supports correct reasoning by enhancing the expressive capabilities of our language. This *logical* notion of truth holds firm some suitable formulation of the T-schema: '"A" is true' is equivalent to *A*.[8] It investigates consistent (non-trivial) ways of characterizing truth by means of suitable quasi-logical principles. Such principles typically need to (i) support the expressive power afforded by the T-schema, and (ii) display theoretical virtues such as strength, unifying power, simplicity.

It's clear that traits (i) and (ii) of the logical notion of truth are not necessarily incompatible with semantic theorizing. Such an incompatibility was argued for by traditional deflationism on independent metaphysical grounds. The present volume features attempts to articulate a *compatibilist* approach to the relationships between semantic applications and the logical notion of truth.

Disquotation, Compositionality, and Reflection. A hallmark of truth-conditional semantics is compositionality. It's useful to restrict our attention to sentences: compositionality requires that the truth-value of a sentence supervenes on the semantic values of its parts (e.g., a conjunction is true if and only if both conjuncts are true). Lavinia Picollo and Thomas Schindler, in *Is Deflationism Compatible with Compositional and Tarskian Truth Theories?*, investigate to what extent compositional principles can be compatible with the logical notion of truth. They propose two desiderata for theories of the logical notion of truth. The first, *Functionality*, enforces a mimimal adequacy requirement for the logical truth predicate. Since the truth predicate is—among other things—a generalizing tool, it has to satisfy the (uniform) T-schema for the class of sentences one wants to generalize over. The second desideratum, *Relative Insubstantiality*, prescribes that the principles characterizing the logical truth predicate should be derivable from the instances of the T-schema and some additional non-truth-theoretic principles. But what are these additional principles? And in what sense are they compatible with the logical truth predicate? Following recent developments in formal theories of truth, Picollo and Schindler consider *proof-theoretic reflection principles* and, in particular, Uniform Reflection Principles. In these recent developments, truth theorists investigate motivations and consequences of combining disquotational truth and proof-theoretic reflection. They show that Uniform Reflection can be used to derive compositional principles from disquotational ones (Horsten and Leigh, 2017; Fischer et al., 2017). Such developments are carefully surveyed in Horsten and Zicchetti's *Truth, Reflection, and Commitment*. Horsten and Zicchetti's contribution is not limited to the interplay between disquotation and reflection, but discusses also the mathematics of reflection principles and some of their recent philosophical applications.

Riki Heck's *Disquotationalism and Compositionality* and Johannes Stern's *Belief, Truth, and Ways of Believing* cast doubts on the compatibility of disquotational and semantic truth. One core tenet of classical disquotationalism is that *A* and '"A" is true' are fully cognitively equivalent. Their equivalence is stricter than material and necessary equivalence. Heck and Stern both argue that the strong equivalence required by disquotationalism overgenerates.[9] In particular, they both present cases to doubt this equivalence between *A* and '"A" is true' in non-extensional contexts.

Stern's criticism focuses on the status of the equivalence between *A* and '"A" is true' in belief contexts. He provides and discusses evidence supporting the claim that believing and believing-true need to be differentiated at the semantic level: for instance, someone may believe that Goldbach's conjecture is true, without believing Goldbach's conjecture, because they lack relevant information concerning the representational status of 'Goldbach's conjecture'. However, Stern also provides a formal semantics that isolates a class of belief reports for which the equivalence envisaged by disquotationalists holds. When one is *aware of* appropriate facts concerning the syntactic representation of the belief under considerations, the disquotationalist's equivalence holds. We will come back to some aspects of Stern's formal model shortly.

Heck focuses on distinct kinds of overgeneration. The common theme of his criticism is the status of compositional principles in the disquotationalist's framework. First, compositional principles such as

(7) for all sentences A: 'A' is not true if and only if '$\neg A$' is true

should be taken to express—according to traditional disquotationalism—the trivial infinite conjunction of all the instances of the schema $\neg A \leftrightarrow \neg A$. However, and this is the first case of overgeneration, it's not clear why this infinite conjunction should not be expressed by

(8) for all sentences A: '$\neg A$ if and only if $\neg A$' is true.

(7) and (8) have very different logical properties. Therefore, they cannot both be taken to express—in the strong sense required by disquotationalism—the infinite conjunction of all instances of $\neg A \leftrightarrow \neg A$.

Heck's second overgeneration argument concerns directly the strategies employed by disquotationalists to recover compositionality from disquotation defended in Picollo and Schindler's contribution and surveyed by Horsten and Zicchetti. According to Heck, the same strategy that enables one to conclude (7) from the schema $\neg \mathrm{Tr}\ulcorner A \urcorner \leftrightarrow \mathrm{Tr}\ulcorner \neg A \urcorner$, also enables one to infer more dubious 'compositional' principles for intensional and hyperintensional connectives such as 'it's necessary that' and 'because'.

Compatibility and Contextualism. The purpose of the logical notion of truth consists in its expressive power. Above all, disquotationalists hold that the truth predicate is indispensable in blind ascriptions ('What Francis said on Sunday is true') and generalizations ('All theorems of Euclidean Geometry are true'). In turn, blind ascriptions and generalizations are necessary to the expression of agreement and disagreement (Field, 2008). Hartry Field forcefully argued that such an expressive role requires the truth predicate to be *transparent*: A and '"A" is true' should be intersubstitutable *salva veritate*. We have seen that transparency may be dubious in intensional and hyperintensional contexts, but disquotationalists maintain that it should at least be uncontroversial in extensional contexts.

In *The Expressive Power of Contextualist Truth*, Julien Murzi and Lorenzo Rossi put this last claim into question. They argue that transparency is not required to perform blind ascriptions and generalizations, and therefore to express agreement and disagreement. This leaves open whether weaker versions of the equivalence of *A* and '"A" is true' may be required for these purposes. In particular, one might still require the truth predicate to be *naïve*, in the sense that it satisfies the rule of inference 'from *A*, infer "'A' is true"', and its converse. Murzi and Rossi also argue that *naïveté* is not required by the logico-linguistic tasks considered by disquotationalists. They propose to employ a *contextualist* approach, and associated contextualist rules to model agreement and disagreement. In particular, they claim that the rules

(9) if '"A" is true' is inferred at context α, infer A at context α

(10) if $\neg A$ is inferred at context α, infer 'it's not the case that
"A" is true' at α

suffice to adequately model the cases of agreement and disagreement that arise in the disquotationalist literature.

Murzi and Rossi's proposal can be seen as an alternative way of achieving some form of compatibility of the logical notion of truth with semantic truth. Contextualist approaches to truth and paradox draw much of their motivation from ideas in philosophy of language and linguistics. A central task for the contextualist is to determine contexts in which specific truth ascriptions have definite semantic values. In particular, according to the contextualist, there are contexts (such as paradoxical ones) in which a given truth ascription is semantically defective because it fails to express a proposition, but it may express a proposition in other contexts. In such an alternative context the truth ascription will have a definite semantic value.

By arguing that this semantically loaded notion of truth can be reconciled with some of the tasks intrinsic to the disquotationalist's truth predicate, Murzi and Rossi create a promising bridge between the two concepts of truth analysed in the first part of the volume.

1.2 Unification: Formal Semantics

Whatever principles of truth one chooses, a core tenet of the Unified Approach investigated in this volume is that truth should naturally interact with modal and doxastic notions. This interaction should then be modelled by a satisfactory formal semantics. These desiderata immediately raise problems. The standard semantic framework for modal and doxastic notions is *possible worlds semantics*. The prominence of possible worlds semantics can be easily explained by some of its theoretical virtues: it's simple and widely adaptable to various domains, it fits strong pre-theoretic intuitions concerning alternative metaphysical and doxastic scenarios, and it has proven to be successful in modelling phenomena in linguistics, philosophy, and computer science. Possible worlds semantics is usually applied to languages in which modal and doxastic notions are formalized as sentential operators. As such, they apply to formulae that have a finite (well-founded) structure.[10] The satisfaction of a formula φ at a world is defined there by a straightforward induction on φ's finite syntactic structure. However, this strategy does not extend so easily to languages containing a truth predicate. Due to the possibility of self-referential constructions, objects to which the truth predicate applies may have an infinite syntactic structure. Therefore, one requires more involved strategies. The inductive strategy of fixed-point models in the style of Kripke (1975) is one example; another is the class of models obtained by revision-theoretic strategies from Gupta and Belnap (1993).

It's however possible to combine semantic constructions for (self-referential) truth predicates and possible worlds semantics. The basic idea is simple. For definiteness, let's focus on the interaction of truth and necessity predicates. One can think of a *frame F* as given by a collection of standard interpretations of our basic language $\mathcal{L}$ (without truth nor necessity)—the collection of *worlds*—together with some binary accessibility relation R on this collection. The interpretations of the truth and necessity predicates are given by a suitable function f that assigns (suitable) sets of sentences to each world in the frame. Given a world w, the intended role played by the evaluation function is to provide both the extension $f(w)$ of the truth predicate at w, and the extension

$$\bigcap_{wRv} f(v)$$

of the necessity predicate at w, that is, the intersection of the extension of the truth predicate at all worlds accessible from w.

To construct suitable evaluation functions, one can apply the strategies employed in the resolution of the liar paradox. For instance, as shown in Halbach and Welch (2009), one can generalize Kripke's fixed-point semantics and define a suitable evaluation function as the fixed point of a positive inductive definition that is performed simultaneously at each world.[11] This procedure will yield suitable evaluations for arbitrary frames. If one does not define the evaluation function via some positive inductive construction, one is not guaranteed that suitable evaluations will be found for arbitrary frames. However, as shown in Halbach et al. (2003), converse well-founded frames always admit evaluations.

In the framework just outlined, many interesting philosophical principles can be validated in full generality. For instance, models based on generalized fixed points will satisfy the factivity of necessity:

(11) $\forall x(\text{Sentence}(x) \wedge \text{Nec}(x) \rightarrow \text{True}(x))$.

Claims such as (11) deal with sentences—or whatever objects of truth and modality one plausibly assumes instead of sentences. They are *de dicto* assertions. Even in expressive frameworks such as the semantics just sketched, *de re* ascriptions are more difficult to treat. In the models considered by Halbach and Welch (2009), Stern (2015), and Stern's contribution to this volume, one has syntactic names for all objects in the domain of discourse. In those happy circumstances, a *de re* version of factivity can be readily formulated because the formal syntax is sufficiently rich to support quantifying into modal contexts:

(12) $\forall x, y(\mathrm{Sentence}(x(y)) \wedge \mathrm{Nec}(x(y)) \rightarrow \mathrm{True}(x(y)))$.

(12) expresses that if the predicate x is necessary of y, then it is also true of y.

The assumption that our language can name all objects, especially when dealing with general claims involving logical and metaphysical necessity, is clearly too strong: our language does not contain names for all possible objects. Volker Halbach, in *The Fourth Grade of Modal Involvement*, investigates languages in which *de re* ascription can be formalized in full generality. The fundamental idea is to generalize Tarski's (1956) treatment of satisfaction to modal predicates. *De re* factivity can now be captured by means of *binary* predicates for necessity and truth:

(13) $\forall x, y(\mathrm{Formula}(x) \wedge \mathrm{Nec}(x, y) \rightarrow \mathrm{True}(x, y))$.

(13) expresses that if the object y possesses the property expressed by x necessarily, then x is also true of y. Halbach outlines different alternatives for a possible worlds semantics for *de re* necessity, and discusses some metaphysical applications of the framework. In (13), the object y may stand for an arbitrary (finite) sequence of objects in the domain of discourse. One crucial aspect of the semantics sketched by Halbach is the strict link between object-linguistic sequences figuring in *de re* ascriptions and metalinguistic sequences of variable assignments. Halbach provides ingenious examples to show that, once the full power of the Fourth Grade of Modal Involvement is available, mathematical assumptions of the metatheory are reflected in one's theorizing in the object-language: if one carelessly adopts the standard metatheory of quantified modal logic, strong theses such as actualism may follow.

In *Belief, Truth, and Ways of Believing*, Johannes Stern considers the interaction of truth and belief. We have already seen how Stern's contribution can be seen as a reaction to some shortcomings of the disquotationalist's position. However, his work also contains a detailed formal semantics for a language with a truth predicate, an awareness predicate, and a belief operator. Stern's model is based on the generalized fixed-point strategy outlined above, but extends it with a suitable interpretation of the belief operator. The main ideas are to extend the notion of evaluation and the associated notion of satisfaction from worlds to *pairs of worlds* (w, v), and to parametrize the accessibility relation to the agents' belief representation. The first component of the pair (w, v) is the one that provides the location in the possible worlds structure at which the basic vocabulary, the truth predicate, and the awareness predicate are evaluated; the second component of the pair fixes the location of the agent's awareness set at a world, which is the set to which the accessibility relation is parametrized. For instance, to evaluate a sentence of the form

(14) *S* believes that *S* believes that 'snow is white' is true

at (w, v), one looks at all of *S*'s doxastic alternatives z with respect to w and at *S*'s awareness set at v to evaluate '*S* believes that "snow is white" is true'. However, to evaluate "'snow is white" is true' at all doxastic alternatives u to z, one will look at the pair (u, v): the location v of the awareness set is kept fixed. In Stern's analysis, this yields a correct way of capturing the semantic resources that the agent has at her disposal in evaluating truth ascriptions: someone may in fact believe a proposition p, but may not believe the truth ascription $\mathrm{Tr}\ulcorner \varphi_p \urcorner$, because she may not be aware that φ_p expresses the proposition p.

The strategies employed to provide a semantics for expressive languages allowing for self-reference can also shed light on some puzzles of rationality. In her contribution *Indeterminate Truth and Credences* Catrin Campbell-Moore focuses on so-called self-undermining credences. A credence is called self-undermining if, in case a rational agent were to adopt it, they would immediately be compelled to revise their attitude. It turns out that in certain scenarios all *precise* credences will be self-undermining, just like all assignments of a classical truth-value to the liar sentence would result in the opposite value. But one may then ask which attitude a rational agent ought to adopt, and it is at this point that the strategies for addressing the semantic and intensional paradoxes considered above come into play.

After developing a revision jump in the style of Gupta and Belnap (1993) for definite assignmements of truth values, Campbell-Moore defines a revision jump for sets of precise credences to model the relevant cases of self-undermining credences: the revision process in such cases does not reach a fixed point. To find non-self-undermining credences, Campbell-Moore then moves to imprecise credences (sets of precise credal functions) and develops a novel, generalized fixed-point semantics based on Kripke's supervaluational jump. Similarly to what happens to the liar sentence in Kripkean fixed points, self-undermining credences are indeterminate, in the sense that no definite credence can be associated with them in this framework. Campbell-Moore's semantics for self-undermining credences provides a rich an interesting model of belief. The framework also stresses that self-reference is a phenomenon that needs to be taken seriously, as it arises in various and sometimes unexpected contexts.

1.3 Unification and Higher-Order Resources

As argued above, the Unified Approach draws from the analysis of quantification in natural language. It appears that in natural language there is only one fundamental kind of quantification, that is quantification into singular-term position. We have seen that it is possible to quantify over nominalized properties and relations in suitable frameworks inspired by the Unified Approach, such as Halbach's framework for *de re* necessity. Some authors forcefully argued that first-order quantification is not sufficient to theorize at the level of generality required by some areas of philosophy. Instead one should adopt a new, irreducible form of quantification over higher-order entities (Williamson, 2003).

Advocates of the higher-order perspective typically motivate their position by noting that a faithful semantics for absolutely general theses such as 'Everything is self-identical' cannot be given in standard (first-order) set theory. Instead, one should provide models that aren't sets, but *sui generis* entities of an irreducibly higher-order type. This irreducibility is typically motivated by higher-order versions of Cantor's theorem, stating that such higher-order entities are also of a different size than first-order ones.

A reaction from the point of view of the Unified Approach may focus on the inherent expressive limitations of the higher-order perspective. Entities are syntactically distinguished in a hierarchy of types: such types are metatheoretic objects that, despite their essential role in the categorization of the type-theorists' ontology, are not part of the domain of higher-order quantifiers. Moreover, the type-theorist would hope that such commitment to types could be reduced to the minimum: namely, that one could resort to a coherent rationale to endorse only arbitrarily many finite types, as in Simple Type Theory. However, there are arguments suggesting that such a rationale may be difficult to find. Linnebo and Rayo (2012) argue that the type-theorist is bound to countenance proper class-many types.

James Studd's *Infinite Types and the Principle of Union* reacts to Linnebo and Rayo's argument. Their argument rests crucially on the claim that the type-theorist does not have at her disposal a principled way to stop ascending the hierarchy of types at any limit ordinal. Studd carefully examines this claim, the Principle of Union, and finds it wanting. In particular, he argues that the union language, say $\bigcup_{n\in\omega}\mathcal{L}_n$ with n a finite type, has a special semantic status. One could provide a semantics for each of its sublanguages, but this is not yet sufficient ground to countenance a language of type $\omega + 1$ without assuming the Principle of Union itself. This is no more justified than granting, without an Axiom of Infinity, a *set* of all finite sets on the basis of the acceptance of all finite sets.

We have seen that higher-order resources, if not strictly regimented in a hierarchy of types, allow for great expressive power that lead to paradoxical phenomena akin to the semantic and modal paradoxes. Andrew Bacon's *Opacity and Paradox* investigates the so-called Prior's Paradox for Thought, which is one instance of the paradoxes of indirect discourse already encountered in this introduction. Prior's Paradox is actually a theorem of propositionally quantified intensional logic and involves an arbitrary sentential operator, which can be interpreted as 'Mary thought at time *t* that'. Prior's Theorem then states that, if Mary thought at time *t* that all she thought at time *t* was false, then Mary thought a true and a false thing at *t*. The puzzling feature of this theorem comes from the fact that it's plausible to think that Mary thought only one thing at *t*, namely that all she thought at time *t* was false. But in that case, she couldn't have thought a true and a false thing.

Bacon proposes a solution to Prior's paradox based on the restriction of Universal Instantiation in opaque contexts such as 'it is thought by Mary at *t* that'. The idea is familiar from the cluster of Frege's puzzles and the puzzles of belief discussed in Stern's contribution: given two semantically equivalent (possibly identical) expressions, this equivalence may fail to be reflected in opaque contexts. He also provides models that invalidate Prior's Theorem (and Universal Instantiation), but preserve many other desirable principles of propositional quantification. Interestingly, such a framework enables him to suggest a resolution of the liar paradox akin to classical gap-theories proposed by Feferman (1991) and Maudlin (2004). Bacon defines a truth predicate on the basis of the (opaque) context 'means that'. The liar sentence *says that* it is not true, but *there's nothing it says*. So, the liar sentence conveys information that can be acted upon, and even known. However, due to the opacity phenomenon, such information does not pick out semantic content: there's no proposition corresponding to the liar.

Acknowledgments

Johannes Stern's research was supported by the European Research Council (ERC) Starting Grant *Truth and Semantics* (TRUST, Grant no. 803684).

Notes

1. Throughout the Introduction we adopt a wide understanding of 'first-order quantification' that, e.g., includes the use of generalized quantifiers. The

relevant contrast is between (first-order) quantification, which, assuming some form of Montague Grammar, binds argument positions of type *e*, and higher-order quantification binding argument positions of higher types. See Heim and Kratzer (1998) or Gamut (1991) for an introduction to Montague Grammar.

2. This may be an oversimplification. Perhaps one needs to accommodate a rich ontology of various attitudinal objects à la Moltmann (2017a, b).
3. That is, unless these expressions are of flexible type.
4. This does not imply that modality and modal notions need to be generally conceived to be of the same logical/grammatical category as truth. It is perfectly acceptable and perhaps desirable to, following Kratzer (1981), analyse modality, i.e., the modal aspect of sentences of natural language via a modal operator but truth as a predicate. The point is that in sentences like 'Everything necessary is true', 'is necessary' needs to be conceived of as a predicate if truth is, i.e., 'is necessary' needs to be construed as a phrase of type $\langle e, t \rangle$.
5. See Stern (2015) for an overview of the various paradoxes.
6. See specifically Stern and Fischer (2015) for a discussion of this point.
7. The paradoxes of indirect discourse have been re-discovered by Brandenburger and Keisler (2006) and are known under the label the Brandenburger-Keisler paradox in the literature on epistemic game theory.
8. How to understand this equivalence is a non-trivial matter. We will come back to this point below.
9. Issues of overgeneration have been raised famously by Gupta (1993) and, more recently, by Nicolai (2020).
10. Formulae are generally modelled after finite, well-founded trees or sequences that unravel their syntactic structure.
11. For the frame $(\{w\}, \langle w, w \rangle)$, a suitable evaluation f consists simply in a function that assigns a Kripke-fixed point to w to both the truth of necessity predicate. The interpretations of the two predicates will diverge in more complex frames, but it will always yield fixed points.

References

Asher, N. (1990). Intentional paradoxes and an inductive theory of propositional quantification. In Parikh, R., editor, *Theoretical Aspects of Reasoning about Knowledge*. Morgan Kaufmann.

Beall, J. C. (2009). *Spandrels of Truth*. Oxford University Press.

Brandenburger, A. and Keisler, H. J. (2006). An impossibility theorem on beliefs in games. *Studia Logica*, 84: 211–240. Special Issue Ways of Worlds II. V. F. Hendricks and S. A. Pedersen, Editors.

Bull, R. A. (1969). Modal logic with propositional quantification. *The Journal of Symbolic Logic*, 34(2): 257–263.

Cobreros, P., Égré, P., Ripley, D., and van Rooij, R. (2013). Reaching transparent truth. *Mind*, 122(488): 841–866.

Cross, C. B. (2001). A theorem concerning syntactical treatments of nonidealized belief. *Synthese*, 129: 335–341.

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

Field, H. (2008). *Saving Truth from Paradox*. Oxford University Press.

Fine, K. (1970). Propositional quantifiers in modal logic. *Theoria*, 36: 336–346.

Fischer, M., Nicolai, C., and Horsten, L. (2017). Iterated reflection over full disquotational truth. *Journal of Logic and Computation*, 27(8): 2631–2651.

Gamut, L. (1991). *Logic, Language, and Meaning, Volume 2: Intensional Logic and Intensional Grammar*. The University of Chicago Press.

Grover, D. (1992). *A Prosentential Theory of Truth*. Princeton University Press.

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

Gupta, A. (1993). A critique of deflationism. *Philosophical Topics*, 21(1): 57–81.

Gupta, A. and Belnap, N. (1993). *The Revision Theory of Truth*. MIT Press.

Halbach, V. (2006). How not to state T-sentences. *Analysis*, 66(4): 276–280.

Halbach, V. (2008). On a side effect of solving Fitch's paradox by typing knowledge. *Analysis*, 68(298): 114–120.

Halbach, V., Leitgeb, H., and Welch, P. (2003). Possible-worlds semantics for modal notions conceived as predicates. *Journal of Philosophical Logic*, 32 (2): 179–223.

Halbach, V. and Welch, P. (2009). Necessities and necessary truths: A prolegomenon to the use of modal logic in the analysis of intensional notions. *Mind*, 118 (469): 71–100.

Heim, I. and Kratzer, A. (1998). *Semantics in Generative Grammar*. Blackwell Publishing.

Horsten, L. and Leigh, G. E. (2017). Truth is simple. *Mind*, 126(501): 195–232.

Horsten, L. and Leitgeb, H. (2001). No future. *Journal of Philosophical Logic*, 30(3): 259–265.

Kratzer, A. (1981). The notional category of modality. In Eikmeyer, H.-J. and Rieser, H., editors, *Words, Worlds, and Contexts: New Approaches to World Semantics*, pages 38–74. de Gruyter.

Kripke, S. (1975). Outline of a theory of truth. *Journal of Philosophy*, 72: 690–712.

Linnebo, O. and Rayo, A. (2012). Hierarchies ontological and ideological. *Mind*, 121(482): 269–308.

Maudlin, T. (2004). *Truth and Paradox: Solving the Riddles*. Oxford University Press.

Moltmann, F. (2017a). A truthmaker theory for modals. *Philosophical Issues*, forthcoming.

Moltmann, F. (2017b). *Truth Predicates, Truth Bearer, and their Variants*. http://friederike-moltmann.com/uploads/TruthPredicates-July-publ-2017.pdf.

Montague, R. (1963). Syntactical treatments of modality, with corollaries on reflexion principles and finite axiomatizability. *Acta Philosophica Fennica*, 16: 153–167.

Mulligan, K. (2010). The truth predicate vs. the truth connective. On taking connectives seriously. *Dialectica*, 64(4): 565–584.

Myhill, J. (1960). Some remarks on the notion of proof. *Journal of Philosophy*, 57(14): 461–471.

Nicolai, C. (2020). Fix, express, quantify. Disquotation after its logic. *Mind*. Online version.

Priest, G. (2006). *In Contradiction: A Study of the Transconsistent*. Oxford University Press.

Prior, A. (1961). On a family of paradoxes. *Notre Dame Journal of Formal Logic*, 2: 16–32.

Prior, A. (1971). *Objects of Thought*. Clarendon Press.

Quine, W. V. O. (1970). *Philosophy of Logic*. Harvard University Press.

Ramsey, F. P. (1927). Facts and propositions. *Proceedings of the Aristotelian Society*, 7: 153–170.

Ramsey, F. P. (1929). The nature of truth. In Rescher, N. and Majer, U., editor, *On Truth: Original Manuscript Materials (1927–1929) from the Ramsey Collection at the University of Pittsburgh*. Kluwer Academic Publishers. Collection published in 1991.

Ripley, D. (2012). Conservatively extending classical logic with transparent truth. *Review of Symbolic Logic*, 5(2): 354–378.

Stern, J. (2015). *Toward Predicate Approaches to Modality*, volume 44. Springer.

Stern, J. and Fischer, M. (2015). Paradoxes of interaction? *Journal of Philosophical Logic*, 44(3): 287–308.

Tarski, A. (1956). Der Wahrheitsbegriff in den formalisierten Sprachen. In *Logic, Semantics, Metamathematics*, pages 152–278. Clarendon Press.

Thomason, R. (1980). A note on syntactical treatments of modality. *Synthese*, 44: 391–395.

Williamson, T. (2003). Everything. *Philosophical Perspectives*, 17(1): 415–465.

2 Half-Truths and the Liar

Paul Égré

2.1 Introduction

The aim of this chapter is to explore some philosophical implications of the fact that the adjective "true" is *gradable* in natural language, in particular concerning the strict-tolerant account of the Liar paradox (see Cobreros et al., 2013). In the literature on truth, "true" is sometimes viewed as a vague predicate, to mean that it admits borderline cases. The idea that "true" is vague features, famously, in Russell's account of vagueness (Russell, 1923), and it is present in various accounts of the Liar paradox (viz. McGee, 1990), including the strict-tolerant account (Cobreros et al., 2015b). The observation that "true" and "false" are gradable adjectives looms large in the tradition of fuzzy logic (viz. Zadeh, 1975; Weatherson, 2005; Smith, 2008), but it was met with skepticism (see Haack, 1980), and it is generally given little philosophical importance (a recent exception is Henderson, 2021). It is, however, of particular interest in order to bridge linguistic and logical considerations about truth.

To say that "true" is gradable means that it supports comparative morphology ("truer", "less true") and also adverbial modification by the intensifier "very" ("very true") (viz. Sapir, 1944). The adjective "true" in English moreover appears to be a special kind of gradable adjective, namely an *absolute* gradable adjective in the sense of Unger (1975) (a feature briefly emphasized in my Égré, 2019, which spurred this chapter). This view, endorsed by Unger himself, means that "true" patterns as a maximum-scale adjective (Kennedy and McNally, 2005; Kennedy, 2007). In particular, unlike relative adjectives such as "tall" or "rich", but like other absolute adjectives such as "flat" or "full", "true" supports modification by the adverbs "completely" and "perfectly" (compare "completely true", "perfectly true", with the ungrammatical "*completely tall", "*perfectly tall").

Independently, Jared Henderson makes the same observation in a paper in which the gradability of truth is discussed, and in which Henderson questions deflationism about truth (see Henderson, 2021). Like

Henderson, in this chapter I elaborate on whether and in what sense "true" is gradable, but with a different objective in mind, mostly to focus on the specific issue of whether the Liar sentence can be considered a "half-truth", namely a sentence that, albeit true to some extent, may be considered less than perfectly true.

The view that the Liar is a half-truth is controversial. For a dialetheist, the Liar is both true and false, but to say this is to consider that the sentence is both perfectly true and perfectly false. The sentence just fails to be *only true* or *only false* (Priest, 2019). Similarly, the strict-tolerant account of truth implies that both the Liar and its negation are true, but the account does not view the Liar as "less true" than more ordinary truths such as "London is in England" or "2 + 2 = 4", but only as having a different assertability status (Cobreros et al., 2013). In what follows, I propose to reconsider this view. Basically, while the strict-tolerant account was initially conceived for vague predicates, its extension to the semantic paradoxes assumed that assertion, but not truth, comes in different degrees. My main argument in what follows is that we get a better explanation for the unified treatment of paradoxes of vagueness and truth offered by ST if we consider that "true" is a special kind of vague predicate indeed, namely an absolute gradable adjective exemplifying a systematic ambiguity between a total and partial interpretation.

2.2 The Strict-Tolerant Account

2.2.1 Vagueness

The strict-tolerant account was originally put forward as an account of the semantics and pragmatics of vague predicates in natural language (Cobreros et al., 2012; van Rooij, 2012). The leading idea is that every vague predicate can be used either in a looser sense (its "tolerant" meaning) or in a stronger sense (its "strict" meaning). As originally laid out, the account is neutral on whether strict and tolerant meaning can vary from speaker to speaker, or even from context to context, though it is compatible with both hypotheses. Basically, a model of vague language is a model of how the idealized, crisp meaning of a predicate (its "classical" meaning) can be tightened up or loosened up relative to relations of indifference that may be speaker-dependent.

An example can help to illustrate this: for a vague predicate like "tall", a speaker may not be able to reliably discriminate between heights below a certain threshold. For instance, assuming "tall" were to denote heights above a given standard (say above 185cm), and that heights that differ by less than 2cm cannot be reliably discriminated, the tolerant meaning of "tall" includes all heights that are indiscriminable from the standard (above 183cm), whereas the strict meaning of "tall" includes all and only heights that are discriminable from the standard (above

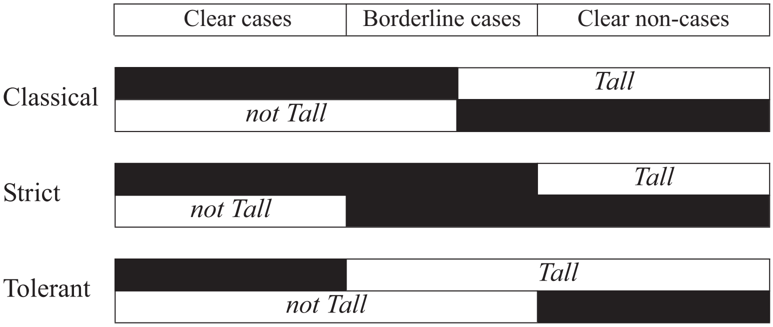


Figure 2.1 Classical, strict and tolerant denotations for "tall"

187cm). The semantics for negation is defined in such a way that in order for an object to belong to the tolerant denotation of the negation of a predicate, it must fail to belong to its strict denotation, and conversely.

The resulting relation between classical, strict and tolerant denotations is depicted in Figure 2.1.[1] As shown in the figure, the strict denotation of "tall" is the narrowest, the tolerant denotation the widest and the classical denotation falls in between. Moreover, the strict denotation of "tall" underlaps with the strict denotation of "not tall", whereas the tolerant denotation of "tall" overlaps with the tolerant denotation of "not tall". This gives an account of borderline status for vague predicates: borderline cases are those that fall in the overlap between tolerant denotations for a predicate and its negation, or dually, that fall in the underlap between strict denotations. The account moreover vindicates soritical reasoning without contradiction: by definition, an argument is ST-valid iff its conclusion holds tolerantly when its premises hold strictly. In agreement with that definition, note that at every point of a sorites series, if *a* is in the strict extension of *P*, then any indiscriminable object *b* from *a* is in the tolerant extension of *P*.

2.2.2 *Truth*

Instead of working with classical denotations, another way of presenting the strict-tolerant account is directly in terms of a trivalent Kleene semantics (see Ripley, 2012, Cobreros et al., 2015b). When *a* is a borderline case of the predicate *P*, *Pa* is assigned the value $\frac{1}{2}$. By definition a sentence *A* holds tolerantly if *A* takes a value 1 or $\frac{1}{2}$, and strictly it is takes the value 1. Using the strong Kleene rules, this implies that for a borderline case *a* of *P*, the sentence $Pa \wedge \neg Pa$ holds tolerantly, whereas *Pa* and $\neg Pa$ each fails to hold strictly.

This machinery was used by Dave Ripley to extend the account to a theory of transparent truth (Ripley, 2012), such that $True\langle A\rangle$ and

A take the same value and are intersubstitutable in every context (see then Cobreros et al., 2013 for philosophical implications). On the strict-tolerant account, the Liar sentence is forced to take the value $\frac{1}{2}$, but so is its negation. This means that the Liar and its negation both hold tolerantly. Given the strict-tolerant definition of validity, this implies that both the Liar and its negation are ST-valid, although without contradiction, due to the relation of ST-validity being nontransitive, an aspect also recruited in the treatment of the sorites paradox.

2.2.3 Handling True *as Vague*

As stressed in Cobreros et al. (2015b), the strict-tolerant account handles the sorites paradox and the Liar paradox in parallel and very similar ways. Philosophically, however, the treatment of vagueness and the treatment of truth my coauthors and I gave differ rather significantly. For vagueness, the theory admits that all vague predicates can be interpreted either strictly or tolerantly. For truth, however, the view put forward in Cobreros et al. (2013, 2015b) is that truth is a unitary notion, only governed by Tarski's disquotation principle, but that assertion comes in degrees. This difference was emphasized as follows:

> As far as we can see, then, there are at least two ways to understand the status paradoxical sentences have on a [strict-tolerant]-based theory like the one we have advanced here. Both ways take paradoxical sentences to fall in between strict and tolerant, but one way takes the distinction between strict and tolerant to be a pragmatic distinction, and the other to be a distinction in meaning.
>
> (Cobreros et al., 2013, p. 857)

That is, for vague predicates, strict and tolerant qualify a difference in meaning; for truth, they qualify a speech act difference. In hindsight, I think this way of contrasting the two accounts is not quite adequate. First of all, the strict-tolerant account of vague predicates accepts that predicates can be used and interpreted in two different ways, the tolerant and the strict way. This means that the selection between strict and tolerant meaning is necessarily a pragmatic matter, even if the distinction between tolerant and strict pertains to the semantic content of predicates.[2]

Secondly, an objection that I think can be made to the assertion-oriented theory of the Liar is that the parallel between the two accounts appears somewhat like a coincidence. Alternatively, a way to maintain a closer link between the two accounts is to handle the predicate "true" as a vague predicate, and to admit that "true" can be interpreted either tolerantly or strictly. The main objection to that approach in Cobreros et al. (2013) was that the account of the Liar would immediately be subject to revenge problems. But as acknowledged in Cobreros et al. (2015b), even an assertion-based account of the Liar cannot be completely immune to revenge paradoxes.

Thirdly, a more fundamental argument to handle "true" as a vague predicate, susceptible to different interpretations, can be located precisely in the fact that "true" is a gradable adjective. Linguistic gradability is generally well-correlated with vagueness, here understood as the admission of borderline cases (a gradable expressions typically implies a non-degenerate degree scale, along which different positions can compete for the assignment of a boundary). Whether gradability is necessary or even sufficient for vagueness is disputed, however. Raffman (2014) argues, against necessity, that a word like "medium" is vague, but that "medium" is not gradable linguistically, due to the apparent oddity of expressions like "very medium" or "more medium". Raffman remains agnostic on whether gradability is sufficient for vagueness. Upon closer inspection, however, "medium" appears to be linguistically gradable. Expressions like "very medium" ("of very medium quality") or "more medium" ("a tiny bit more medium than I wanted") are attested, despite being infrequent. In fact, "medium" appears to pattern as an *absolute* adjective ("perfectly medium" is attested). For absolute adjectives, as we shall see in greater detail, modification by "very" or "more" can be marked due to absolute adjectives being semantically attached to a specific degree on the scale. But the admission of those modifiers is a reliable indicator of vagueness. "True" is very similar to "medium" in that regard: expressions like "more true" or "very true" are likely parasitic on an absolute meaning, but they indicate vagueness.

The strict-tolerant account of vagueness used the gradable adjective "tall" as a paradigmatic example of vague predicate. Importantly, "tall" is a *relative* gradable adjective, whereas "true", as we will review in greater detail in the next section, is an *absolute* gradable adjective. Hence we may expect to find differences. For example, whereas "tall" is sorites-susceptible, "true" is not obviously sorites-susceptible. If we follow Kennedy's (2007) account of absolute gradable adjectives, Kennedy thinks of absolute gradable adjectives indeed as not being sorites-susceptible, due to the fact that the meaning of those adjectives is determined by a maximum or minimum standard along some degree scale. A different account can be found in Burnett's typology of gradable adjectives (Burnett, 2017). According to Burnett, whether a gradable adjective is sorites-susceptible or not depends on the granularity of the scale that is contextually relevant, and also on the directionality of the sorites sequence. For her, however, even absolute adjectives are sorites-susceptible, although the sorites-susceptibility of absolute adjectives is indeed more constrained than for relative adjectives.

In Cobreros et al. (2015b), our account did follow McGee in admitting that "true" is vague, but mostly for lack of determinate rules in the language to adjudicate the status of the Liar sentence. Here, and

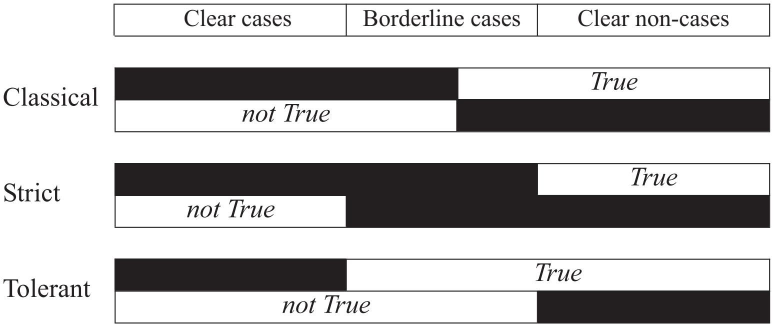


Figure 2.2 Classical, strict and tolerant denotations for "true"

by analogy with the strict-tolerant account of borderline cases for "tall", the view is that for "true" we ought to get a picture much like the previous one, where "true" can be interpreted tolerantly or strictly, and where the Liar stands as a borderline case (see Figure 2.2). As already mentioned, however, "true" and "tall" have different scale structures (not represented in either figure so far), and so we need to say more about the gradability of "true" proper.

2.3 *True* as Absolute Gradable

Typological work on adjectives indicates that they fall into three broad classes (viz. Kennedy, 2007 and Burnett, 2017): *non-scalar* (like "hexagonal", "pregnant", "prime"), *relative* gradable (like "tall", "rich") and *absolute* gradable (like "full", "dangerous"). Basically, non-scalar adjectives have a sharp and context-invariant meaning. The underlying scale for them is binary and degenerate. Relative adjectives are strongly context-sensitive: the scale is typically richer and their meaning is not attached to a fixed degree on it. Absolute adjectives on the other hand are only weakly context-sensitive: their meaning is relative to a maximum or minimum degree on the scale, though pragmatically they pattern like relative adjectives to some extent.[3]

The question this section seeks to clarify is in which of those three classes "true" falls. Like Henderson (2021), I think the evidence weighs clearly in favor of Unger's observation that "true" is an absolute adjective. This view is not uncontroversial. Soon after Unger introduced the distinction, Haack (1980) examined whether true is gradable, but her conclusion was that it is neither a relative nor an absolute gradable adjective; instead she put it in a different category (of "achievement predicates").

My view is that Haack was right to deny that "true" patterns anything like "tall", namely as a relative adjective, but that she was wrong to

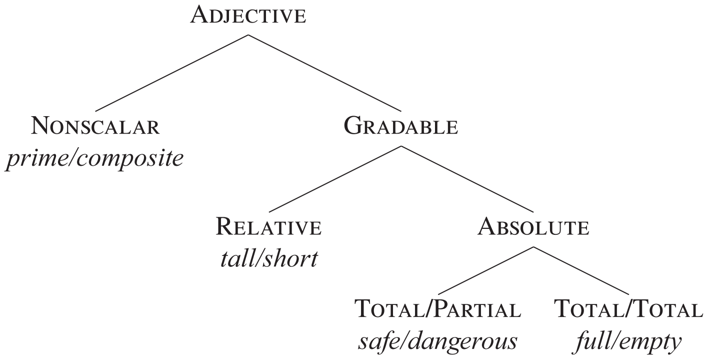


Figure 2.3 Adjectival typology with antonym pairs

reject Unger's description of it as an absolute adjective. What I show in this section, following the extended typology of absolute adjectives proposed by Cruse (1980) and Yoon (1996), is that while "true" and "false" typically pattern as *total* absolute adjectives, they can also pattern as *partial* absolute adjectives (see Figure 2.3). This behavior supports the idea that "true" can be attached a strict as well as a tolerant meaning.

2.3.1 Is True *Non-Scalar?*

"True" is a gradable adjective by the admission of the comparative form ("more true", "less true"), and of the intensifier "very" ("very true"). Moreover, it is an absolute adjective given modification by the adverb "completely" ("completely true"). Haack's (1980) main argument against "true" being an absolute gradable adjective is that "true" does not appear to support modification by "extremely", unlike the absolute adjective "flat". However, Haack concedes that "true" admits modification by "completely", and it is now widely considered that "completely" is a more reliable indicator of absoluteness than "extremely". Indeed, "extremely" conveys an element of surprise, and can moreover be found combined with relative adjectives, as in "extremely tall".[4] Another argument Haack uses against "true" being absolute is that "very flat" generally implicates "not (perfectly) flat", whereas "very true" need not implicate "not (perfectly) true". But that alleged asymmetry too is fragile, since for both adjectives "very + ADJ" appears to trigger the same cancellable implicature to the negation of "perfectly + ADJ".

Despite Haack's arguments being inconclusive, one should ask whether "true" is not like "pregnant" or "hexagonal", namely fundamentally a non-scalar adjective, but which can be coerced in some contexts into a gradable adjective. Indeed, comparatives like "more pregnant" and "more hexagonal" are attested, and likewise intensified expressions such as "very pregnant" and "very hexagonal" (Burnett, 2017). As argued in (Burnett, 2017), however, those are better described as cases of coercion of non-scalar adjectives into gradable adjectives (for instance, "very pregnant" means "very advanced in pregnancy", or "showing very clear signs of pregnancy", but not "*being more fertilized"). Couldn't "more true" or "very true" be cases of coercion in the same way?

"True" has "false" as a lexical antonym, whereas neither "pregnant" nor "hexagonal" appear to have one, compatibly with "true" and "false" denoting opposite regions on a degree scale. For "pregnant" and "hexagonal", an explicit negation seems needed to denote the opposite ("non-pregnant", "non-hexagonal"). Unfortunately, the existence of a natural antonym seems neither necessary nor sufficient to conclude that an expression is gradable. Against necessity, color adjectives like "red" and "blue" are clearly gradable, but do not have obvious lexical antonyms. Moreover, whereas "non-hexagonal" could in principle refer to a variety of shapes, "non-pregnant" appears to denote a specific state, one for which a simple lexeme might exist in some languages. Against sufficiency, a counter-example is the pair "prime-composite" for numbers. "Prime" and "composite" are antonyms, but they remain non-gradable, for expressions like "more prime" or "very prime" have only very marginal uses, indicative of coercion.

A better argument to conclude that "true" and "false" are scalar terms rather than non-scalar is the admission of adverbial modifiers ruled out by "prime" or "pregnant". It is fine to say of a sentence that it is "not completely true", but expressions like "not completely prime" or "not completely pregnant" sound inappropriate. "Hexagonal" is more tricky, because a figure may be described as "not completely hexagonal" or "not perfectly hexagonal", in the same way in which a proposition may be described as "not completely true" or "not perfectly true".

Another test concerns modification by the adverbial "to some extent". A number cannot be "prime to some extent" to mean that it is partly prime and partly non-prime. Likewise, "pregnant to some extent" or "hexagonal to some extent" sound degraded, whereas "true to some extent" and "false to some extent" seem to be very common expressions, not parasitic on a coerced use of "true" or "false". Similar judgments are found with the adverb "partly": a statement can be partly true, but a number cannot be partly prime, a figure partly hexagonal, or a woman partly pregnant.

As admitted by Burnett (2017), the frontier between non-scalar and absolute adjectives is thin. Hence, even if "true" and "false" were better described as cases of non-scalar adjectives coerced into gradable adjectives, to describe them as absolute gradable is to admit an element of contextual invariance shared with non-scalar adjectives.

2.3.2 True *Is Total and Partial*

Granting that "true" and "false" are absolute gradable adjectives, the next question concerns their associated scale structure.

In the literature on gradable adjectives, two categories of absolute adjectives have been distinguished, following the typology of Cruse (1980) and Yoon (1996), reviewed in Rotstein and Winter (2004), namely *total-total* pairs (T-T pairs) and *total-partial* pairs (T-P pairs). A pair like "full-empty" is of type T-T, since both adjectives are modifiable by "completely", in agreement with the fact that both are maximum-scale adjectives. A pair like "safe-dangerous" is of type T-P, for although the first adjective is modifiable by "completely", the second is not (completely safe/*dangerous). The antonym, on the other hand, is modifiable by the adverb "slightly" ("slightly dangerous"), unlike the first (*slightly safe), and unlike relative adjectives (*slightly tall).[5] We must therefore ask if the pair "true-false" is more similar to "full-empty", or to "safe-dangerous".

Henderson describes "true" and "false" as closed-scale adjectives, each denoting opposite endpoints on the scale. In support of Henderson's (2021) description, comparison suggests that the pair "true-false" comes closer to a T-T pair than to a T-P pair: like "true", "false" is indeed modifiable by "completely" (viz. "those allegations are completely false"). A difficulty for that analysis is that some occurrences of "slightly false" appear felicitous, and similarly for "slightly true". Admittedly, the expression "slightly false" can mean the same thing as "slightly insincere", when applied to a specific kind of behavior (for instance a smile).[6] "Slightly false" seems less common when predicated of an utterance, but is used to convey that the utterance is misleading or not completely accurate ("what John told you is slightly false"). Beside "false", another antonym for "true" is "untrue". In English, the expression "slightly untrue" is attested in that same sense, but "completely untrue" is used as well. For "untrue" as well as for "inexact", modification by "slightly" and by "completely" thus appears permissible. "Slightly true" is much less common than "completely true", but it is found in surveys that put it on a scale with "not at all true" and "completely true" at opposite ends.[7] "Partly true" is quite common, on the other hand, but it suggests that "true" can behave like "filled", or indeed like "full" when modified with "partly".

A possibility is that "true" and "false" show a pattern of systematic ambiguity. Basically, whereas "true" and "false" fundamentally denote endpoints of a top and bottom-closed scale, the complementary region may be described as a region of partial falsity or partial truth. The pattern comports with the strict-tolerant account of vague predicates, and can be depicted graphically (see Figure 2.4): the strict interpretation of "true" is the top of the scale, and the tolerant interpretation of the interval ruling out the bottom point.

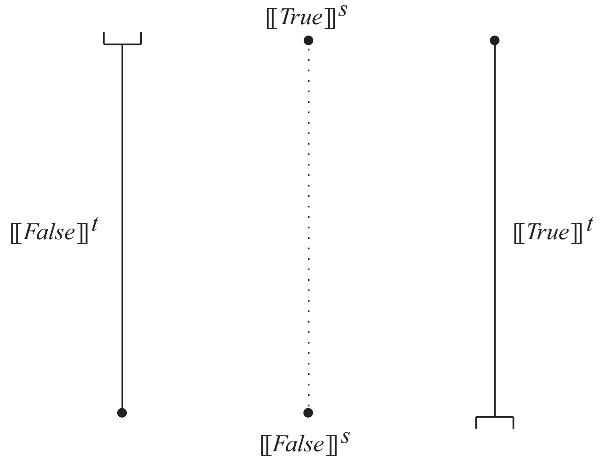


Figure 2.4 Strict representation and tolerant representation for scalar "true" and "false"

The figure presents an analogy with the opposition between "empty" and "full". For a container, "empty" strictly speaking denotes the minimum, zero degree of being filled, and "full" the maximum degree. However, "half-empty" implies that the container is not completely empty, and likewise "half-full" that it is not completely full. Both expressions imply that the container is full to some extent, and empty to some extent. So one may understand the strict meaning of "empty" to mean that *all* of the container contains nothing (is empty *simplirciter*), whereas one may associate a tolerant meaning to "empty" to mean that *some of its content* contains nothing. For "full", the strict meaning should be that all of its content is filled, and the tolerant meaning that some of its content is filled.

Another case of absolute antonyms whose interpretation appears to vary more systematically between total and partial is given by Jeremy Zehr and is the pair "transparent/opaque".[8] A glass is *completely opaque* when no light goes through it. But it can also be described as *slightly opaque* when the light does not perfectly go through. Conversely, a glass is *completely transparent* when the light goes through with no significant loss of luminosity. But when it is not perfectly opaque it may be described as *slightly transparent*, to mean that light goes through to some extent.

Either way the meaning of "opaque" and "transparent" is relative to opposite endpoints on the associate scale. But more clearly than for "full' and "empty", the meaning of "opaque" can either denote the point of maximum opacity, or the region that excludes the point of maximum transparency. Zehr points out that when a container is described as "full" simpliciter, we typically understand that it is filled near its maximum, not that it is "not perfectly empty". By contrast, when we hear that a glass is "opaque", we can hesitate between understanding "not perfectly transparent" or "perfectly opaque". For "full-empty", the default interpretation is therefore strict-strict, whereas for "opaque-transparent" the interpretation can vary more freely between strict and tolerant.

What about "true" and "false"? When we hear or say that an utterance is "true", is the default interpretation that the utterance is perfectly true, or can it be that it is true to some extent? The answer to this question is not as obvious as it might seem. Quite plausibly the strict interpretation of "true" is the default interpretation, for we typically say "not false" to indicate that an utterance is not perfectly true. But "not false" does convey that the utterance is true to some extent, or contains some element of truth.

2.4 True in Some Respects

Further evidence can be given in favor of the alternation of "true" and "false" between a strict and a tolerant meaning. One concerns modification of either adjective by "almost". The other concerns quantification over respects.

2.4.1 Almost

Beside "completely" and "slightly", another modifier studied by Rotstein and Winter (2004) in relation to total adjectives is "almost", which can be found attached to both "true" and "false". Here are two occurrences found on the web:[9]

(1) As to "I can start or finish when I want". False but almost true. Yes, technically, I can choose to start my day at 5am, and finish at 2pm.... In reality, well ... no: I don't start or finish when I want.[10]

(2) Keep this in mind as far as the size [of the organizer goes]: They claim it is 27 inches in height. Technically, this is true but almost false. The height for my closet between the shelf and the roof is 30 inches. For whatever reason they make you put the circular attachments at the top.[11]

As mentioned by Rotstein and Winter, "almost" can be found with partial adjectives ("almost dangerous"), but according to them for a pair of type T-P, "almost P" seems generally to entail "not T" (viz. "almost dangerous" to "not safe"). In the previous examples, "almost false" does not entail "not true", and likewise "almost true" does not entail “not false”. In that regard, both examples would indicate that each adjective is a total adjective.

However, “true but almost false” in the previous example conveys that the sentence, despite being true, is not completely or perfectly true, and similarly for “false but almost true”, which suggests that the sentence is not perfectly false. This is similar to saying that an activity is “safe but almost dangerous”, to convey that it is not perfectly safe. It sounds odd, by contrast, to assert: “this activity is perfectly safe, but almost dangerous”.

2.4.2 Respects

The previous two examples indicate that “true” and “false” also pass some of the tests put forward by Sassoon (2012) in relation to multidimensionality. Sassoon points out that the (absolute) gradable adjective “healthy” allows modification by “with respect to”, as in “healthy with respect to blood pressure, but not with respect to glucose level”. In the same way, sentence (1) is presented as false with respect to “reality”, although as true with respect to what is technically or legally permissible. Likewise, that the organizer is 27 inches in height is presented as true without respect to the attachments, though as false with respect to them. That “true” and “false” pass Sassoon’s test can be independently confirmed by explicit modification by “in every/some respect”, as in:

(3) The current belief that fact finders must come with a blank slate is false in every respect save one.[12]

(4) Although this was written 14 years ago it is still true in every respect.[13]

What does it mean to say that a sentence is “true in some but not all respects” then? I believe that a sentence will be called “true in some respect” or “false in some respect” only if that sentence contains some element of semantic or pragmatic indeterminacy, such that it could express different propositions depending on how the indeterminacy is resolved. “True in some respect” thus appears to mean the same thing as “true in some sense”, where the sense in question is a sense of the sentence of which “true” is predicated.

A case to see this concerns sentence (3), “fact finders must come with a blank slate”. Allen (1993) declares the sentence “false in every respect save one”. The author makes clear that he is in fact talking about the belief that “jurors must have no knowledge about the case”. He goes on to write (emphasis mine):

> The belief is false *in the technical sense* that … only knowledge that would qualify a person as a witness disqualifies the person as a juror. The conventional belief about the necessary ignorance of jurors is

> false *in a deeper sense*. Juridical decisions makers come to trial with a vast storehouse of knowledge, beliefs and modes of reasoning that are necessary to permit communication to occur simply and efficiently.
> (p. 1157)

That is, the sentence "jurors must have no knowledge about the case" fails to specify the type of knowledge in question. One precisification of this sentence is therefore "jurors must have no *witness* knowledge about the case". That precisification would make the sentence true. But the precisification "jurors must have no knowledge *whatsoever* about the case" would make the sentence false.

The same analysis can be given of sentence (2), "the organizer is 27 inches tall". One precisification of this sentence is "the organizer *without the attachments* is 27 inches tall", which is true. Another is "the organizer *with the attachments* is 27 inches tall", which is false. The sentence may be judged "technically true", to mean "true in the respect that excludes the attachments", but "practically false", to mean "not true in the respect that includes the attachments". A similar analysis can be produced for sentence (1) above, "I can finish work whenever I want", depending on whether the latter is modified by "in practice" or by "in principle".

2.5 More True

A (vague) sentence is true tolerantly if there is at least one context where the proposition expressed by the sentence is true simpliciter. This is what the expressions "true to some extent", "true in some sense" and "true in some respect" all convey. Yet truth simpliciter must be a non-scalar, yes-or-no matter. This does not prevent "true" from supporting the comparative form, but this apparent paradox concerns all absolute gradable adjectives.[14]

For example, to say that a surface is flat is to have a standard of precision in mind, such that the degree of flatness of the surface falls within that standard. Let the meaning of "flat" be "with no bumps deeper than *n* mm from the same horizontal line". Given a value for *n*, either a surface is flat, or it is not. As Lewis (1979) writes:

> on no delineation of the correlative vagueness of "flatter" and "flat" is it true that something is flatter than something that is flat.
> (p. 353)

It is still possible to say that a surface is flatter than another, but on Lewis's analysis to say so is to change the standard of precision. Formally, a surface is flatter than another if it remains flat simpliciter in all delineations of the meaning of "flat" in which the former is flat, and more. Equivalently, if it remains flat for smaller values of *n*.

The same account can be applied to the comparative "more true". Adapting Lewis's supervaluationist analysis, a sentence may be considered "more true" than another if it is true simpliciter in all delineations in which the former is true simpliciter, and more.[15] For instance, consider two ways of measuring organizers, one with and one without their circular attachments at the top. Suppose organizer A comes out taller than 27 inches irrespective of the measurement method, whereas organizer B comes out not taller than 27 inches without the attachment, but taller with the attachments. Can we say that the sentence "organizer A is taller than 27 inches" is *more true* than "organizer B is taller than 27 inches" in this case?

Doubtless the sentence is *more determinately true* than the second, or *more clearly true* than the second. One can find several occurrences of "more true" in the expression "even more true" to back up such uses. An example is:

(5) Mr. Speaker, what was true on day one is even more true now and that is that Canadians no longer trust the Conservatives to protect the environment.[16]

The proportion of Canadians used to make that vague generalization is left unspecified by the speaker. One interpretation of this sentence is that the predicate "no longer trust the Conservatives to protect the environment" is true of more Canadians than it was before. As a result, every precisification (proportion of Canadians) that made the sentence true initially makes it true in the posterior context, but not conversely.

If "more true" can be used to mean "more clearly true", however, one may object that the underlying gradable property is "clearly true" instead of "true". But this need not be the case. Also common is the expression "more true" followed by "to say" to compare the truth status of two vague statements, as in the following quote from F. W. Robertson:

(6) It is more true to say that our opinions depend upon our lives and habits, than to say that our lives and habits depend on our opinions.

The two generic sentences "opinions depend upon our lives and habits" and "our lives and habits depend on opinions" are both assertable in this context, and both may even be considered to have the same degree of clarity. Robertson appears to mean that the extent to which opinions depend on lives and habits is greater than the extent to which lives and habits depend on opinions. By way of consequence: the extent to which it is true that opinions depend upon lives and habits is greater than the extent to which it is true that lives and habits depend on opinions.

The Lewisian analysis of the vagueness of gradable adjectives therefore carries over to "true". Lewis (1979) himself admitted that truth comes in different extents when he wrote:

> we treat a sentence more or less as if it is simply true, if it is true over a large enough part of the range of delineations of its vagueness. (For short: if it is *true enough*.)
>
> (p. 352)

Lewis's view here is fully compatible with the recognition that "true" admits a tolerant meaning beside its strict meaning. Granted, what Lewis calls "simply true" may be understood to mean "true over the whole range of delineations of its vagueness", in agreement with the supervaluationist perspective. But Lewis's view is more flexible. "Simply true" can mean "tolerantly true", in agreement with the subvaluationist perspective, in cases in which truth over a proper part of the range still counts as enough.[17] Half-truths are precisely cases of this kind.

2.6 Half-Truths and the Liar

2.6.1 Half-Truths

In ordinary language, we talk of "half-truths" for utterances that fail to tell the whole truth (in violation of the Gricean Maxim of Quantity), or for utterances that are true in some respects, but not in every respect (in violation of the Gricean Maxim of Quality). Both properties often co-occur, since using a sentence that is true in only one out of several senses can be a way to mislead or to hold back information (Engel, 2016; Égré and Icard, 2018). Hence what defines a half-truth is fundamentally the fact for a vague sentence to be true in only some of its relevant senses.[18] Moreover, a half-truth generally fails to be simply true because the senses in which it is false cannot be ignored.

A good example of a half-truth concerns the biblical response made by Abraham to Abimelech concerning his wife Sarah (Augustine, ca. 420 and Stokke, 2018 for a recent discussion). Abraham declares to the king "Sarah is my sister". In one sense she is because Sarah is his half-sister (his father's daughter), but in another she is not, because Sarah is not Abraham's full sister (his mother's daughter). Another example is Bill Clinton's statement "I never had sexual relations with Monica Lewinsky". In one sense this utterance is true, namely the contrived and overly specific sense submitted by the judge who interrogated Clinton (see Tiersma, 2004; Égré and Icard, 2018). But in another sense this utterance is not true, namely the more mundane sense in which receiving oral sex is sufficient to have a sexual relation.

Both examples cohere with the discussion given earlier of the gradability of "true". Neither Abraham's utterance nor Bill Clinton's utterance are *perfectly true*, because for each of them there is a respect or sense in which the utterance is also not true.[19] Importantly the utterances are not completely false either (they are not strictly false). Moreover, they are not sentences that would be ineligible for a judgment of truth of falsity, as happens in cases of presupposition failure. Instead they happen to be true in one sense, and false in another.

2.6.2 The Liar

The Liar is the sentence saying of itself that it is not true. The proposal here is that the Liar is a half-truth in the same way in which the previous examples are half-truths. In one sense, the Liar is true, in fact the tolerant sense of "true"; but in another it is not, in effect the strict sense of "true". This view agrees with the dialetheist analysis, on which the Liar is both true and false, except that "true" and "false" need to be qualified.

One way of substantiating this idea this is to define the semantics of "true" and "false" along the lines of the degree-theoretic analysis of scalar expressions proposed by Kennedy (2007). For a relative adjective like "tall", the degree semantics says that "tall" denotes the property $\lambda x.f_{Tall,M}(x) \geq \theta_{Tall,M}$, where $f_{Tall,M}$ is a function mapping individuals to degrees (for instance heights), and where $\theta_{Tall,M}$ is a threshold value for "tall". Importantly, for "tall" the scale of heights is open-ended, and the threshold value is strongly context-sensitive, meaning that it can vary with comparison class and language-user.

To capture the alternation between strict and tolerant meaning for all gradable adjectives *Adj*, it is natural to introduce a convex interval of admissible thresholds $\mathcal{I}_{Adj,M}$, to represent the range of values for which the adjective applies to some but not all extent. For "tall", for instance, $\mathcal{I}_{Tall,M}$ may select all values between 183cm and 187cm included in M, but this could be a different interval in a different model. This serves to mimic the strict-tolerant semantics outlined in Section 2.2.1.[20] The tolerant extension of "tall" in context M, $[\![Tall]\!]^{t,M}$, can be defined as $\{d \in M|\ \exists\theta \in \mathcal{I}_{Tall,M} : f_{Tall,M}(d) \geq \theta\}$ (tall to some extent), and its strict extension $[\![Tall]\!]^{s,M}$ as $\{d \in M|\ \forall\theta \in \mathcal{I}_{Tall,M} : f_{Tall,M}(d) \geq \theta\}$ (tall to all extent). One may then apply the recursive machinery of Cobreros et al. (2012), so that for first-order sentences A, $[\![\neg A]\!]^{t,M} = 1 - [\![A]\!]^{s,M}$, $[\![\neg A]\!]^{s,M} = 1 - [\![A]\!]^{t,M}$; $[\![A \wedge B]\!]^{t/s,M} = \min\{[\![A]\!]^{t/s,M}, [\![B]\!]^{t/s,M}\}$, and $[\![\forall x A]\!]^{t/s,M} = \inf\{[\![A[d/x]\!]^{t/s,M}); d \in M\}$.[21]

The same semantics can be adapted to absolute adjectives. For those, the degree scale will need to be top-closed or bottom-closed, and moreover the interval of admissible thresholds must be defined relative to a context-invariant value. For example, with the absolute adjective "transparent", the scale will represent degrees of transparency varying between 0 (the minimum) and 1 (the maximum), and $f_{Transparent,M}$ is a function mapping objects to their transparency degree on the scale. The total meaning of "transparent" is attached to the degree 1, and the partial meaning is given by the interval excluding 0.

In the case of "true", we let $f_{True,M}(x)$ be an interpretation function mapping sentences to their truth value in some closed set $\mathcal{V}$. Given our observations, $\mathcal{V}$ needs to be at least three-valued, and we may assume $\mathcal{V}$ to correspond to the real interval [0, 1]. To get the meaning of "true", we start from the property $\lambda x.f_{True,M}(x) \geq \theta$.[22] In effect, this says that for a sentence to be true, the sentence needs to be be true enough, namely to exceed a given threshold. In order to get the absolute meaning of "true" (see Figure 2.4), $\mathcal{I}_{True,M}$ should include the top value 1 in every model M, and exclude the value 0 in every model M. We may therefore pick $\mathcal{I}_{True,M} = \mathcal{V}\backslash\{0\}$, to represent the maximum range of degrees for which a sentence is true to some extent.[23] The partial meaning of "true", $[\![True]\!]^{t,M}$ can be defined as $\{A|\ \exists\theta \in \mathcal{I}_{True,M} : f_{True,M}(A) \geq \theta\}$, and its total meaning $[\![True]\!]^{s,M}$ as $\{A|\ \forall\theta \in \mathcal{I}_{True,M} : f_{True,M}(A) \geq \theta\}$. The interpretation of complex sentences is defined recursively as before.

Obviously, the strictly true sentences are those that take the value 1, and the tolerantly true sentences those that do not take the value 0. By saying of itself that it is not true, the Liar can only be tolerantly true, and it must express of itself that it is tolerantly not true. This holds if we assume that $f_{True,M}(\neg A) = 1 - f_{True,M}(A)$ (value-inverting negation), and that for every θ, $f_{True,M}(True\langle A\rangle) \geq \theta$ iff $f_{True,M}(A) \geq \theta$ (identity of truth). Under those assumptions, the Liar sentence λ, being equivalent to $\neg True\langle\lambda\rangle$, can only take the value 0.5.[24] Moreover, the semantics of "false", being the antonym of "true", can be defined in a dual way, from the property $\lambda x.f_{False,M}(x) \leq \theta$, and assuming $\mathcal{I}_{False,M} = \mathcal{V}\backslash\{1\}$. Tolerant and strict meaning for "false" can be defined correspondingly, making the Liar sentence tolerantly true and tolerantly false.

2.7 Comparisons

The present account fundamentally agrees with dialetheism, and also with the strict-tolerant analysis of truth and the Liar presented in Cobreros et al. (2013). It differs in several ways, however.

Regarding dialetheism, Priest (2019) recently contends that although the Liar isn't "just true" or "just false", it should not be viewed as less than perfectly true for that matter. Priest moreover doubts that truth is vague, and also that talk of degrees is relevant in this and similar cases.[25] I agree with Priest on the latter point if, like Haack, what he means is that "true" is unlike "tall" and other relative gradable adjectives. But by treating "true" as an absolute gradable adjective we are led to a different picture of "true" from the one entertained by the dialetheist, one on which "true" shows a pattern of systematic ambiguity between strict and tolerant sense. The fact that "true" applies to the Liar in the tolerant sense also answers the objection about vagueness: the Liar is a borderline case of truth in the sense in which Cobreros et al. (2012) characterize borderline cases of vague predicates more generally (see Figure 2.2).

Regarding the version of the strict-tolerant account of truth given in Cobreros et al. (2013), the difference concerns not the framework or the logic, but the semantics-pragmatics division attached to it. First of all, here we have a proper basis to distinguish strict truth and tolerant truth: the gradable predicate "true" is susceptible of distinct interpretations, which are grammatically visible; it is no longer a coincidence if "true" can be treated along the same lines as other vague predicates. Secondly, the difference between strict and tolerant interpretation is not just at the level of assertion; it supervenes on a difference in meaning.

One way to probe this interpretation is to ask if failure of perfect truth for the Liar is visible in strict-tolerant logic. On the naive theory of truth, the truth predicate *True* satisfies the Tarskian disquotation principle whereby $True\langle A\rangle \leftrightarrow A$ for every sentence A. In relation to logical validity, moreover, logical truth classically satisfies the suppressibility property whereby $\vdash True\langle C\rangle$ implies: if C, $A \vdash B$ then $A \vdash B$. That property may be called Fregean, by reference to what Frege calls the True.[26] In terms of inference, it reflects the idea that a logical truth is uninformative, that it makes no difference when taken as a premise.

Now, the strict-tolerant account defines validity as follows: when all premises are strictly true, the conclusion is tolerantly true. It follows from the semantic status of the Liar sentence λ that on the *ST*+ theory of validity applied to sentences involving the truth predicate, $\models^{ST+} True\langle\lambda\rangle \leftrightarrow \lambda$, and that $\models^{ST+} True\langle\lambda\rangle$. In other words, the Liar, like other sentences, satisfies the disquotation principle, but moreover it is a logical truth (an *ST*+-validity). However, $\lambda, A \models^{ST+} B$ fails to imply $A \models^{ST+} B$ in general. This is so because $\lambda \models^{ST+} \bot$, but $\not\models^{ST+} \bot$. In other words, the Liar is not suppressible, despite being a logical truth.

Of course, one may object that once we recognize "true" to be ambiguous between tolerant and strict, one should no longer expect logical truth to remain a simple notion, governed by universal principles. Maybe suppressibility is only a property of a subclass of logical truths, those that are strictly true. This is right, but this makes the point only more vivid: the Liar fails to be suppressible on the ST account precisely because it fails to ever be strictly or perfectly true. These considerations, possibly, may be used to foster Henderson's claim that the absolute character of "true" is at odds with a purely deflationary view of truth (Henderson, 2021). If perfect truth, that is strict truth, is fundamentally constraining the meaning of "true", then it may be argued that there is more at bottom of the naive conception of truth than the transparency of truth.

2.8 Conclusion

The main point of this chapter is that "true" should be considered in the same way in which other vague predicates have been considered on the strict-tolerant account, namely as showing a systematic difference in meaning between a weak and a strong interpretation. Pace Haack, and in agreement with Henderson, I have argued that "true" and "false" are absolute gradable adjectives with a closed-scale structure, but moreover that they pattern as total or as partial adjectives depending on the context. When a sentence declares its own lack of truth, as the Liar does, it is therefore natural to question whether the sentence declares its lack of partial truth, or its lack of total truth, and whether it is to be evaluated according to the latter or the former notion of truth.

From this I have drawn the consequence that the Liar sentence is a half-truth. One virtue of this classification is that it groups the Liar together with vague sentences that are recognizably truth-evaluable, but not fully true on account of their borderline status. Like the Liar such sentences are equivocal, they admit different interpretations and they are only partly true. In this regard, recognizing that "true" is vague comports with the role of the truth predicate in most theories, whether based on the idea of transparency or on the idea that truth ought to supervene on non-semantic facts. If truth is to apply not just to precise or context-insensitive sentences (like "2 + 2 = 4"), but also to vague and context-sensitive sentences (like "John is tall"), then we can expect truth to reflect the vagueness of the sentence to which it is attributed.

Several issues remain open. Consider again a vague sentence like "John is tall". We may introduce an operator "tolerantly" in the language, such that: "John is tall" is true tolerantly iff "John is tolerantly tall" is true (strictly). With such an operator, however, we can easily create a Strengthened Liar (see Cobreros et al., 2015b). One insight into this problem is that once we take seriously the idea that "true" is vague, we should take seriously the idea that "true" can be higher-order vague, too. The mechanisms used to handle first-order vagueness would need to be iterated in order to deal with higher-order vagueness and revenge phenomena in a satisfactory way.

Also, I have assumed that the degree scale for "true" and "false" is given. But for gradable adjectives at large, degrees are more plausibly a construction from the non-scalar meaning of those expressions (see Klein, 1980; Burnett, 2017). Similarly, the present account does not use the mechanism of indifference relations and classical extensions originally in play in Cobreros et al. (2012)'s account of vague predicates. It is also connected to the supervaluationist notion of precisification and to the idea of quantification over contexts in an indirect way only. Future work should seek to establish whether even tighter parallels can be built with each of these approaches.

Acknowledgments

I thank audiences in Paris (Pascal Engel's EHESS seminar), Salzburg (workshop "Semantic paradox, context, and generality"), Munich (Lorenzo Rossi's seminar on truth), Bochum (conference "LP after 40 years") and Beijing (Tsinghua University's online colloquium, April 2020) for comments and questions. Thanks also to the program van Gogh 42589 PM and the participants of the 2019 workshop on "Fiction, Lies, and Pretense", organized with Emar Maier and Louis Rouillé. I am particularly grateful to the Editors of this volume, Carlo Nicolai and Johannes Stern, for their comments on the first version of this chapter, and to Susanne Bobzien, Colin Caret, Pascal Engel, Diego Feinmann, Jared Henderson, Andreas Kapsner, Fenrong Liu, Emar Maier, Julien Murzi, Hitoshi Omori, Graham Priest, David Ripley, Lorenzo Rossi, Heinrich Wansing and Jeremy Zehr for helpful feedback and criticisms. I dedicate this chapter to the memory of Serge Déjean (1950–2020): Κρῆτες άεὶ ψεῦσται · σὺ δ’ ού θάνες, έσσὶ γὰρ αίεί.

Notes

1. Thanks to Jeremy Zehr who created this figure.
2. See Cobreros et al. (2015a) and Égré and Zehr (2018) for more on mechanisms of selection between strict and tolerant meanings.
3. See in particular Burnett (2017, p. 89) for a summary. Note that the term "non-gradable" could be used instead of "non-scalar", since the latter is meant to refer to whether an expression comes with a non-degenerate degree scale attached to it (that is, with a scale not reduced to two values).
4. Besides, native speakers report to me that "extremely true" is acceptable in some contexts.
5. Incidentally, the pair "flat/bumpy" constitutes a pair of T-P absolute adjectives (viz. *completely flat/slightly bumpy*). When Haack asked whether truth is flat or bumpy, she inadvertently assumed "bumpy" to be a relative adjective. What she meant to ask in her paper was actually whether "true" is more like "flat" or more like "tall". The point of this section may be summarized provocatively as follows: *truth is flat and bumpy*.
6. An example is provided by the Cambridge Dictionary, which paraphrases "suavely" as "in a way that is polite, pleasant, and usually attractive, but often slightly false".
7. See for instance Redman (2003, p. 325).
8. J. Zehr, private communication.
9. Sentence (1) is translated from French, sentence (2) comes from a US customer review. Both appear written by native speakers of their own language.
10. See http://www.mariegraindesel.fr/?p=3460. The original French text is: "Quant au "je ferais les horaires que je veux". Faux mais presque vrai. Oui, techniquement, je peux choisir de commencer ma journée à 5h du matin : et à 14h, fini … Mais dans la réalité euh … ben non : je ne fais pas les horaires que je veux."
11. From https://www.amazon.com/Park-Purse-Organizer-no-tools-assembly/product-reviews/B06WVBXPWW.
12. From Allen (1993).

[13] From https://academic.oup.com/jof/article-abstract/46/4/282/4707959.

[14] See Burnett (2017) for an extended discussion of this problem. Here and elsewhere, I restrict myself to "true" and "more true" applied to sentences. I set aside issues of truth and comparative truth pertaining to theories (sets of sentences), which introduce further complications.

[15] Weatherson (2005) accepts that "true" is gradable, but he dismisses the supervaluationist view, though without much argument, essentially because according to him, "it assumes that we can independently define what is an admissible precisification, and this seems impossible". As argued by Lewis (1979), however, the vagueness of what to count as an admissible precisification does not imply that the notion is not operative.

[16] From https://www.ourcommons.ca/DocumentViewer/en/39-1/house/sitting-139/hansard.

[17] See Hyde (1997).

[18] The comparaison between "half-true" and "half-full" suggests a potential disanalogy. When we talk of a half-full glass, we have in mind a precise extent to which the glass needs to be filled, namely half of its capacity. When we talk of a "half-truth", do we necessarily mean that the sentence is true in exactly half of its precisifications, or does it suffice that it be true in just some of them? The problem is that the class of respects may be open-ended. "True" might be closer to "healthy" in this regard: it is less natural to declare someone "half-healthy" than to declare a glass "half-full" because quantifying the respects in which one can be healthy is more difficult than quantifying the extent to which a container can be filled. In the Abraham case and in the Clinton case, there turns out to be only two main precisifications of the predicates "sister" and "having sexual relations" in relation to either utterance, but this feature is probably inessential. The fact that the expression "half-true" is so common may indicate that the respects that make a sentence true roughly balance out the respects that make it not-true, and a single relevant respect for truth can sometimes suffice to offset multiple respects for untruth.

[19] Interestingly, Bill Clinton was also questioned by journalists as follows: "If she [Monica Lewinsky] told someone that she had a sexual affair with you beginning in November of 1995, would that be a lie?". His response was: "It's certainly not the truth. It would not be the truth". By thus answering, Clinton avoids committing to the full falsity of the sentence expressed (he only implicates it), while denying that the sentence in question is perfectly true.

[20] See Égré (2019). There, I distinguish quantification over respects and over degrees. What follows here is a simplified account where the mapping from respects to degrees is taken for granted, and where the degree scale is given (see the Conclusion).

[21] As in Cobreros et al. (2012), this presentation assumes for simplicity that all objects of the domain have a name in the language.

[22] More precisely, and using the notation of Heim and Kratzer (1998), we may define the basic meaning of "true" to be: $\lambda\varphi : \varphi$ is a sentence. $f_{True,M}(\varphi) \geq \theta$. This means that when φ is not a sentence, f would return a value $\sharp$ separate from other values in $\mathcal{V}$, to represent presupposition failure.

[23] One may narrow this interval, letting the strict extensions of "true" and "false" denote upsets and downsets excluding .5 and including 1 and 0, respectively. See Cobreros et al. (2019).

[24] See Rossi (2019) for a proof using equation systems.

[25] Priest (2019): "it is not clear that the truth predicate is a vague predicate".

26. See Chemla and Egré (2019) for more on the connection between suppressibility and the property Chemla and I call polarization for semantic values. This notion of polarization can be linked to Frege's conception of truth values as specific objects.

References

Allen, R. J. (1993). Expertise and the Daubert decision. *Journal of Criminal Law and Criminology*, 84: 1157.

Augustine (ca. 420). Contra Mendacium. *Corpus Scriptorum Ecclesiasticorum Latinorum (CSEL)*, Vol. 41, ed. Joseph Zycha, Vienna, F. Tempsky, 1900.

Burnett, H. (2017). *Gradability in Natural Language: Logical and Grammatical Foundations*, volume 7. Oxford University Press.

Chemla, E. and Égré, P. (2019). Suszko's problem: Mixed consequence and compositionality. *The Review of Symbolic Logic*, 12(4): 736–767.

Cobreros, P., Égré, P., Ripley, D., and van Rooij, R. (2012). Tolerant, classical, strict. *The Journal of Philosophical Logic*, 41(2): 347–385.

Cobreros, P., Égré, P., Ripley, D., and van Rooij, R. (2013). Reaching transparent truth. *Mind*, 122(488): 841–866.

Cobreros, P., Égré, P., Ripley, D., and van Rooij, R. (2015a). Pragmatic interpretations of vague expressions: Strongest meaning and nonmonotonic consequence. *Journal of Philosophical Logic*, 44(4): 375–393.

Cobreros, P., Égré, P., Ripley, D., and van Rooij, R. (2015b). Vagueness, truth and permissive consequence. In T. Achouriotti, H. Galinon, and J. Martínez, editor, *Unifying the Philosophy of Truth*, pages 409–430. Springer.

Cobreros, P., Égré, P., Ripley, D., and van Rooij, R. (2019). Tolerance and Degrees of Truth. Unpublished manuscript.

Cruse, A. (1980). Antonyms and gradable complementaries. In *Perspektiven der lexikalischen Semantik: Beiträge zum Wuppertaler Semantikkolloquium vom 2–3 Dec. 1977*, pages 14–25.

Égré, P. (2019). Respects for contradictions. In Baskent, C. and Ferguson, T., editors, *Graham Priest on Dialetheism and Paraconsistency*, pages 39–57. Springer.

Égré, P. and Icard, B. (2018). Lying and vagueness. In Maibauer, J., editor, *The Oxford Handbook of Lying*, pages 354–369. Oxford University Press.

Égré, P. and Zehr, J. (2018). Are gaps preferred to gluts? A closer look at borderline contradictions. In Castroviejo, E., McNally, L., and Sassoon, G., editors, *The Semantics of Gradability, Vagueness, and Scale Structure*, pages 25–58. Springer.

Engel, P. (2016). Demi-vérités et demi-mensonges. In Wieworka, M., editor, *Mensonges et vérités*, pages 15–29. Sciences Humaines Editions.

Haack, S. (1980). Is truth flat or bumpy? In *Deviant Logic, Fuzzy Logic*, pages 243–258. The University of Chicago Press.

Heim, I. and Kratzer, A. (1998). *Semantics in Generative Grammar*, volume 1185. Blackwell Oxford.

Henderson, J. (2021). Truth and Gradability. *Journal of Philosophical Logic*. https://doi.org/10.1007/s10992-020-09584-3

Hyde, D. (1997). From heaps and gaps to heaps of gluts. *Mind*, 106(424): 641–660.

Kennedy, C. (2007). Vagueness and grammar: The semantics of relative and absolute gradable adjectives. *Linguistics and Philosophy*, 30(1): 1–45.

Kennedy, C. and McNally, L. (2005). Scale structure, degree modification, and the semantics of gradable predicates. *Language*, 81(2): 345–381.

Klein, E. (1980). A semantics for positive and comparative adjectives. *Linguistics and Philosophy*, 4(1): 1–45.

Lewis, D. (1979). Scorekeeping in a language game. *Journal of Philosophical Logic*, 8: 339–359. Reprinted in D. Lewis, *Philosophical Papers*, vol. 1.

McGee, V. (1990). *Truth, Vagueness, and Paradox: An Essay on the Logic of Truth*. Hackett Publishing.

Priest, G. (2019). Respectfully yours. In Baskent, C. and Ferguson, T., editors, *Graham Priest on Dialetheism and Paraconsistency*. Springer. Chapter 27, Section 4, Reply to P. Égré.

Raffman, D. (2014). *Unruly Words: A study of Vague Language*. Oxford University Press.

Redman, B. K. (2003). *Measurement Tools in Patient Education*. Springer Publishing Company.

Ripley, D. (2012). Conservatively extending classical logic with transparent truth. *The Review of Symbolic Logic*, 5(2): 354–378.

Rossi, L. (2019). A unified theory of truth and paradox. *The Review of Symbolic Logic*, 12(2): 209–254.

Rotstein, C. and Winter, Y. (2004). Total adjectives vs. partial adjectives: Scale structure and higher-order modifiers. *Natural Language Semantics*, 12(3): 259–288.

Russell, B. (1923). Vagueness. *The Australasian Journal of Psychology and Philosophy*, 1(2): 84–92.

Sapir, E. (1944). Grading, a study in semantics. *Philosophy of Science*, 11(2): 93–116.

Sassoon, G. W. (2012). A typology of multidimensional adjectives. *Journal of Semantics*, 30(3): 335–380.

Smith, N. J. J. (2008). *Vagueness and Degrees of Truth*. Oxford University Press.

Stokke, A. (2018). *Lying and Insincerity*. Oxford University Press.

Tiersma, P. (2004). Did Clinton lie: Defining "sexual relations". *Chicago-Kent Law Review*, 79(3).

Unger, P. (1975). *Ignorance: A Case for Scepticism*. Oxford University Press, USA.

van Rooij, R. (2012). Vagueness, tolerance, and non-transitive entailment. In Cintrula, P., Fermüller, C., Godo, L., and Hájek, P., editors, *Understanding Vagueness: Logical, Philosophical and Linguistic Perspectives*, pages 205–222. College Publications.

Weatherson, B. (2005). True, truer, truest. *Philosophical Studies*, 123(1–2): 47–70.

Yoon, Y. (1996). Total and partial predicates and the weak and strong interpretations. *Natural Language Semantics*, 4(3): 217–236.

Zadeh, L. A. (1975). Fuzzy logic and approximate reasoning. *Synthese*, 30(3–4): 407–428.

3 Is Deflationism Compatible With Compositional and Tarskian Truth Theories?

Lavinia Picollo and Thomas Schindler

3.1 Introduction

For a number of reasons, deflationists about truth favour a formal treatment of the notion. What requirements deflationary formal theories of truth must satisfy is, thus, an important issue for deflationism. It is widely believed that compositional and Tarskian theories convey substantial concepts of truth or are otherwise unacceptable for the deflationist. Call this claim the 'incompatibility thesis'. Since compositional and Tarskian theories are often seen as superior to purely disquotational theories, the incompatibility thesis, if true, would provide support for substantial theories of truth over their deflationary rivals. Assessing whether the arguments for the incompatibility thesis are correct is therefore of great philosophical importance.

Here is the plan of the chapter. After some preliminaries (Section 3.2), we will rehearse six arguments for the incompatibility thesis from the literature (Section 3.3). We contend that most of these arguments issue from an overly narrow understanding of what role formal theories of truth are supposed to play. In Section 3.4, we introduce an important but often overlooked distinction between theories that are intended for a *descriptive* purpose (roughly, a theory that provides a faithful account of the basic usage of 'true') and those that are intended for a *logical* purpose (roughly, a theory that characterises the correctness of inferences involving 'true').

The notion of a logical purpose raises the question what the role of 'true' exactly consists in, and what truth principles are needed to carry it out. Drawing on earlier work (Picollo and Schindler, 2018a), we suggest (Section 3.5) that this role is best understood as enabling us to mimic sentential and predicate quantification within a first-order framework, and extract a criterion of functionality from that. However, not any theory that allows the truth predicate to fulfil its function might be acceptable to a deflationist: among other things, such a theory must not convey a substantial notion of truth. What this is supposed to mean is of course a controversial issue. We will not be able to provide an absolute criterion of substantiality, though we will propose (Section 3.6) a relative one: under certain circumstances, adding certain truth-theoretic principles to a deflationary theory will not inflate the notion of truth. In Section 3.7 we will defend this criterion against a popular objection.

In concluding this chapter (Section 3.8), we will survey a variety of formal truth theories and assess them in light of our criteria. It will be seen that a number of compositional and Tarskian truth theories are, plausibly, acceptable from a deflationary point of view and, therefore, do not encapsulate a substantial notion of truth. We conclude that the incompatibility thesis is false. Interestingly, our account also suggests that some popular compositional truth theories on the market are in fact not acceptable from a deflationary point of view. As we will argue, this does not constitute an embarrassment for deflationism, as there are good reasons to reject these theories on independent grounds.

3.2 Deflationism and the Orthodoxy

The variant of deflationism that will be the focus of this chapter consists of two fundamental claims. Some of its proponents are Field, Horsten, and Horwich, although their views might differ from each other in other, more satellite aspects.

The first core thesis of deflationism is that 'true', as it is deployed in theoretical contexts, is a primitive term governed by some form of equivalence between each truth ascription and the sentence or proposition itself to which truth is attributed to, i.e. by a so-called *transparency principle*. We will refer to this as the 'equivalence thesis'. This thesis is taken to suggest that there is no need or possibility of further conceptual analysis, no point in the search for an explicit definition of truth in terms of simpler, fundamentally more basic concepts—i.e. a *substantial account*. This is what distinguishes (this version of) deflationism from robust or substantive approaches to truth, such as the correspondence and the coherence theories, according to which there is a hidden nature of truth to uncover by means of an explicit definition, in which truth is analysed in terms of simpler concepts. Thus, deflationists sometimes claim that truth is not an ordinary or substantive property.

The second fundamental thesis of deflationism is that the sole reason for having a truth predicate in natural language is that it plays an indispensable logico-linguistic role. We will refer to this claim as the 'logico-linguistic function thesis'. For instance, the truth predicate allows us to endorse a single statement without explicitly articulating it, as in 'Goldbach's conjecture is true', or several, even infinitely many statements at once, as in 'All theorems of arithmetic are true'. It is this second thesis that distinguishes modern deflationism from its predecessor, the redundancy theory of truth. While in sentences such as '"Snow is white" is true' the truth predicate is easily eliminable and, therefore, dispensable, this is not so in the case of the two examples given above. For we may not know what Goldbach's conjecture is or what the theorems of arithmetic are. And in the latter case, even if we did, there are too many of them to assert them one by one.

There are several reasons why the deflationary account of truth motivates a formal treatment of the notion. Some authors outright assert that truth is a primitive undefinable notion that must be axiomatised (cf. Halbach and Horsten, 2005). Moreover, the so-called transparency principle that, according to the equivalence thesis, governs the truth predicate is simple, schematic, and reminiscent of those governing logical vocabulary. In addition, despite its simplicity, transparency is riddled with paradoxes when unrestricted and formulated over sufficiently strong logics and base theories. To successfully avoid contradictions, precise formulations are needed. Finally, and perhaps most importantly, the study of the logico-linguistic function which deflationists—and many non-deflationists as well—attribute to the truth predicate based on the inferential behaviour of truth obviously demands a formal treatment.

Indeed, recent years have seen a proliferation of formal truth theories, both in connection with and independently of deflationism. There has been much subsequent discussion about which formal properties a deflationary truth theory can and should have. Most agree that deflationists should opt for axiomatic systems, which thus will be the focus of this chapter. Although we don't directly address semantic theories, some of the arguments below can be applied equally to them.

Axiomatic truth theories consist of axioms for truth formulated over a base theory that contains a sufficient amount of syntax to provide the specific objects we will ascribe truth to, the truth-bearers, which, as is customary, we take to be sentences—or numbers that code sentences.[1] Let $\mathscr{L}$ be a first-order language, the language of the base theory, and let $\mathscr{L}_{\mathrm{T}}$ extend $\mathscr{L}$ with a monadic predicate T, for truth. We assume $\mathscr{L}$ contains enough vocabulary to express certain syntactic properties, relations, and functions of expressions of $\mathscr{L}_{\mathrm{T}}$ to be specified, and a quote name φ for each formula φ of $\mathscr{L}_{\mathrm{T}}$. Let Σ, the base theory, be a recursively axiomatised system formulated in $\mathscr{L}_{\mathrm{T}}$ containing a syntax theory for $\mathscr{L}_{\mathrm{T}}$ itself, which we assume is strong enough to relatively interpret first-order Peano arithmetic. For simplicity, we assume $\mathscr{L}$ has a term for every object in the domain of its intended interpretation and that Σ proves this, although all of our claims can be easily generalised if a satisfaction predicate is adopted instead. We also assume that only logical or syntactic principles containing T are derivable in Σ, but no truth principles. An axiomatic truth system Γ is then a recursive extension of Σ with axioms governing T.

What formal theories of truth can and should deflationists endorse? What truth axioms can and should a theory Γ consist of? The orthodoxy dictates that Γ should extend the base theory Σ only with a transparency principle. These are given by instances of so-called principles of (local) disquotation, that is, either the following schema:

$$\mathrm{T}\ulcorner\varphi\urcorner \leftrightarrow \varphi \tag{T-schema}$$

or the inference rules

$$\varphi \vdash \mathrm{T}\ulcorner\varphi\urcorner \tag{T-Intro}$$

$$\mathrm{T}\ulcorner\varphi\urcorner \vdash \varphi \tag{T-Elim}$$

possibly restricted to a class Δ of sentences of $\mathscr{L}_{\mathrm{T}}$, which may but need not necessarily coincide with $\mathscr{L}_{\mathrm{T}}$.

Some philosophers have claimed that the deflationist's truth axioms should consist of *all* instances of disquotation for sentences of $\mathscr{L}_{\mathrm{T}}$, including those that contain the truth predicate. Due to the semantic paradoxes, this would preclude the use of classical logic and force the adoption of weaker systems instead, adding yet another entry to the long list of restrictions imposed on deflationary theories. In Picollo and Schindler (2018b) we have given some reasons for believing that this restriction cuts too deep. We will not rehearse these arguments here, but simply assume that deflationists can adhere to classical logic. Horwich, one of the most vocal deflationists, clearly shares our view on this matter.

As anticipated in the introduction, it is usually maintained that compositional truth theories should be excluded from the deflationary picture. These theories get their name from their axioms, some of which are not instances of disquotation but compositional principles, such as

$$\forall x \forall y\, (\mathrm{Sent}_{\Delta}(x) \land \mathrm{Sent}_{\Delta}(y) \land \mathrm{Sent}_{\Delta}(x \underset{\cdot}{\land} y) \to (\mathrm{T}x \underset{\cdot}{\land} y \leftrightarrow \mathrm{T}x \land \mathrm{T}y)) \tag{$\mathrm{T}\land\upharpoonright\Delta$}$$

where $\mathrm{Sent}_{\Delta}(x)$ is a predicate that holds only of sentences in Δ and $\underset{\cdot}{\land}$ is a symbol for the function that maps every pair of formulae of $\mathscr{L}_{\mathrm{T}}$ to their conjunction (and similarly for the other logical connectives). $\mathrm{T}\land\upharpoonright\Delta$ states that if x, y, and their conjunction belong to Δ, then the conjunction is true just in case both conjuncts are. Similar principles can be given for the other logical connectives and the quantifiers.

The orthodox view also maintains that no Tarskian truth theory shall be endorsed by a deflationist. These theories extend Σ with an axiom of the form

$$\forall x\, (\mathrm{T}x \leftrightarrow \Phi(x)) \tag{$\mathrm{T}\upharpoonright\Delta$}$$

where $\Phi(x)$ holds only of sentences in Δ and T occurs in $\Phi(x)$ only applied to expressions of less complexity than x—sometimes considered to be a recursive (or explicit, if T doesn't occur in Φ at all) definition of

T. We will occasionally refer to principles of the form $T \restriction \Delta$ as ‘Tarskian principles’, or, if intended as definitions, as ‘Tarskian definitions’.

Next we will consider and discuss a series or arguments in favour of the orthodoxy, and show that, at best, they have limited reach.

3.3 Arguments for the Incompatibility Thesis

In this section we will rehearse six arguments that have been given in favour of the incompatibility thesis.

Argument 1. A reason often given for restricting the deflationist’s truth axioms to locally disquotational principles—of which Horwich (1998) is perhaps the most vocal promoter but many others have echoed him—stems from the equivalence thesis. According to the latter, the only basic facts about truth from a deflationist viewpoint are instances of transparency; they are “the whole truth about truth” (Stoljar and Damnjanovic, 2014). Thus, many have concluded, the axioms of a deflationary formal truth theory should consist exclusively of these *basic* principles, and every other fact about truth should be explained by—i.e. follow from—them. In support of this conclusion, consider the following remark by Horwich (2005): “the minimalist thesis is that the *basic* facts (i.e. the axioms of the theory that explains *every* other fact about truth) will all be instances of the [equivalence] schema” (p. 76).

Argument 2. A related argument often wielded against compositional and Tarskian truth theories *qua definitions* is also based on the equivalence thesis, which suggests that a definition of ‘true’ is *neither necessary nor possible*:

> For ‘true’ is a primitive term; so the only interesting account that can be given of its meaning is one that identifies which underlying property of the word (i.e. which aspect of our use of it) is responsible for its possessing that meaning. In particular, our truth predicate means what it does . . . in virtue of our underived commitment to the equivalence schema.
>
> (Horwich, 2005, pp. 75–76)

Thus, even if extensionally adequate, compositional and Tarskian truth theories cannot provide *real* definitions of truth.

Argument 3. Another argument commonly offered against the compatibility between deflationism and Tarskian truth theories stems from the logico-linguistic function thesis. If our truth predicate could be given by a Tarskian definition, then the language would already have the resources to formulate a predicate satisfying the relevant transparency principles. In this case, truth would be eliminable via the definiens, and thus the truth predicate would be *dispensable*. But according to the logico-linguistic function thesis, ‘true’ plays an indispensable role in (theoretically informed) natural language. Thus, Halbach and Horsten (2005) write: "definable notions of truth are not of primary interest to the deflationist because they are always just notions of truth for at best a part of our 'real' language" (p. 204).

ARGUMENT 4. Yet another reason given against compositional and Tarskian truth theories is that they *only work for simple, formal languages*. While Tarskian theories may be able to explain how the truth conditions of sentences of certain formal languages depend on the referents of their parts, it is not clear how they could deal with sentences of a natural language: "nobody has been able to show, for sentences involving 'that'-clauses, probabilistic locutions, attributive adjectives, or mass terms, how their truth could be explained by as a consequence of the referents of their parts" (Horwich, 2005, p. 77).

ARGUMENT 5. Following the line of thought of Argument 1, it has been argued that compositional and Tarskian truth theories are not available to the deflationist because they cannot be *derived* from what the deflationist considers to be the basic facts about truth. A particularly forceful objection of this kind is due to Gupta (2000), and has generated much discussion in the literature. Of course, it was essentially for this reason that (Tarski, 1935, p. 257) rejected an axiomatisation of truth based purely on instances of T-schema.

ARGUMENT 6. The sixth argument for the incompatibility thesis is an argument from substantiality. It has been claimed that compositional and Tarskian truth theories encapsulate substantial conceptions of truth because such theories are often *non-conservative* over their base theory, i.e. they allow us to prove claims in the language of the base theory that are not already provable in the base theory. In other words, Tarskian and compositional truth theories often allow us to gain more knowledge about the objects the base theory is about; thus, their truth predicate must be playing an explanatory role and, therefore, they must convey a substantial notion of truth.

At first glance, these arguments look convincing. At any rate, it appears that they have been accepted by many opponents of deflationism. Indeed, in light of the previous quotes by Horwich, one would think that (some) deflationists themselves have accepted the incompatibility thesis. However, there is a tension: there is a considerable amount of textual evidence that deflationists do in fact reject the incompatibility thesis. Field's work is a clear example, as he systematically advocates truth theories that validate compositional principles,[2] as does Horsten.[3] Moreover, Field (1999) himself has offered a forceful response to Argument 6, defending compositional truth theories against the charge of substantiality.

In addition, Horwich explicitly notes that the notions of truth, reference, and satisfaction actually do interact in the way indicated by Tarski, at least for certain fragments of English. He does not object to

Tarski's theory on the ground that its axioms are *incorrect*. Rather, he claims that these axioms "should not be treated as explanatorily basic, but should be explained in terms of simple, separate, minimal theories of truth, reference, and satisfaction" (Horwich, 1998, pp. 111–112). Similarly, Horwich explicitly endorses several compositional principles of truth, such as that a conjunction is true if and only if both conjuncts are true. Again, the reason that they do not feature among the axioms of his theory of truth is simply that he doesn't consider them as explanatorily basic.

In order to dissolve this tension and to show that the incompatibility thesis is incorrect, it will be helpful to have a closer look at the different purposes formal theories of truth can serve.

3.4 What Is a Formal Truth Theory Good for?

As many have pointed out, formal truth theories can serve various purposes. Soames (1984), for instance, distinguishes between three things a truth theory can do. First, it can serve as a faithful account of the behaviour of our natural language truth predicate. Call this a 'descriptive purpose'. As Soames points out, not many philosophers have attempted to provide a truth theory suited for descriptive purposes; rather, this is seen as the proper domain of linguistics. Philosophers, instead, have been mostly concerned with truth theories that put forward a new, precise, and consistent truth-like predicate intended as a replacement for our (possibly defective) natural language truth predicate. Soames gives the example of Tarskian truth theories as an illustration of theories of this kind. The third purpose he discusses involves cases where a notion of truth, taken to be antecedently understood, is deployed to explicate other related concepts such as meaning or knowledge or some general metaphysical view. A prominent example here is the use to which Davidson attempted to put Tarski-style truth theories in giving an account of natural language semantics.

In addition to these three purposes that a truth theory can serve, we would like to propose a fourth, which should be very close to the deflationist's heart. According to the deflationist's logico-linguistic function thesis, the truth predicate serves a role akin to that of the logical connectives. If deflationism is right, the truth predicate—roughly like conjunction, the conditional, the universal quantifier, etc.—plays an important expressive or inferential role. For deflationists and other philosophers who believe that truth plays such a role, it is only reasonable to want a formal truth theory capable of characterising the validity or correctness of inferences involving the notion of truth. As an analogy, it is helpful to compare the way in which, for instance, calculi for first-order logic play the role of characterising the validity or correctness of inferences involving negation, conjunction, quantifiers, etc. When a theory of truth plays this role, let us say that it serves a 'logical purpose'.[4]

The language of a formal truth theory intended to serve a logical purpose should be extensive or extensible enough that we can formalise (most of) our arguments involving the truth predicate (and other logical terms), just as first-order languages do for their logical terms. The aim, then, is to provide a theory that diagnoses an argument (suitably formalised and regimented) as valid just in case its premises entail the conclusion against the background of the truth theory.

In the remainder of this section we will argue that, while the first four arguments considered in the previous section have significant force when applied to formal truth theories intended to play a *descriptive* purpose, their force considerably diminishes when applied to formal theories intended to play a *logical* purpose instead.

If one is working with a broadly descriptive goal in mind, it is natural to impose certain constraints on one's formal truth theory Γ. Most significantly, it will be required that the truth-theoretic component of Γ closely reflects the actual usage or meaning of 'true'. Of course, it is almost inevitable in practice that even a truth theory offered in a descriptive spirit will be idealised in various ways; but the point is that the main criterion of success is fidelity to established usage. Similarly, the theory should not lapse into oversimplification. As Argument 4 suggests, we should plausibly expect that a descriptive theory Γ satisfactorily describes not only the behaviour of the truth predicate taken alone, but also within complex environments, e.g. within 'that'-clauses, probabilistic locutions, environments containing attributive adjectives, mass terms, etc., as these constructions are prevalent in (even theoretically informed) natural language.

Plausibly, given the emphasis that deflationism places on the equivalence and the logico-linguistic function thesis, one's deflationist commitments will impose additional constraints on theories put forward to serve the descriptive project. One is that no real definition of truth is possible, as Argument 2 suggests. Another is that no descriptively adequate truth theory will put forward an even nominally *definable* and, therefore, *eliminable* truth predicate, as prescribed by Argument 3. Finally, given the basic and exhaustive role the equivalence thesis ascribes to formal transparency principles in governing our usage of 'true', deflationists are committed to the claim that a broadly descriptively adequate truth theory will consist of instances (perhaps within a restricted class of sentences) of local disquotation. In particular, as Argument 1 suggests, all other truth-theoretic principles must then be derived from those instances; there seems to be no room for compositional or Tarskian axiomatisations within the descriptive project, at least as carried out by deflationists.

It is this line of reasoning, we believe, that lends a spurious plausibility to Arguments 1–4 in the previous section. To the extent that deflationists are attempting to offer an axiomatic truth theory capable of playing a *descriptive* role, these arguments can be endorsed and the constraints they propose can be taken as genuine ones. However, we believe that the plausibility of these arguments diminishes substantially when applied to a truth theory intended to serve a logical purpose, as we will now explain.

Assume we are attempting to formulate a formal theory of truth capable of serving a logical purpose. Naturally, the first thing we require is that the theory contains or entails principles governing the truth predicate sufficient to allow it to serve its logico-linguistic role.

At first, it may seem as if a truth theory of this kind must satisfy the same conditions that deflationists impose on their *descriptive* truth theories. After all, if a formal truth theory adequate for logical purposes were not also descriptively adequate, it is not clear how one could properly formalise natural language arguments involving the truth predicate. Moreover, if deflationism is right about the role of the (theoretically informed) natural language truth predicate, it seems that the truth predicate of a theory that is faithful to our usage should also be capable of serving that role.

However compelling these points may seem, we will argue that they do not adequately take into account the fact that simplification and idealisation are considerably more admissible in a theory intended for logical purposes than in a purportedly descriptive account. Moreover, for theories serving a logical purpose, it is less important to be faithful to the precise way that the meaning of the truth predicate is fixed in English. To make this point clear, we turn once again to the analogy with first-order languages and calculi.

Note first that first-order languages do not admit indexicals, 'that'-clauses, probabilistic locutions, attributive adjectives, mass terms, etc.; they work, as it were, with eternal or context-independent sentences only. For instance, if one wishes to formalise an English argument in a first-order language, one first needs to replace all indexicals with names referring to their referents (in the context of utterance) and make corresponding amendments. Arguably, this is a small price to pay for perspicuity and elegance, which is evidenced by how widely first-order logic is deployed in the analysis of the validity of arguments. It would seem equally reasonable for us to pay this price in the case of a formal truth theory we wish to adopt for logical purposes. *Pace* Argument 4, a theory of truth (be it a deflationary one or not) should not be expected to account for the interaction between truth and indexicals, 'that'-clauses, and other natural language oddities. Simplification and idealisation are permissible to a larger degree if one's purpose is not fundamentally descriptive.

Second, note that for logical purposes, whether the axioms of our theory coincide with the most basic principles governing our usage of the truth predicate in natural language does not matter. Adverting again to the analogy with logical constants, note that natural language usage is often ignored, just as e.g. Hilbert-style or sequent calculi for first-order logic make no pretence of capturing the most psychologically basic patterns of inference in their basic axioms and rules. All that matters is that, taken together, they provide us with an adequate account of validity for arguments involving truth (modulo simplification, and idealisation). Thus, whilst Arguments 1 and 2 might be compelling when considering formal truth theories intended for descriptive purposes, they aren't so when we want our theories for logical purposes instead.

Finally, if we are interested in making inferences involving only certain expressions, it would seem permissible to restrict our logical or truth-theoretic principles in such a way that our logical terms or truth predicate interacts exclusively with the relevant class of expressions. Of course, as an account of validity for a more encompassing class of expressions, the resulting theories will not be satisfactory; their use would be limited. In the case of truth, this could amount, for instance, to restricting the sound instances of disquotation (whichever these are) to a proper subclass. As a result, the truth predicate of the theory could turn out to be nominally definable. However, this does not conflict with the fundamental tenets of deflationism, as Argument 3 suggests, provided that the truth predicate of certain *extended* truth theories is not definable. It is compatible with the indefinability of our natural language truth predicate that when transparency is restricted to a subclass of expressions in our formal theories, the resulting predicate admits a nominal definition.

Despite the fact that Arguments 1–4 of the previous section fail to apply to truth theories intended for a logical purpose, theories of this kind should nevertheless still be expected to satisfy certain other conditions. We would like to mention two fundamental requirements: the *functionality* criterion and the *insubstantiality* criterion.[5]

The functionality criterion, indicated a few paragraphs above, demands that the axioms of the theory allow the truth predicate to perform the logico-linguistic function of truth—at least to a reasonable extent. Of course, which axioms are sufficient to achieve this goal depends entirely on the precise nature of the logico-linguistic function of truth, so an account of the latter is needed. We have developed such an account in Picollo and Schindler (2018a). In Section 3.5 we briefly review it and extract a precise criterion of functionality from it.

The insubstantiality criterion demands that truth theories do not convey a non-deflationary notion of truth, i.e. they should not entail that truth is a substantial property. What counts as a substantial truth property is naturally a very controversial issue which we are not able to resolve here. However, in Section 3.6 we will argue that if one starts with a truth theory that is taken to be insubstantial, then adding certain principles which, in a sense to be explained, follow from the axioms of the theory does not render the relevant notion of truth substantial.

Taken together, these criteria will allow us to recognise certain compositional and Tarskian theories as being deflationary, and hence to refute the two remaining arguments—Arguments 5 (cf. Section 3.6) and 6 (cf. Section 3.7). Consequently, we will argue that the thesis that compositional and Tarskian theories are necessarily committed to a substantial notion of truth can be put to rest, at least for now (cf. Section 3.8).

3.5 The Functionality Criterion

Our first criterion on an axiomatisation of truth intended for logical purposes is that such a theory must enable the truth predicate to fulfil its logico-linguistic role. In the present section, we will provide a precise formulation of this criterion.

It is striking that the purported logico-linguistic function the truth predicate is supposed to play has rarely been the subject of study in the literature, even by logicians or deflationists. One of the few articles providing a positive and formally precise account of the function is Halbach (1999). In Picollo and Schindler (2018b) we discussed this account and others that are hinted at in the literature, and argued they are unsuccessful; in Picollo and Schindler (2018a) we put forward our own positive account. Inspired by a tradition that originated in Ramsey, and to which Quine, Grover, and Azzouni, among others also belong,[6] we argued that the function deflationism ascribes to the truth predicate is best understood as enabling us to simulate sentential and predicate quantification within a first-order framework. In other words, the truth predicate lets us quantify into sentence and predicate position in an indirect way, i.e. without introducing sentential or predicate quantifiers. In still other words, the truth predicate and sentential and predicate quantifiers serve the same purpose.

For instance, to assert all theorems of first-order Peano arithmetic one could work with a monadic operator $\Box$ expressing provability in this theory plus sentential quantifiers, and write $\forall\alpha\,(\Box\alpha \rightarrow \alpha)$. Alternatively, one could use a provability predicate $\mathrm{Prov}(x)$ and a truth predicate, and assert $\forall x\,(\mathrm{Prov}(x) \rightarrow \mathrm{T}x)$. Similarly, one can generalise on $\varphi(t) \vee \neg\varphi(t)$ using second-order quantifiers, as in $\forall X(Xt \vee \neg Xt)$, or one can turn to the truth predicate and say

$$\forall x\,(\mathrm{Form}_1(x) \rightarrow (\mathrm{T}x(\ulcorner t \urcorner) \vee \neg\mathrm{T}x(\ulcorner t \urcorner)))$$

where $\mathrm{Form}_1(x)$ expresses the property of being a formula with only one free variable and $x(y)$ is the result of substituting in x the free variable with the term denoted by y.

In order to explain and substantiate our claims, in Picollo and Schindler (2018a) we have offered a series of formal results that establish that every theory in a language with sentential or predicate quantifiers—second- or higher-order, predicative or impredicative—can be 'naturally reformulated' in a language containing a purely disquotational truth predicate instead.[7] We first indicated how to translate every formula of a higher-order language into a first-order language with a truth predicate in a natural and effective way, i.e. along the lines of our examples in the previous paragraph. Instances of sentential comprehension—be they predicative or impredicative—translate into instances of local disquotation, provided the comprehension instances contain no free sentential variables, and into instances of *uniform* disquotation otherwise. On the other hand, instances of predicate comprehension always require uniform disquotation. The latter is a principle that generalises local disquotation to formulae with free variables. For instance, the following does so for formulae with one free variable:

$$\forall t\,(\mathrm{T}\ulcorner\varphi(\underset{\cdot}{t})\urcorner \leftrightarrow \varphi(t^{\circ})) \qquad \text{(Uniform T-schema)}$$

Here, $\forall t\,\psi$ abbreviates $\forall v\,(\mathrm{ClTerm}(v) \rightarrow \psi)$ for a suitable variable v, where $\mathrm{ClTerm}(v)$ expresses the property of being a closed term; $\ulcorner\varphi(\underset{\cdot}{t})\urcorner$ denotes the result of substituting t for the free variable in φ, and t° denotes the value of the term t.

We then proved that the proposed translation is a relative interpretation of the higher-order calculus into a—classical and consistent—disquotational truth theory, as it maps every higher-order derivation into a derivation in the truth theory that extends first-order classical logic with a suitable syntax theory and all instances of the (Uniform) T-schema for formulae in the range of our translation. This result is novel in so far as it establishes that (uniform) disquotation can even interpret *full impredicative predicate comprehension*, i.e. principles of the form

$$\exists X\,\forall v\,(Xv \leftrightarrow \varphi)$$

where φ itself may contain bound predicate variables. It shows that the proof-theoretic power of truth is much greater than previously thought. Moreover, we also showed that all *inferences* between translations that can be carried out in this truth theory are derivable in the calculus for higher-order logic.

Sentential and predicate quantifiers allow us to directly generalise over all sentences and formulae in the higher-order language. The truth predicate, we concluded, can bring about the same logical power: if (uniform) disquotation for translations of higher-order formulae is available, we can simulate quantification over the latter using their translations as proxies. More generally, one can use a truth predicate to simulate sentential and predicate quantification over a given class of expressions as long as the instances of disquotation for the expressions in this class—or their translations—are available. Our account of the function of truth confirms the common but rarely substantiated claim that (uniform) disquotation is both sufficient and necessary for the truth predicate to fulfil its role.

As a consequence, local disquotation for the class of sentences we wish to generalise over is desirable in our truth systems, and uniform disquotation even more so, especially if we wish to generalise into predicate position. In general, we would like to put forward the following adequacy criterion for formal truth theories intended for logical purposes:

Functionality A formal theory of truth intended for logical purposes should entail all instances of (uniform) disquotation for the class of expressions one wishes to generalise over.

Note that this criterion implies that generalising over the whole class of expressions of the language of the theory itself is not possible if classical logic is assumed in the background. For that would require that all instances of disquotation for sentences containing the truth predicate are derived, and triviality would follow. If one wishes to generalise unrestrictedly over all expressions of the language, one should probably look into non-classical truth theories instead. However, this might turn out to be not as straightforward as it seems. It is not entirely clear to us what inferences the truth predicate should validate in that case, as our results only establish the relative interpretability of *classical* higher-order theories in a disquotational truth theory. On the one hand, classical higher-order theories seem to be too strong to be relatively interpretable in a non-classical truth theory. On the other hand, very little is known about non-classical systems of higher-order quantification. In any case, the general lesson of our discussion should be clear: one first needs to determine what axioms and rules for truth are needed in a particular logic for it to fulfil its logico-linguistic function, and then derive a criterion of functionality from that.

3.6 The Insubstantiality Criterion

In the previous section we formulated a criterion of functionality that any formal truth theory intended for logical purposes ought to satisfy. However, not any such theory will do—for example, inconsistent or trivial theories are excluded, as they would obviously fail to adequately characterise the validity or correctness of inferences involving the notion of truth. Moreover, as we anticipated towards the end of Section 3.4, truth theories that can be legitimately endorsed by deflationists for logical purposes should also satisfy the insubstantiality criterion: they must not encapsulate a substantial notion of truth.

What encapsulating a substantial notion of truth amounts to is of course a matter of controversy, and we will not engage with the general metaphysical question of what a substantial property is. Instead, we will show that if one starts with an insubstantial theory, then the addition of certain compositional and Tarskian principles will not inflate that notion of truth. This will be the case, for instance, if the latter principles generalise on a schematic consequence of the starting theory. Again, we will not say much about what constitutes an insubstantial theory of truth. However, if deflationism is correct, then there must be at least one such theory—e.g. the theory consisting of all correct instances of disquotation. The purpose of this section is to show that if one starts from such an insubstantial and restricted truth theory, then adding certain compositional or Tarskian principles will not lead to an inflated notion of truth.

Despite its aspirations for generality, the T-schema cannot be stated by a single, universally quantified claim of the form $\forall x\,(\mathrm{T}x \leftrightarrow \ldots)$, as each instance has a sentence φ occurring inside quotes on the left-hand side and outside them on the right-hand side. The fact that each φ is both used and mentioned in its corresponding instance of the T-schema precludes a straightforward generalisation of the disquotational principle.

However, there are salient schematic principles that follow from the T-schema together with background syntactic assumptions, and which can easily be generalised. A simple warm-up example is given by the Uniform T-schema, which we already discussed in the previous section. Let Δ be the class of sentences for which an instance of local disquotation is available. Assume that, for some formula $\varphi(x)$, the sentence $\varphi(t)$ is in Δ for every closed term t. Thus, all the instances of the following are available:

$$\mathrm{T}\ulcorner\varphi(t)\urcorner \leftrightarrow \varphi(t)$$

Then it is easily seen that the relevant instance of uniform disquotation, i.e.

$$\forall t\,(\mathrm{T}\ulcorner\varphi(\dot{t})\urcorner \leftrightarrow \varphi(t^{\circ}))$$

is a straightforward generalisation of the schematic principle.

Let us look at another example. Consider the set of sentences such that both they and their negations are in Δ. Then, for each such sentence φ we can prove:

$$\mathrm{T}\,\underset{\cdot}{\neg}\ulcorner\varphi\urcorner \leftrightarrow \neg\mathrm{T}\ulcorner\varphi\urcorner$$

Since $\ulcorner\varphi\urcorner$ is a singular term, we can generalise on this principle as follows:

$$\forall x\,(\mathrm{Sent}_{\Delta}(x) \wedge \mathrm{Sent}_{\Delta}(\underset{\cdot}{\neg}\,x) \rightarrow (\mathrm{T}\,\underset{\cdot}{\neg}\,x \leftrightarrow \neg\mathrm{T}x)) \qquad (\mathrm{T}\neg\restriction\Delta)$$

Analogously, all the instances of the following principle (where φ, ψ, and $\varphi \wedge \psi$ are in Δ) follow from the T-schema as well:

$$\mathrm{T}\ulcorner\varphi \underset{\cdot}{\wedge} \psi\urcorner \leftrightarrow \mathrm{T}\ulcorner\varphi\urcorner \wedge \mathrm{T}\ulcorner\psi\urcorner$$

Replacing all occurrences of $\ulcorner\varphi\urcorner$ with x and those of $\ulcorner\psi\urcorner$ with y we can generalise on this schema by the following:

$$\forall x \forall y\, (\mathrm{Sent}_\Delta(x) \land \mathrm{Sent}_\Delta(y) \land \mathrm{Sent}_\Delta(x \underset{.}{\land} y) \rightarrow (\mathrm{T}x \underset{.}{\land} y \leftrightarrow \mathrm{T}x \land \mathrm{T}y)) \qquad (\mathrm{T}{\land}{\upharpoonright}\Delta)$$

Analogous principles for the other propositional connectives can be obtained likewise. Similarly, compositional principles for the quantifiers can be seen as generalisations of schematic consequences of local disquotation.[8]

Provided that Δ is a T-free *sublanguage* of $\mathscr{L}_\mathrm{T}$ (i.e. Δ is closed under logical predicates and operators) containing finitely many predicate symbols, one can also generalise on the T-schema by means of a so-called Tarskian definition, $\mathrm{T}{\upharpoonright}\Delta$. For instance, if all closed terms of $\mathscr{L}_\mathrm{T}$ occur in formulae in Δ, identity is the only predicate symbol, and $\neg$, $\land$, and $\forall$ are the only logical operators occurring in formulae in Δ, the following principle just 'puts together' the instance of uniform disquotation for the identity predicate and the compositional principles for the logical terms:

$$\begin{aligned}\forall x\, (\mathrm{T}x \leftrightarrow \mathrm{Sent}_\Delta(x) \ \land \ (&\exists s \exists t\, (x = (s \underset{.}{=} t) \land s^\circ = t^\circ) \ \lor \\ &\exists y\, (x = \underset{.}{\neg} y \land \neg \mathrm{T}y) \ \lor \\ &\exists y \exists z\, (x = y \underset{.}{\land} z \land \mathrm{T}y \land \mathrm{T}z) \ \lor \\ &\exists y \exists z\, (x = \underset{.}{\forall} y\, z \land \forall t\, (\mathrm{T}z(\underset{.}{t})))))\end{aligned}$$

We have seen how uniform disquotation, as well as compositional and Tarskian principles, can be 'extracted' from the T-schema by generalising on certain schematic principles that follow from it. They are general principles all of whose instances are already entailed by the latter. Arguably, these principles just provide more general ways of presenting the T-schema itself or some of its schematic consequences. If this is on the right track, then it is hard to see why they should be more substantial than the principles we started with. In this context, it is interesting to note that the way in which uniform disquotation generalises on local disquotation is not too different from the way the compositional principles do, so it is surprising that nobody has disputed the suitability of uniform disquotation as a deflationary truth principle. As we see things, if uniform disquotation is acceptable, then so are compositional principles.

There is one obvious worry regarding our reasoning above. Compositional, Tarskian, and uniform disquotation principles don't follow *logically* from their corresponding instances, but are usually (proof-theoretically) stronger than them, due to the compactness of the logical consequence relation. The possibility that this additional content inflates the notion hasn't been completely ruled out, despite the fact that these principles merely generalise schematic consequences of local disquotation.

Horwich's method for dealing with this objection is well known. Non-basic facts about truth need to be explained in terms of transparency principles together with *further* explanatory factors, i.e. principles that have nothing specifically to do with the truth predicate (cf. Horwich, 1998, p. 24; Horwich, 2005). We believe this strategy is essentially sound. Horwich himself appears to appeal to some form of ω-rule as an additional principle, which has provoked some criticism due to its infinitary character (cf. Raatikainen, 2005). Fortunately, there are other suitable principles. We will first describe what these principles are, and then discuss whether they are available to the deflationist.

We can bridge the gap between generalisations such as compositional, Tarskian, and uniform disquotation principles and their instances by informing the truth theory we are working with that, whenever it schematically proves all instances of a certain formula, the inference to the general claim that all instances of this formula hold is permissible (see Halbach, 2001a and Horsten and Leigh, 2017 for some formal results). Let $\mathrm{Prov}_\Gamma(x)$ express in $\mathscr{L}$ that x is a theorem of the formal theory Γ. Our gap-bridging principles take then the following form:

$$\forall t\,\mathrm{Prov}_\Gamma(\ulcorner \varphi(\underset{.}{t}) \urcorner) \to \forall t\,\varphi(t^\circ) \tag{GBP(Γ)}$$

Principles of this kind—not provable in Γ for familiar Gödelian reasons—allow us to formalise the 'extraction' of a general claim from its instances into a proper derivation. For example, let Γ extend the base theory Σ with all instances of local disquotation for sentences in Δ. Since Γ schematically derives all instances of

$$\mathrm{Sent}_\Delta(t) \wedge \mathrm{Sent}_\Delta(\underset{.}{\neg}\, t) \to (\mathrm{T}\,\underset{.}{\neg}\, t \leftrightarrow \neg \mathrm{T}t)$$

for every closed term t, adding GBP(Γ) to Γ delivers

$$\forall t\,(\mathrm{Sent}_\Delta(t^\circ) \wedge \mathrm{Sent}_\Delta(\underset{.}{\neg}\, t^\circ) \to (\mathrm{T}\,\underset{.}{\neg}\, t^\circ \leftrightarrow \neg \mathrm{T}t^\circ))$$

which, together with the fact—provable in Σ—that each sentence in Δ is denoted by a term in the language, entails the compositional principle $\mathrm{T}\neg{\restriction}\Delta$. Applying a similar reasoning, we can derive the Uniform T-schema and compositional principles restricted to Δ for the other propositional connectives in Γ extended with GBP(Γ). Finally, compositional principles for the quantifiers can be derived in Γ'+ GBP(Γ'), where Γ' = Γ+GBP(Γ).

A similar argument can be given in the case of so-called Tarskian definitions, $\mathrm{T}{\restriction}\Delta$. Let Γ be as before. If, additionally, Δ is a T-free sublanguage of $\mathscr{L}_\mathrm{T}$ as before, then $\mathrm{T}{\restriction}\Delta$ follows in Γ from uniform disquotation and the compositional principles for all logical terms, both restricted to Δ, plus the following:

$$\forall x\,(\mathrm{T}x \rightarrow \mathrm{Sent}_{\Delta}(x)) \tag{$\beta\restriction\Delta$}$$

Thus, $\mathrm{T}\restriction\Delta$ follows in Γ from (iterated applications of) GBP(Γ) together with $\beta\restriction\Delta$, which states that only sentences in Δ can be true.

Does GBP(Γ) qualify as a suitable additional principle that the deflationist can employ in explaining certain facts about truth? Suppose we work with classical logic, as Horwich does, and assume for a moment that we firmly endorse the deflationary acceptable truth theory Γ: when I learn that some sentence is provable in Γ, I have good reasons to believe it. Now let $\varphi(x)$ be a formula of $\mathscr{L}_{\mathrm{T}}$ and consider the following instance of excluded middle:

$$(\forall t\,\mathrm{Prov}_{\Gamma}(\ulcorner\varphi(\underset{.}{t})\urcorner) \rightarrow \forall t\,\varphi(t^{\circ}))\ \vee\ \neg\,(\forall t\,\mathrm{Prov}_{\Gamma}(\ulcorner\varphi(\underset{.}{t})\urcorner) \rightarrow \forall t\,\varphi(t^{\circ}))$$

Which of the two disjuncts should we endorse? Consider the second disjunct. Accepting it commits us to the claim that although $\varphi(t)$ is provable in Γ for every closed term t, nonetheless there is a closed term t such that $\neg\varphi(t)$. This entails that we should not accept some consequences of Γ! Since we firmly endorse Γ, we should reject the second disjunct, and therefore accept the first disjunct. But the latter is just an instance of GBP(Γ).

Let us clarify one point, before dealing with some objections. Given an instance of excluded middle, one can in general remain agnostic about which disjunct obtains. For example, a classical set theorist is committed to the claim that either the continuum hypothesis or its negation holds, but she may remain agnostic about which disjunct holds barring new evidence. However, the present case is different. The second disjunct entails that some consequences of Γ don't hold. Thus, if you firmly endorse Γ, you ought to reject it and accept the first disjunct, even if the statement is independent of Γ. Anything else would be incoherent. But now, once you have accepted GBP(Γ) as an additional (non-truth-theoretic) principle, other truth-theoretic principles follow.

We cannot see any good reason why the truth-theoretic principles that follow from adding GBP(Γ) to our truth theory Γ should inflate the notion of truth. We have assumed that the truth-theoretic principles of Γ are insubstantial. In arguing for GBP(Γ), we have not appealed to the notion of truth, let alone a substantial notion of truth. Moreover, GBP(Γ) itself isn't formulated in terms of truth. In what follows, we anticipate three possible objections.

Objection 1. The argument assumes that $\mathrm{Prov}_{\Gamma}(x)$ 'expresses' the property of being provable in Γ. The standard explanation of why it does so involves the notion of truth in the standard model of the base theory Σ—e.g. the standard model of arithmetic. However, the latter is not admissible to a deflationist, because on their account truth is characterised through transparency.

Our reply to this objection is essentially identical to that given by Cieśliński (2017, p. 153). Very roughly, $\mathrm{Prov}_{\Gamma}(x)$ ‘expresses’ the property of being provable in Γ because the way the predicate is defined structurally resembles the way how ‘provable in Γ’ is defined in the metalanguage of Γ. We find this response especially plausible in this context because deflationists usually rely on a use theory of meaning—rather than on truth-conditional semantics—according to which the meaning of ‘provable in Γ’ must be given through some rules for using that expression.

Objection 2. GBP(Γ) is a schematic principle, and according to deflationists the sole purpose of the truth predicate is to generalise sentence places in our language. Thus, we ought to formulate GBP(Γ) as a single statement deploying the truth predicate. But then it becomes apparent that our additional principle is of a truth-theoretic nature after all.

We do not find this objection very convincing. First, it is not generally the case that whenever we generalise a schema using the truth predicate the resulting statement is a truth-theoretic statement. The claim that everything the Pope said is true or that all theorems of arithmetic are true is not a truth-theoretic statement, although it involves the notion of truth. According to the logico-linguistic function thesis (the second core tenet of deflationism), such generalisations do little more than express all papal assertions or all theorems of arithmetic in a compact way.

Second, even if the truth predicate allows us to express the schema in a single statement, we are certainly not obliged to do so. At any rate, it is hard to see how the fact that we can derive compositional principles of truth using GBP(Γ)—which is not stated in terms of truth—could be undermined by the fact that we can generalise GBP(Γ) using the notion of truth.

Third, we know that due to the paradoxes it is not possible to generalise over all sentence places in our language (at least as long as we adhere to classical logic). We can only do so for a restricted class of sentences. But GBP(Γ) is a schema that ranges over all sentences. Thus it is not even clear that we can generalise GBP(Γ) using the notion of truth. (It might be thought that all this shows is that the deflationary account of truth is incompatible with the use of classical logic. We have argued in Picollo and Schindler (2018b) that this is not the case.)

Objection 3. GBP(Γ) is unacceptable because it is inconsistent with certain unrestricted compositional axioms for truth.

We are not particularly worried by this objection either. On our view, deflationists ought to reject unrestricted compositional axioms for truth on quite independent grounds already, so their inconsistency with GBP (Γ) cannot cast doubt on the latter. Very roughly, the reason why deflationists ought to reject unrestricted compositional axioms for truth is *precisely* because not all of their instances are generally entailed by restricted disquotational principles of truth. If only instances of disquotation for a given class of expressions Δ are available, it is hard to see how compositional or Tarskian principles whose instances go beyond Δ can be justified on the basis of the original theory, even if additional non-truth-theoretic principles are invoked.[9] We return to this point in Section 3.8, at the end of the chapter.

Our preceding argument for GBP(Γ) relies on the law of excluded middle and so might not be available to all deflationists. It is difficult to say something in general here, as the matter will depend on the details of the non-classical system. At any rate, since our goal is merely to show that deflationism is compatible with compositional and Tarskian truth theories, it is sufficient if we can make our point in the case where the deflationist account is based on classical logic.

To sum up, we maintain that the addition of certain compositional, Tarskian, and uniform disquotation principles does not thicken the notion of truth conveyed by a deflationary adequate truth theory. First, we pointed out that certain compositional principles and Tarskian definitions are mere generalisations of schematic consequences of a class of instances of local disquotation, so it is hard to see how they could possibly inflate the notion of truth. We then pointed at the existence, under certain given conditions, of derivations of the more general principles from local disquotation plus other non-truth theoretic claims deflationists may reasonably endorse. (Of course, if such proofs are not available—which will largely depend on the restrictions imposed on local disquotation and the background logic—there is no guarantee of the legitimacy of the general principles.) This motivates the following criterion:

Relative Insubstantiality The (truth-theoretic) axioms of a formal truth theory are insubstantial if they are derivable in an insubstantial locally disquotational theory of truth together with additional non-truth-theoretic principles a deflationist may reasonably endorse.

The qualification ‘derivable in an *insubstantial* locally disquotational theory etc.’ is important: not every class of instances of local disquotation is necessarily insubstantial. For example the class that comprises all the instances, being inconsistent, entails every truth principle whatsoever, even those one would readily call inflationary, e.g. that truth is correspondence with fact (if expressible in the language). Other consistent subsets of this class will also be inadmissible for similar reasons, for although they will not entail every sentence of the language, some of them will entail substantial claims about truth, as will be seen in

Section 3.8. As we have said before, we won't offer a definition of what constitutes an insubstantial disquotational theory of truth, but if deflationism is correct, such theories do exist—the theory consisting of all correct instances of disquotation being one of them.

3.7 The Argument From Conservativeness

There is one objection that one could mount against our criterion of insubstantiality. This is the argument from conservativeness, mentioned in Section 3.3, namely, Argument 6. We will now deal with this objection.

The equivalence thesis commits deflationism to the idea that any attempt to uncover the nature of truth beyond disquotation, the quest for a real definition of truth in terms of simpler notions is futile. According to deflationism, truth cannot be defined or further analysed; it is a *sui generis* property, if a property at all. This is often expressed by saying that truth has no nature, is metaphysically thin, or is otherwise insubstantial, but of course these are just metaphors. Many, however, have taken them to be a—and even the—defining feature of deflationism. Moreover, some understand the insubstantiality of truth to entail that truth cannot have any explanatory power. Shapiro (1998), for instance, claims that "[i]f truth/satisfaction is not substantial—as the deflationist contends—then we should not need to invoke truth in order to establish any results not involving truth explicitly" (p. 497). Formally, this translates in a natural way into what is known as the 'conservativeness requirement': deflationary truth theories should be conservative over their respective base theories—which should contain some amount of syntax (cf. Halbach, 2001b)—i.e. the addition of truth principles to a base theory should not allow us to prove new theorems in the language without the truth predicate. This requirement has been argued for by e.g. Horsten (1995), Shapiro (1998), and Ketland (1999).

Another—related—road to conservativeness draws from the function deflationism assigns to truth. Its only purpose, as stated by the logico-linguistic function thesis, is a logico-linguistic one. Thus, it has been argued, there is no room for an explanatory role of truth within deflationism. In Horwich's words:

> A deflationist attitude toward truth is inconsistent with the usual view of it as a deep and vital element of philosophical theory. Consequently the many philosophers who are inclined to give the notion of truth a central role in their reflections in metaphysical, epistemological, and semantic problems must reject the minimalist account of its function. Conversely, those who sympathize with deflationary ideas about truth will not wish to place much theoretical weight on it. They will maintain that philosophy may employ the notion only in its minimalist capacity—that is, as something enabling the

> formulation of certain generalizations—and that theoretical problems must be resolved without it.
>
> (Horwich, 1998, p. 52)

Again, if deflationary truth must not play a role in the resolution of theoretical issues, then the conservativeness requirement follows (or so it argued).

Most compositional theories of truth on the market are, however, not conservative over their respective base theory (cf. Halbach, 2014; Horsten, 2011). Thus, if the conservativeness requirement is right, these theories are not deflationary. But also many *untyped disquotational* theories aren't conservative over their base theory either, some of which seem fairly attractive from a deflationary perspective, as the restriction they impose on the instances of disquotation can be justified from a philosophical point of view (cf. Picollo, 2019, for instance).

On our view, however, the conservativeness requirement not only does not follow from the core theses of deflationism outlined in the introduction of this chapter but also is not a reasonable requirement to be imposed on deflationary truth theories. Indeed, we claim that the conservativeness requirement is the result of (a) inferring too much from the metaphor of insubstantiality and (b) failing to see what the function of truth really amounts to. The analysis of this function, briefly sketched in Section 3.5, actually points (in many cases) in the opposite direction.

Let us focus briefly on sentential and predicate quantifiers. While their role—whether logical or quasi-logical, we would not like to enter this dispute here—is merely expressive, their addition to a first-order base theory does not always yield a conservative extension. Now, in Picollo and Schindler (2018a) we've argued that the logico-liguistic function deflationism ascribes to the truth predicate is best understood as enabling us to simulate sentential and predicate quantification in a first-order setting, as mentioned in Section 3.5. In other words, from a deflationist perspective, the truth predicate—together with the first-order quantifiers—has the *same function* as sentential and predicate quantifiers. As a consequence, we should not expect a formal truth theory well suited for functional purposes to conservatively extend its base theory either. On the contrary, non-conservativeness is just a feature of the truth predicate fulfilling its role. The conservativeness requirement cannot stem from the logico-linguistic function thesis; a 'mere' expressive role is compatible with the violation of conservativeness.

Can the equivalence thesis support an argument for conservativeness? If so, it would be devastating for deflationism: while one of its core theses would point to conservative theories, the other points in the opposite direction. Are the two fundamental theses of deflationism incompatible with each other? We believe this is not the case. If we look closely at the equivalence thesis, there is good reason to believe that the insubstantiality metaphor is just meant to indicate that the truth predicate, unlike other predicates, does not play a *descriptive* role in our language; truth ascriptions are not descriptions of the truth-bearers involved. To quote Frege (1956), "nothing is added to the thought by my ascribing to it the property of truth" (p. 293), so the latter is not an ordinary or substantial property. As such, truth cannot play the explanatory role ordinary properties play, i.e. to highlight an aspect of the object of study that would explain some of the characteristics of this object. But this doesn't exclude the possibility that the truth predicate plays an explanatory role of a different kind, i.e. in proofs. Indeed, sentential and predicate quantifiers can lead to new knowledge as well and therefore have explanatory value (assuming that proofs can have explanatory value), without being in any way descriptive. Their explanatory value derives solely from their role as a logico-linguistic device; the fact that they have explanatory value does not indicate in any way that they are 'substantial'. Since the truth predicate plays the same function as these quantifiers, similar considerations apply to it. Thus, we echo Field (1999) when he says that "any use of 'true' in explanations which derives solely from its role as a device of generalization should be perfectly acceptable" (p. 537).

We therefore conclude that the conservativeness requirement should be given up; it cannot be used as an argument against the admissibility of certain truth-theoretic axioms for deflationism.

3.8 Revisiting the Incompatibility Thesis

It is time to take stock. We have looked at a number of arguments for the incompatibility of deflationism, on the one hand, and compositional and Tarskian truth theories, on the other. We have pointed out that the majority of these arguments ostensibly presuppose a particular purpose, i.e. to *describe* the basic usage of the truth predicate in natural language. This is a legitimate enterprise and we do not necessarily disagree with some of the objections if judged against this purpose. However, we were quick to point out that the deflationist may want a formal theory of truth for a slightly different purpose, that is, to provide an account of the validity or correctness of arguments involving the truth predicate.

We have formulated two constraints that any formal truth theory intended to serve a logical purpose ought to satisfy: functionality and insubstantiality. A formal truth theory intended for logical purposes should entail all instances of (uniform) disquotation for the class of expressions one wishes to generalise over and, moreover, its axioms should be insubstantial. Although we did not provide a general criterion of insubstantiality, we argued that a formal truth theory is insubstantial if its axioms are derivable in a locally disquotational truth theory which is itself insubstantial together with additional non-truth-theoretic principles a deflationist may reasonably endorse. With these constraints at hand, let us now have a look at some of the classic formal truth theories one can find in the literature and see whether they can be endorsed by a deflationist.[10]

Let us start with what is probably the best-known and most simple formal truth theory: the theory that extends its base theory with all T-free instances of the T-schema—usually known as TB, for 'Tarski Biconditionals'. This theory satisfies our functionality criterion: if one merely wishes to quantify into sentence position over the class of T-free sentences, TB will do. Moreover, it is widely believed to convey an insubstantial notion of truth. Based on this, the uniform version of TB, UTB (for 'Uniform Tarski Biconditionals'), can also be seen to be deflationary because its axioms follow from TB together with additional non-truth-theoretic principles an advocate of TB may reasonably endorse, e.g. GBP(TB). Since UTB entails all instances of uniform disquotation for T-free predicates, it improves on TB, as it also allows us to quantify into predicate position over this class of formulae.

Similar considerations also apply to other locally disquotational theories: if the local theory is in good standing, so will be its uniform version. Note, however, that some locally disquotational theories might actually not be in good standing. This is obviously the case of the (classical) theory containing an instance of the T-schema for each sentence of $\mathscr{L}_\mathrm{T}$, as it is inconsistent. But there are other purely disquotational theories that are consistent and yet violate some of our criteria. Assume φ expresses a substantial truth principle—e.g. that truth is essentially correspondence with the facts. Deploying a trick of McGee (1992), we know there is an instance of local disquotation that is provably equivalent (in the base theory) to φ. Hence, any theory containing that instance will be substantial.

Let us now turn to compositional truth theories, i.e. systems in which instances of disquotation are only given for atomic expressions—or sometimes also negations of atomic expressions—whereas other truth axioms are compositional. As we have argued, the latter are admissible if they follow from Γ and suitable non-truth-theoretic principles. Such is the case of the axioms of CT, which extends the base theory with uniform disquotation for each *primitive* predicate in the T-free fragment of the language and compositional principles for the connectives and quantifiers, also restricted to sentences without T. CT is acceptable because it follows from GBP(UTB) (cf. Halbach, 2001a), which we already have seen to be acceptable.

Still, one might wonder what the use of compositional theories like CT would be, given that they merely generalise on instances of disquotation, which are already sufficient for the function of truth. Since all these instances of disquotation are derivable in the compositional theory, there seems to be no reason not to endorse it. But are there any positive reasons?

There are at least two—intertwined—motives why compositional theories could be preferable to corresponding disquotational systems. First, compositional principles allow us to reason more generally about truth. This can, in turn, provide us with simpler and shorter proofs (cf. Fischer, 2014). Second, compositional principles can be used to provide us with a finite or more concise theory. If the T-free fragment of the language contains finitely many primitive predicates, CT can be seen as a finite and more general way of "formulating" the truth-theoretic part of both TB and UTB. If there are infinitely many primitive predicates in the language instead, CT also contains infinitely many axioms, but is still more general and concise than its disquotational counterparts, as e.g. it doesn't contain one instance of disquotation for each negated expression but all negations are dealt with by a single axiom in a general manner, and similarly for the other logical terms.

For analogous reasons, Tarskian truth theories can be deflationary admissible for logical purposes and even preferable to local or uniform disquotational theories for the same class of expressions: they are more general and concise than the latter. Furthermore, since they have the form of a (recursive) definition, we know they do not introduce any inconsistencies to the base theory, which is clearly a theoretical advantage.

This shows that the incompatibility thesis—i.e. that deflationism, on the one hand, and compositional and Tarskian theories, on the other, are not compatible—is mistaken after all. However, so far we have only given evidence of the admissibility of typed theories of truth. Let us therefore conclude the chapter by briefly surveying some untyped theories.

Let us first consider the well-known system KF, Feferman's axiomatisation of Kripke's fixed-point theory of truth in classical logic. KF extends the base theory with uniform disquotation for atomic and negation of atomic formulae that don't contain T, plus "positive" compositional principles for *every* sentence of the language, including those containing T, and an axiom governing attributions of untruth. No axiom of the theory states that truth commutes with negation, but compositional axioms for double negations, conjunctions, disjunctions, universal claims, etc., and negated conjunctions, negated disjunctions, negated universal statements, etc. belong to KF. For instance, the compositional axiom for negated disjunctions is the following:

$$\forall x \forall y\, (\mathrm{Sent}_{\mathscr{L}_\mathrm{T}}(x) \wedge \mathrm{Sent}_{\mathscr{L}_\mathrm{T}}(y) \rightarrow (\mathrm{T}\, \underset{\cdot}{\neg}(x \underset{\cdot}{\vee} y) \leftrightarrow (\neg \mathrm{T}x \wedge \neg \mathrm{T}y))) \qquad (\mathrm{T}\neg\vee)$$

Let us now ask whether KF satisfies the criteria we set out. Is it functional? KF implies instances of (uniform) disquotation for a certain class of expressions Δ (including all T-free sentences), so if one's goal is to quantify over expressions in Δ, functionality is satisfied. Is it insubstantial? We have not provided an absolute criterion of insubstantiality, but one way to show it to be insubstantial would be to look for an insubstantial disquotational theory that implies the axioms of KF, given additional non-truth-theoretic principles a deflationist may reasonably endorse.

Note that KF's compositional axioms are unrestricted; that is, they govern the interaction of the truth predicate and the logical operators as they apply to every expression of the language. Thus, a natural theory of disquotation that implies them (given additional non-truth-theoretic principles) would be the theory containing all instances of the T-schema. But this class of sentences is obviously not in good standing, for it leads to triviality. Could some other disquotational theory do the job?

The short answer is yes, trivially. Recall that McGee's trick entails that every sentence of $\mathscr{L}_{\mathrm{T}}$ is provably equivalent to an instance of the T-schema in the base theory. Thus, for every truth theory, whether compositional, Tarskian, disquotational, or else, there is a purely disquotational theory that proves the same theorems. *A fortiori*, there is a disquotational theory that has exactly the same consequences as KF (even without any gap-bridging principles). However, since these theories are otherwise highly unmotivated, we have little reason to believe that they are themselves in good standing.

Perhaps more interestingly, as Horsten and Leigh (2017) have shown, the axioms of KF can be derived by iterating GBP twice over the theory PTB, which extends the base theory with an instance of local disquotation for each sentence of $\mathscr{L}_{\mathrm{T}}$ in which the truth predicate occurs only positively—i.e. under the scope of an even number of negations. However, whether this theory is in good standing is rather doubtful. Restricting the T-schema to positive instances is quite *ad hoc*. It isn't based on any well-motivated criterion of what an acceptable instance is, but merely on the observation that the liar sentence and other paradoxical expressions aren't positive. Just like positive set theory, which avoids Russell's paradox by restricting comprehension to positive instances, this leads to a mathematically interesting theory, but to a rather strange picture of truth (sets).[11] Of course, one could justify PTB by pointing out that its axioms are derivable from KF, as Halbach (2014) observes, but this is of little use in the present context. Overall, we have little reason to believe that KF qualifies as a deflationary theory of truth.

Similar considerations apply to FS, though in this case one can actually give positive reasons to reject it. FS is the classical theory extending the base theory with uniform disquotation for T-free atomic expressions and compositional axioms for the connectives and the quantifiers just like CT's, except the restriction to T-free sentences is lifted. Additionally, FS contains two 'meta'-rules of inference that allow us to attach the truth predicate to and remove it from every theorem of the theory. As is well known, FS is ω-inconsistent, i.e. it proves all instances $\varphi(t)$ of a formula $\varphi(x)$ but, at the same time, it also entails $\neg\forall x\varphi(x)$. Thus, FS is in a sense unsound, as is every disquotational theory that entails the axioms of FS (with or without additional non-truth-theoretic principles). So no such disquotational theory appears to be in good standing.

In general, we are suspicious that *classical* theories containing unrestricted compositional axioms—i.e. axioms applying to *all* sentences of the language, including those with the truth predicate—can be shown to follow from some insubstantial disquotational theory together with additional non-truth-theoretic principles. In most cases, the only disquotational theories that come to mind here are those obtained by McGee's trick, for which it is quite doubtful that they are in good standing. Thus, as far as classical type-free theories are concerned, it would seem to be more promising to search for systems that restrict disquotational or compositional principles to a proper subclass of sentences of the language of truth, such as e.g. the grounded ones.[12] It is no coincidence that the theories of truth proposed by the authors, e.g. Picollo (2019) or Schindler (2014), have gone in that direction. In this respect, non-classical theories might be at an advantage, insofar as they might have all instances of disquotation at their disposal, though this requires some further investigation.

Acknowledgments

A first draft of this chapter was written while Thomas Schindler received support from the European Commission within the project *The Logical Function of Property Talk* (LOFUPRO, grant agreement no. 792202). The final version was written while Thomas Schindler received support from the European Research Council (ERC) within the project *Truth and Semantics* (TRUST, grant agreement no. 803684). In addition, both authors acknowledge the support of the German Research Council (DFG, Deutsche Forschungsgemeinschaft) within the project *Reference Patterns of Paradox* (GZ: PI 1294/1-1).

Notes

1. Should propositions be preferable to sentences, one could understand our truth predicate as applying not directly to the sentences but to what these sentences express.
2. See, for instance, Field (2003, 2008).
3. See, e.g. Horsten (2011).
4. We use this terminology for lack of a better alternative: in particular, in using it we do *not* wish to suggest that truth is a distinctively logical notion; rather, we use it to emphasise the aim of laying down general principles governing the validity or correctness of inferences involving truth.

5. There have been several attempts to formulate general requirements on axiomatic theories of truth, e.g. Leitgeb (2007) and Sheard (2002); a list of desiderata specifically designed for deflationists has been proposed by Halbach and Horsten (2005). Although reasons of space prevent a direct comparison, it should be emphasised that our desiderata differ decidedly from theirs.
6. See, for instance, Ramsey (1927, p. 158), Quine (1970), Grover (1972), Grover et al. (1975), and Azzouni (2001).
7. This result relies on the assumption, mentioned in Section 3.2, that every object in the domain has a name. Again, that restriction can be lifted if we work with a satisfaction rather than a truth predicate.
8. Recall that we assumed that we can prove in the base theory that for every object there is a term denoting this object. Again, if one wants to lift that restriction, one needs to work with a satisfaction predicate instead.
9. A similar point was made by Armour-Garb and Beall (2005, Section 5.1).
10. For an overview of axiomatic truth theories, see Halbach (2014) or Horsten (2011).
11. See Schindler (2015, pp. 398–399) for further arguments against PTB.
12. See Schindler (2020, sec. 3–4) for further discussion.

References

Armour-Garb, B. and Beall, J. C. (2005). Minimalism, epistemicism, and paradox. In Armour-Garb, B. and Beall, J. C., editors, *Deflationism and Paradox*, pages 85–96. Oxford University Press.

Azzouni, J. (2001). Truth via anaphorically unrestricted quantifiers. *Journal of Philosphical Logic*, 30: 329–354.

Cieśliński, C. (2017). *The Epistemic Lightness of Truth. Deflationism and its Logic*. Cambridge University Press.

Field, H. (1999). Deflating the conservativeness argument. *Journal of Philosophy*, 96: 533–540.

Field, H. (2003). A revenge-immune solution to the semantic paradoxes. *Journal of Philosophical Logic*, 32: 139–177.

Field, H. (2008). *Saving Truth from Paradox*. Oxford University Press.

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

Frege, G. (1956). The thought: A logical inquiry. *Mind*, 65: 289–311.

Grover, D. L. (1972). Propositional quantifiers. *Journal of Philosophical Logic*, 1: 111–136.

Grover, D. L., Camp, J. L., and Belnap, N. D. (1975). A prosentential theory of truth. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, 27: 73–125.

Gupta, A. (2000). Minimalism. *Philosophical Perspectives*, 7: 359–369.

Halbach, V. (1999). Disquotationalism and infinite conjunctions. *Mind*, 108: 1–22.

Halbach, V. (2001a). Disquotational truth and analyticity. *Journal of Symbolic Logic*, 66: 1959–1973.

Halbach, V. (2001b). How innocent is deflationism? *Synthese*, 126: 167–194.

Halbach, V. (2014). *Axiomatic Theories of Truth*. Cambridge University Press, 2nd edition.

Halbach, V. and Horsten, L. (2005). The deflationist's axioms for truth. In Armour-Garb, B. and Beall, J. C., editors, *Deflationism and Paradox*. Oxford University Press.

Horsten, L. (1995). The semantical paradoxes, the neutrality of truth and the neutrality of the minimalist theory of truth. In Cortois, P., editor, *The Many Problems of Realism*, volume 3 of Studies in the General Philosophy of Science, pages 173–187. Tilburg University Press.

Horsten, L. (2011). *The Tarskian Turn: Deflationism and Axiomatic Truth*. MIT Press.

Horsten, L. and Leigh, G. (2017). Truth is simple. *Mind*, 126: 195–232.

Horwich, P. (1998). *Truth*. Oxford University Press, 2nd edition.

Horwich, P. (2005). A minimalist critique of tarski on truth. In Beall, J. C. and Armour-Garb, B., editors, *Deflationism and Paradox*, pages 75–84. Oxford University Press.

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

Leitgeb, H. (2007). What theories of truth should be like (but cannot be). *Philosophy Compass*, 2(2): 276–290.

McGee, V. (1992). Maximal consistent sets of instances of Tarski's schema. *Journal of Philosphical Logic*, 21: 235–241.

Picollo, L. (2019). Reference and truth. *Journal of Philosophical Logic*. https://doi.org/10.1007/s10992-019-09525-9.

Picollo, L. and Schindler, T. (2018a). Deflationism and the function of truth. *Philosophical Perspectives*, 32: 326–351.

Picollo, L. and Schindler, T. (2018b). Disquotation and infinite conjunctions. *Erkenntnis*, 83: 899–928.

Quine, W. V. O. (1970). *Philosophy of Logic*. Harvard University Press.

Raatikainen, P. (2005). On Horwich's way out. *Analysis*, 65: 175–177.

Ramsey, F. P. (1927). Facts and propositions. *Proceedings of the Aristotelian Society*, 7: 153–170.

Schindler, T. (2014). Axioms for grounded truth. *Review of Symbolic Logic*, 7: 73–83.

Schindler, T. (2015). A disquotational theory of truth as strong as Z-. *Journal of Philosophical Logic*, 44: 395–410.

Schindler, T. (2020). A note on Horwich's notion of grounding. *Synthese*, 197: 2029–2038.

Shapiro, S. (1998). Proof and truth: Through thick and thin. *Journal of Philosophy*, 95: 493–521.

Sheard, M. (2002). Truth, provability and naive criteria. In Halbach, V. and Horsten, L., editors, *Principles of Truth*, pages 169–181. Hänsel-Hohenhausen.

Soames, S. (1984). What is a theory of truth? *Journal of Philosophy*, 81: 411–429.

Stoljar, D. and Damnjanovic, N. (2014). The deflationary theory of truth. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2014 edition.

Tarski, A. (1935). The concept of truth in formalized languages. In *Logic, Semantics, Metamathematics*, pages 152–278. Clarendon Press.

4 Truth, Reflection, and Commitment

Leon Horsten and Matteo Zicchetti

4.1 Introduction

Proof-theoretic reflection principles have been discussed in proof theory ever since Gödel's discovery of the incompleteness theorems. But these reflection principles have not received much attention in the philosophical community. The aim of the present chapter is to survey some of the principal meta-mathematical results on the iteration of proof-theoretic reflection principles, and also to investigate these results from a logico-philosophical perspective; we will concentrate on the epistemological significance of these technical results and on the epistemic notions involved in the proofs. In particular, we will focus on the notions of *commitment to* and *acceptance of* a theory. Special attention is given to the connection between proof-theoretic reflection and axiomatic truth theories.

After distinguishing between different types of proof-theoretic reflection principles, we review some proof-theoretic results concerning extensions of formal theories by (iterated) reflection principles. As basis theories we concentrate on standard arithmetical and elementary axiomatic truth theories. We then go on to explore the epistemological significance of these results. In this investigation we aim at showing that epistemic notion of acceptance of (or commitment to) a theory plays a crucial role in the philosophical argumentation for reflection principles and their iteration.

The structure of this chapter is as follows. In Sections 4.2 and 4.3, iterated reflection over arithmetical theories is discussed. In Section 4.4, we discuss reflection over axiomatic truth theories—here we concentrate on theories of disquotational and of compositional truth. The philosophical background for our discussion in Sections 4.3 and 4.4 is given by Feferman's theory of implicit commitment. However, as we will show, the epistemic notions involved in the investigation of reflection principles presented in Sections 4.2, 4.3, and 4.4 are never made explicit; they are employed only informally in the philosophical argumentation for reflection principles. In Section 4.5 we turn to Cieśliński's formal analysis of the process of reflection on implicit acceptance of a formal theory.

As we will show, in this approach the epistemic notion of acceptance of a theory is made fully explicit via the use of a modal predicate. We will analyse Cieśliński's approach and indicate some problems and questions. We close this chapter with some general philosophical remarks on the nature and role of reflective processes in mathematics.

We try to keep our notation as standard as possible. Concerning proof-theoretical background, we presuppose some familiarity with a few basic formal systems of arithmetic, such as Peano Arithmetic (PA, and its language $\mathcal{L}_{\mathsf{PA}}$) and Elementary Arithmetic (EA). Moreover, although we will present some basic facts about Kleene's notation system $\mathcal{O}$, we will presuppose some familiarity with ordinal notations, the Veblen hierarchy, and related notions. Concerning truth theory, we assume a passing acquaintance with a handful of the main truth theories, such as the compositional theory CT, the Kripke-Feferman system KF, and the Partial Kripke-Feferman system PKF. Nevertheless, for the benefit of readers who are not familiar with these systems, we include footnote references to places where they are defined and discussed.

4.2 Reflection Principles and Progressions of Theories

We concentrate on theories that are formulated in the language of first-order arithmetic or an extension thereof, and at least as strong as Elementary Arithmetic (EA). We are interested in the iteration of proof-theoretic reflection principles over these theories, where a proof-theoretic reflection principle for a given theory S is a formalised soundness statement for S: it expresses that everything provable in S is also true.

By Tarski's theorem of the undefinability of truth, the language of arithmetic does not contain its own truth predicate. So in the language of arithmetic this guiding idea can only be approximated to varying degrees. We can distinguish the following types of reflection principles (for a given theory S):

(i) Con_S (consistency)
(ii) $Prov_S \ulcorner \varphi \urcorner \rightarrow \varphi$ (local reflection)
(iii) $Prov_S \ulcorner \varphi(\dot{x}) \urcorner \rightarrow \varphi(x)$ (uniform reflection)

Here $Prov_S$ is a standard provability predicate for the given theory S. The formula Con_S expresses the consistency of S in terms of $Prov_S$: it can be taken to be the formula

$$Prov_S \ulcorner 0 = 1 \urcorner \rightarrow 0 = 1.$$

Local reflection for a theory S is denoted as Rfn_S, and uniform reflection is denoted as RFN_S. Restricted versions for these principles are also considered: one can consider Rfn_S (RFN_S) for sentences (formulae) of a specific syntactic complexity. Π^0_1-Rfn_S, for instance, is local reflection for the Π^0_1 fragment of S and is equivalent to Con_S.

We can iterate the procedure of adding a reflection principle to a given theory S. For a given theory S and a given reflection principle $\mathcal{R}$ we let $\mathcal{R}[S]$ mean "the reflection principle $\mathcal{R}$ over S". Then we can define the iteration of reflection in the following way by letting:

> $\mathcal{R}^0[S]$ be S;
> for α a successor ordinal, $\mathcal{R}^{\alpha+1}[S]$ be $\mathcal{R}[\mathcal{R}^\alpha[S]]$;
> for λ a limit ordinal, $\mathcal{R}^\lambda[S]$ be the union of all $\mathcal{R}^\alpha[S]$ for $\alpha < \lambda$.

The first proof-theoretic results that we will discuss concern *progressions of theories* generated via iteration (into the transfinite) of reflection principles.

However, before presenting the notion of a progression of theories and the results, we introduce a few notions concerning Kleene's $\mathcal{O}$. We call $|a|$ the ordinal denoted by an ordinal notation a in Kleene's notation system $\mathcal{O}$, which is partially ordered by the relation $<_\mathcal{O}$. We have $a <_\mathcal{O} b$, for two ordinal notations a and b, if and only if $|a| < |b|$.

A *path* P is a subset of $\mathcal{O}$ such that (i) for any a, $b \in P$ either $a \leq_\mathcal{O} b$ or $b \leq_\mathcal{O} a$, (ii) if $b \in P$ and $c \leq_\mathcal{O} b$ then $c \in P$. For any $a \in \mathcal{O}$, a set $P = \{ b \mid b <_\mathcal{O} a \}$ is called a *path within* $\mathcal{O}$. The *length* of a path P is the ordinal of the restriction of $<_\mathcal{O}$ to P. For any path P within $\mathcal{O}$, the order type of P, denoted as $|P|$, is less than ω_1^{CK}. A path P is a path *through* $\mathcal{O}$ if $|P| = \omega_1^{CK}$, where $\omega_1^{CK} = \sup\{|a| : a \in \mathcal{O}\}$. The relation $<_\mathcal{O}$ is not recursively enumerable; indeed, it is Π^1_1-complete. However, for any a, the restriction of $<_\mathcal{O}$ to $\{ b \mid b <_\mathcal{O} a\}$ is recursively enumerable.

Now we introduce the notion of a progression of theories. A progression of a theory S is a primitive recursive mapping taking any ordinal notation a in some path in Kleene's ordinal notation system $\mathcal{O}$ to a Σ^0_1-formula φ_a that recursively enumerates the axioms of a theory S_a, such that

1. $S_0 = S$;
2. $S_{suc(a)} = S_a + \mathcal{R}^a[S]$;
3. $S_{lim(a)} = \bigcup_{b\,<\,a} S_b$.

In words: the starting theory S_0 is just S, the successor stage of the progression is, for any notation a, just the previous theory S_a plus a reflection principle $\mathcal{R}^a$ for S_a, and at limit stages we take unions.

Any transfinite progression yields a *progressive reflection sequence*, which is a sequence of theories of the form

$$S_0, S_1, \ldots S_\omega, S_{\omega+1}, \ldots S_\alpha, \ldots,$$

where $S_{\alpha+1}$ is an extension by reflection of S_α, and S_λ, for limit ordinals λ, has as axioms the union of the axioms of earlier theories.

In the following section we will survey two main results: Turing's completeness theorem for consistency progressions and Feferman's completeness theorem for uniform reflection progressions. Moreover, we will briefly touch upon Feferman's results about autonomous progressions of formal theories.

4.3 Mathematical Reflection

Turing used consistency progressions in an attempt to reduce incompleteness in arithmetic. He proved the following theorem:

THEOREM 4.3.1 (Turing, 1939). For any true Π^0_1 sentence φ there is an $a \in \mathcal{O}$ such that $|a| = \omega + 1$ and $S_a \vdash \varphi$. Moreover, there is a primitive recursive function that associates such an a with each true Π^0_1 sentence φ.

Turing (1939) suggests that the transition from a theory S_a to $S_{suc(a)}$ invokes some sort of reflection:

> We were able, however, from a given system to obtain a more complete one by the adjunction as axioms of formulae, seen intuitively to be correct, but which the Gödel theorem shows are unprovable in the original system; from this we obtained a yet more complete system by a repetition of the process, and so on.
>
> (p. 198)

However, the epistemological import of Turing's completeness theorem is limited. Theorem 4.3.1 only tells us that for any true Π^0_1 sentence φ there is a consistency progression with length $\omega + 1$, such that $S_{\omega + 1}$ proves φ. As Franzén (2004b, Section 6) already pointed out, Turing's result does not provide us with a method of *recognising*, for any true Π^0_1 sentence φ, that it is true. Turing's proof indeed associates with every true Π^0_1 sentence φ a consistency reflection sequence of length $\omega + 1$ that ends in a theory $S_{\omega + 1}$ that proves φ. However, the axioms of S_ω have a non-canonical definition; the trick of Turing's proof consists in defining S_ω in such a way that its consistency entails that φ is true. Even though Turing's clever definition of ω and "canonical" definitions of ω extensionally coincide, no S_n proves that this is so.[1]

Feferman realised that in order to strengthen Turing's completeness result, uniform reflection progressions rather than consistency or local reflection progressions are needed. He proved:

THEOREM 4.3.2 (Feferman, 1962). There is a uniform reflection progression based on PA such that for any true arithmetical sentence φ there is an $a \in \mathcal{O}$ such that $|a| \leq \omega^{\omega^{\omega} + 1}$ with $S_a \vdash \varphi$.[2]

This is known as Feferman's completeness theorem. His proof generates a *path P* within $\mathcal{O}$ of length $\omega^{\omega^{\omega} + 1}$ such that the union of all theories associated with the notations in this path is arithmetically complete.

As with Turing's completeness theorem, and for the same reasons, the epistemological import of Feferman's completeness proof is limited. Following Franzén, we can see that it would be wrong to say that Turing's and Feferman's results show that we will eventually obtain every arithmetical truth by iterating reflection principles.[3]

4.3.1 Autonomous Progressions

The proof of Turing's completeness theorem (and the proof of Feferman's completeness theorem) shows that there is a sense in which progressions as defined in the previous section fail to capture how systems of a higher ordinal level are warranted "from below". For this reason, Kreisel (1958) argued that progressions should satisfy an additional *autonomy* requirement: for every S_a that is in a progression, it should be provable in some S_b with $b <_{\mathcal{O}} a$ that a is in $\mathcal{O}$. A progression that satisfies this additional criterion is called an *autonomous progression*. Before surveying the results about the autonomous progressions, we will introduce briefly the notions of Veblen functions and Veblen hierarchy.

Veblen functions are a hierarchy of normal functions (continuous strictly increasing functions from ordinals to ordinals). If φ_0 is any normal function, then for any ordinal $\alpha > 0$, φ_α is the function enumerating the common fixed points of φ_β for $\beta < \alpha$. These functions are all normal. In the special case when $\varphi_0(\alpha) = \omega^\alpha$ this family of functions is known as the Veblen hierarchy. The function φ_1 is the same as the ε function: $\varphi_1(\alpha) = \varepsilon_\alpha$. The first ε ordinal number ε_0 is $\sup\{1, \omega, \omega^2, \ldots, \omega^\omega, \ldots, \omega^{\omega^{\omega^{\cdots}}}\}$ and is the least fixed point of φ_0, so that $\omega^\alpha = \alpha$. And then $\varphi_2(0)$ is the least ordinal α, such that $\varepsilon_\alpha = \alpha$.

The Feferman–Schütte ordinal Γ_0 can be defined as the smallest ordinal that cannot be obtained by starting with 0 and using the operations of ordinal addition and the Veblen functions $\varphi_\alpha(\beta)$. That is, it is the smallest α such that $\varphi_\alpha(0) = \alpha$. Feferman (1964) and Schütte (1964, 1965) investigated autonomous progressions of predicative theories of analysis. In particular, Feferman (1964) investigated autonomous progressions via uniform reflection based on the systems H and R,[4] determining the *limit* of predicative reasoning. In a nutshell, he showed that the ordinal Γ_0 is the least ordinal that "cannot be reached" predicatively. Or in other words, it is the least ordinal greater than all autonomous a in the progression.[5]

But one can also consider autonomous reflection progressions over first-order arithmetic. The following is a typical result, which is apparently "folkore":[6]

THEOREM 4.3.3. The autonomous uniform reflection progression based on Peano Arithmetic is the first-order fragment of the system of Ramified Analysis up to (but not including) level ω, and the length of this progression is $\varphi_2(0)$.

These theorems are epistemologically more significant than the completeness theorems of Turing and Feferman. In contrast to the non-autonomous progressions, the autonomy condition assures that *we recognise* by means of a proof in a previous stage of the progression that for a limit a, a is an ordinal notation. In this sense, propositions such as Theorem 4.3.3 show *what we can come to know* in reflection progressions. Of course a strong idealisation is involved here: *we* are only able to go through a finite number of stages of an autonomous progression before we die.[7]

In this chapter we are interested in the informal notions involved in the transition from a theory S_a to $S_{suc(a)}$, that is, in the addition of the reflection principles. Like Turing, Feferman (1962) claims that the transition from a theory S_a to $S_{suc(a)}$ is obtained via a process of reflection. He states that our acceptance of a reflection principle for our base theory (and iterating this procedure) rests on our pre-theoretic *attitude*:

> In contrast to an arbitrary procedure for moving from A_K to A_{K+1}, a reflection principle provides that the axioms of A_{K+1} shall express a certain *trust* [our emphasis] in the system of axioms A_K.
>
> (p. 261)

We observe that Feferman's appeal to trust differs from Turing's appeal to mathematical intuition; if we look at the previous quote by Feferman, we see that a reflection principle for a theory S does not only express the soundness of S, but has also an *epistemic* component. In later work, Feferman (1991) continued to emphasise in that reflection principles have an epistemic component:

> Gödel's theorems show the inadequacy of single formal systems [for the purpose of formal analysis of mathematical thought]. However at the same time they point to the possibility of systematically generating larger and larger systems whose *acceptability is implicit in acceptance of the starting theory.*
>
> (p. 2, our emphasis)

Feferman here sketches an epistemological route from knowledge of the axioms of a weaker system to knowledge of the axioms of a stronger system. One starts by believing the axioms of a system S. If one's reasons for doing so are good and S is true, then these beliefs amount to knowledge of the axioms of S. When one is in such a situation, one is implicitly committed to reflection principles for S, such as Con_S. By explicitly endorsing such implicit commitments, one can come to accept, and perhaps even to know, the axioms of a stronger system S'.

4.4 Reflecting on Truth

We will now leave reflection over purely arithmetical theories behind, and concentrate on the iteration of reflection principles over theories of truth (and falsity) that are formulated in an expansion of the language of PA or EA with a fresh truth (and falsity) predicate.

4.4.1 Axiomatic Truth Theories

Pioneers of the investigation of proof-theoretic reflection principles pointed out that the concept of *truth* is involved in the concept of reflection:

> By a "reflection principle" for a formal system S we mean, roughly, the formal assertion stating the soundness of S:
>
> > *If a statement φ (in the formalism S) is provable in S then φ is valid.*
>
> (Kreisel and Lévy, 1968, p. 98)

This was regarded as a problem:

> Literally speaking, the *intended* reflection principle cannot be formulated in S itself by means of a single statement. This would require a *truth definition* T_S, with a variable a over (Gödel numbers of, or, simply, over) formulas of S, and a definition of the proof relation $Prov_S(p,\ a)$ (read: p is (the Gödel number of) a proof of a in S). The reflection principle for S would be
>
> $$\forall p \forall a[Prov_S(p,a) \rightarrow T_S(a)].$$
>
> Such a truth definition T_S, does not exist.
>
> (Kreisel and Lévy, 1968, p. 98)

This difficulty can be (and was) circumvented by *approximating* the intended reflection principle by means of the purely arithmetical principles Rfn_S and RFN_S. But this is not the only possible way forward. Instead, a primitive truth predicate T can be added to the language of arithmetic, thus generating the language $\mathcal{L}_T = \mathcal{L}_{\mathsf{PA}} \cup \{T\}$, and new axioms governing the behaviour of the truth predicate can be added to the background arithmetical theory. This is what some proof theorists started to do in the late 1970s. Moreover, the resulting formal systems were related to a philosophical discussion about the function or role of the concept of truth.

One important role for the concept of truth is to express and reason with generalisations over statements. For this purpose, the use of the truth predicate as a device of quotation and of disquotation is essential. This means that Tarski-biconditionals, i.e., formulae of the form $T \ulcorner \varphi \urcorner \leftrightarrow \varphi$, play a pivotal role in truth theory.

A distinction is made between *typed* and *untyped* (or type-free) Tarski-biconditionals. In the typed case, the truth predicate is not itself allowed to occur in φ. If we start with PA as a base theory and add to PA the collection of all *typed* Tarski-biconditionals $T \ulcorner \varphi \urcorner \leftrightarrow \varphi$ for $\varphi \in \mathcal{L}_{\mathsf{PA}}$, the resulting theory is called TB.[8] If one wants to add to PA a collection of *untyped* Tarski-biconditionals, then, in order to avoid the liar paradox, one can either weaken the background logic, or restrict the collection of Tarski-biconditionals and preserve full classical reasoning. One consistent way of weakening the logic that keeps the full Tarski-biconditionals is to work in *Basic De Morgan* logic (*BDM*).[9] The untyped truth theory formulated in *BDM*, where the Tarski-biconditionals are completely unrestricted, is called TS_0 and is discussed in Fischer et al. (2017).

If one wants to preserve classical logic, then there are different options for restricting the Tarski-biconditionals to avoid inconsistency. Here we discuss two such possible restrictions. One possibility is to restrict the Tarski-biconditional scheme to the sentences φ in which the truth predicate only occurs *positively* (i.e., in the scope of an even number of negation symbols). If we add this collection to PA, the resulting truth theory is called PTB.[10] A natural way of extending this theory is to expand the language of the truth theory ($\mathcal{L}_T$) with a primitive *falsity* predicate, thus generating the language $\mathcal{L}_{T,F}$. We then consider the sublanguage $\mathcal{L}^+_{T,F}$, which is obtained by allowing the negation symbol from $\mathcal{L}_{T,F}$ only to prefix atomic arithmetical formulas. Moreover, we consider the truth biconditionals $T \ulcorner \varphi \urcorner \leftrightarrow \varphi$ with φ restricted to $\mathcal{L}^+_{T,F}$, and the falsity biconditionals $F \ulcorner \varphi \urcorner \leftrightarrow \bar{\varphi}$, where $\bar{\varphi}$ is the *dual* of φ. We can define duals recursively: the dual of an atomic arithmetical formula is its negation; the dual of an atomic formula of the form Tt is Ft, and vice versa, the dual of $A \wedge B$ is the disjunction of the dual of A and the dual of B, and so on.[11] PA plus these two collections of biconditionals is called TFB.

4.4.2 Compositionality and Implicit Commitment

The philosophical question now arises whether the *content of the concept of truth* is given by some such collection of Tarski-biconditionals. An affirmative answer to this question is defended, for instance, in Horwich (1990), Halbach (2001), and Horsten and Leigh (2016). This position is called *disquotationalism*, as it asserts that the content of the concept of truth is captured by a relatively simple and natural collection of Tarski-biconditionals, i.e., by a disquotational theory of truth. If disquotationalism is correct, then the concept of truth really is at bottom merely a device for quotation and disquotation, as Quine maintained.

A standard objection against this, which traces back to Davidson, is that truth is *compositional*. According to this view, truth theories should be able to prove intuitive semantic principles, for instance that any conjunction is true if and only if its conjuncts are both true, and so forth. But these compositional truth clauses cannot be derived from a set of Tarski-biconditionals. In this way it seems that disquotationalist views fall short of capturing the content of the concept of truth.

The standard typed compositional truth theory is called CT.[12] The most popular compositional type-free truth theory in classical logic is KF; the most popular type-free compositional truth theory in non-classical logic is PKF.[13] The Davidsonian objection against disquotational truth theories applies to all the theories mentioned above: the message is that compositional typed (type-free) truth outstrips disquotational typed (type-free) truth by proving *core* principles governing the concept of truth that disquotational theories cannot prove. Without further resources, it seems that there is no way out for the disquotationalist.

At this point, reflection principles enter the philosophical debate. The idea is that the compositional principles might be *implicit* in some collection of Tarski-biconditionals and that *reflection* can bridge the gap between disquotational and compositional truth.

This is indeed the case. In the typed context, Halbach observed that iterating uniform reflection over TB twice recovers typed compositional truth (Halbach, 2001, Section 4):

Theorem 4.4.1. $RFN^2[\mathsf{TB}] \vdash \mathsf{CT}$.

This phenomenon extends to the classical type-free context (Horsten and Leigh, 2016, Theorem 7):

Theorem 4.4.2. $RFN^2[\mathsf{TFB}] \vdash \mathsf{KF}$.

Theorem 4.4.2 has to be taken, however, with a grain of salt. Even though the version of KF that is used by Horsten and Leigh (2016) is closely related to the usual formulations of KF (for instance, the version given in Halbach, 2014, Definition 15.2), it is not outright equivalent to them. In *Pos*(KF) (*positive* KF), the version of KF derivable via two iterations of reflection from TFB, the compositional axioms are restricted to the positive fragment of the language, whereas in the case of the usual KF the compositional axioms are completely unrestricted. Therefore, although these two versions of KF are equivalent for the arithmetical part of the language, their truth predicate behaves somewhat differently. In Zicchetti (2020) it has been shown that TFB and the version of KF adopted in Horsten and Leigh (2016), i.e., the version of KF that we obtain in Theorem 4.2.2 *via reflection from the theory* TFB, can be consistently closed under unrestricted rules of *Necessitation* and *Conecessitation* for the truth and falsity predicates to the theory *Pos*(KF)*, whereas the version of KF given in Halbach (2014) is inconsistent with the addition of the two rules.

The recovery of compositionality through reflection also extends to the type-free non-classical context (Fischer et al., 2017, Corollary 1, Section 3.2):

Theorem 4.4.3. $\mathcal{R}^2[\mathsf{TS}_0] \vdash \mathsf{PKF}$,

where the uniform reflection principle $\mathcal{R}$ is formulated as a rule instead of an axiom. The reflection principle used in the proof of Theorem 4.3.3 is the following:

$$\frac{\Rightarrow Prov^*_{\mathsf{TS}_0}\ulcorner\Gamma(\dot{x}) \Rightarrow \Delta(\dot{x}), \Phi(\dot{x}) \Rightarrow \Psi(\dot{x})\urcorner \quad \Gamma(x) \Rightarrow \Delta(x)}{\Phi(x) \Rightarrow \Psi(x)} \quad (\mathcal{R})$$

where the $Prov^*_{\mathsf{TS}_0}$ expresses that the rule from $\Gamma(x) \Rightarrow \Delta(x)$ to $\Phi(x) \Rightarrow \Psi(x)$ is an admissible rule of TS_0.

Again, following the general idea that the acceptance of a theory generates the possibility to accept stronger theories of which the acceptability is implicit in the acceptance of the weaker theory, we can see that, if we commit ourselves to disquotational typed (type-free) truth theories, then we *implicitly* commit ourselves to compositional typed (type-free) truth theories.[14]

However, iterating reflection does not only recover compositional principles from disquotational ones. As it is shown in Leigh (2016, Theorem 1.4, Theorem 1.5, Section 1), iterating the process of reflection also increases the amount of provable transfinite induction.

We fix a natural notation system for ordinals up to and not including Γ_0 that can be presented as an *elementary ordinal notation system* in the sense of Rathjen (1997), and call it **O**. Then both **O** and the ordering relation $\prec$ on ordinals defined by elements of **O** are definable in first-order arithmetic.

Definition 4.4.4 [Transfinite induction]. Let A be a formula.

1. Transfinite induction for A up to any $\alpha < \Gamma_0$, denoted as $TI(A, \alpha)$, is the formula

   $Prog(\lambda x A) \rightarrow A(t)$,

   where t is a notation in **O** for α, and $Prog(\lambda x A)$ states that A is progressive along $\prec$, i.e.,

   $\forall x \in \mathbf{O}[\forall y \prec x A(y/x) \rightarrow A(x)]$.

2. For a language $\mathcal{L}$ and ordinal $\alpha < \Gamma_0$, the schema of transfinite induction up to α, $TI_{\mathcal{L}}(< \alpha)$, is the collection of formulae

   $\{TI(A, \beta) \mid A \in \mathcal{L} \wedge \beta < \alpha\}$.

Definition 4.4.5. For a theory S and an (elementary) ordinal κ, let S^{κ} denote the extension of S by $TI_{\mathcal{L}}(< \kappa)$.

Definition 4.4.6. For a theory S and (elementary) ordinal κ, let $RFN^{\kappa}[S]$ denote the theory $\mathsf{EA} + \kappa$ times iterated uniform reflection over S.

Now suppose that we start from a disquotational theory that is based on the weak arithmetical theory EA instead of on full PA. In particular, let TB_0, TFB_0 be just like TB, TFB, respectively, except that they have EA instead of PA as their arithmetical background component. Then we have (Leigh, 2016, Theorem 1.4):

Theorem 4.4.7. For all $\kappa \in \mathbf{O}$ with $\kappa > 0$:

1. $\mathsf{CT}^{\varepsilon_\kappa} = RFN^{1+\kappa}[\mathsf{TB}_0]$;
2. $\mathsf{KF}^{\varepsilon_\kappa} = RFN^{1+\kappa}[\mathsf{TFB}_0]$.

Moreover, if we look at the consequences of these theories for the restricted language $\mathcal{L}_{\mathsf{PA}}$, then we have the following result (Leigh, 2016, Theorem 6.24):

Theorem 4.4.8. For all $\kappa \in \mathbf{O}$ with $\kappa > 0$:

1. If A is an $\mathcal{L}_{\mathsf{PA}}$-formula provable in $RFN^{1+\kappa}[\mathsf{TB}_0]$, $RFN^{\kappa}[\mathsf{CT}]$, or $\mathsf{CT}^{\varepsilon_\kappa}$, then A is a theorem of $\mathsf{EA} + TI(< \varepsilon_{\varepsilon_\kappa})$.
2. If A is an $\mathcal{L}_{\mathsf{PA}}$-formula provable in $RFN^{1+\kappa}[\mathsf{TFB}_0]$, $RFN^{\kappa}[\mathsf{KF}]$, or $\mathsf{KF}^{\varepsilon\kappa}$, then A is a theorem of $\mathsf{EA} + TI(< \varphi_{\varepsilon\kappa}(0))$.

The situation in the non-classical settings is similar. In Fischer et al. (2017, Proposition 3.3.3) it is shown that two acts of uniform reflection over the theory called *Basic*, which is EA formulated in the language with the truth predicate $\mathcal{L}_T$ with an induction rule for Δ^0_0-formulae and in *BDM* logic,[15] proves the principle of transfinite induction for the language $\mathcal{L}_T$ for all ordinals up to and including ω^{ω}:

Theorem 4.4.9. $\mathcal{R}^2[Basic] \vdash TI_{\mathcal{L}_T}(\omega^{\omega})$.

Iterating reflection into the transfinite proves even more transfinite induction, as it is shown in Fischer et al. (2017, Corollary 3, Subsection 3.3):

Theorem 4.4.10. $\mathcal{R}^{\omega}[Basic] \vdash TI_{\mathcal{L}_T}(< \omega^{(\omega^2)})$.

In other words, transfinitely many iterations of uniform reflection over a non-classical truth theory still proves much less transfinite induction than just two iterations of uniform reflection over classical logic. This is because *Basic* is formulated in the non-classical logic *BDM*. Some interpret this as a defect of (truth) theories in non-classical logic: they cannot reproduce (possibly not even with reflection) the same mathematical reasoning that classical theories offer (Halbach and Nicolai, 2018).

4.4.3 Global Reflection

The reflection principles involved in the theorems that have been discussed so far merely *approximate* the correct way of formalising soundness. This correct way of formalising soundness was already articulated by Kreisel and Lévy (1968):[16] it is the *Global Reflection Principle* (*GRP*), which can be defined as follows:

Definition 4.4.11. The global reflection principle for a theory S, denoted as GRP_S, is the formula

$$\forall x[Sent_S(x) \wedge Prov_S(x) \rightarrow T(x)].$$

From a "typed" perspective on truth, one mark against global reflection is the fact that already one iteration of global reflection over a typed truth theory violates typing. But from a "type-free" perspective, GRP_S may be a plausible way of making the commitment that is implicit in accepting type-free truth theory S explicit.

If we look at theories formulated in non-classical logic such as TS_0, then we get (Fischer et al., 2017, Proposition 1):

Theorem 4.4.12. The uniform reflection principle and the global reflection principle are provably equivalent over TS_0.

Since TS_0 is arithmetically sound when uniform reflection is added, global reflection over TS_0 is likewise sound. Moreover, this procedure can then consistently be repeated. In other words, TS_0 is *coherent* with its implicit commitment.

The situation in classical logic is different. The closure of classical truth theories under *GRP* for the whole language often forces some kind of inconsistency. This can either be outright inconsistency, or what is called *internal inconsistency*, i.e., the existence of a sentence φ, such that it is provable that $T\varphi \wedge \neg\varphi$. In Halbach (2014) it is shown that FS is inconsistent with $GRP_{\mathsf{FS}}[\mathsf{FS}]$; in Fischer et al. (forthcoming, p. 8) it is observed that the standard axiomatisation of KF is internally inconsistent with $GRP_{\mathsf{KF}}[\mathsf{KF}]$.[17] Indeed, KF is internally inconsistent even with GRP_{FOL}, where FOL is first-order logic formulated in $\mathcal{L}$. This phenomenon has been interpreted by some to indicate that standard theories of type-free truth in classical logic are implicitly incoherent.

In our discussion so far, we have taken the implicit acceptance of or commitment to a theory S to be made explicit via the addition (and iteration) of reflection principles. However, in the previous approaches the epistemic notion of acceptance had been only made indirectly explicit via the notions of provability and truth. In what follows, we will discuss a different procedure to make the implicit acceptance of a theory explicit.

4.5 Reflecting on Acceptance

Instead of taking for granted the idea that proof-theoretic reflection principles express trust or acceptance, one might decide to investigate the notion of acceptance of a given theory T directly, with the aim of spelling it out without the help of reflection principles or the concept of truth. In this case, the concept of *accepting a theory T* should be made precise.

An attempt at doing this was made by Galinon (2014), who focusses on the weakest reflection principle: consistency. In his explication of the reflection process, Galinon uses two key principles. The first of these is the *Principle of (first-person) Responsibility*:

> If a rational agent accepts a collection T of propositions, then she must accept "T is acceptable".
>
> (Galinon, 2014, p. 328)

Second, he endorses the following principle:

> A rational agent must accept that if a collection propositions is acceptable, then that collection is coherent.
>
> (Galinon, 2014, p. 325)

Using these principles, Galinon (2014) develops the following argument for the acceptance of consistency statements. Suppose a rational agent unconditionally accepts a mathematical theory T. Then, using the Principle of Responsibility, she must accept "T is acceptable". And from this, using the second principle, the agent is rationally obliged to infer that T is consistent (p. 329).

In this chapter we cannot do justice to the philosophical complexity of the issues that are relevant here, so we restrict ourselves to a brief discussion of one of Galinon's key principles.[18] The Principle of Responsibility seems a demanding requirement. One might wonder if reflecting on one's acceptance of T might not, in some cases, lead one to abandon rather than to accept one's acceptance of T. Of course this does not exclude that there are cases where we reflect on our acceptance of a theory T and *legitimately* conclude that T is acceptable. If that is so, then maybe Galinon and Feferman go too far when they claim that one is *rationally obliged* to accept reflection principles for theories that one accepts. Perhaps the claim should rather be that there are cases where an agent is *rationally permitted* to accept, on the basis of reflecting on a theory T that she already accepts, reflection principles for T.[19]

Cieśliński (2018, 2017) provides an alternative analysis of reflection on one's mathematical beliefs. He first spells out which informal notion of acceptance of S is relevant, and then proposes the following informal understanding of acceptance of S:

> For any sentence φ, if I believed that φ has a proof in S and I had no independent reason to disbelieve φ, then I would be ready to accept φ.
> (Cieśliński, 2018, p. 1087, notation has been adapted to ours)

Cieśliński (2018) provides an axiomatic theory of believability that employs the informal notion of acceptance presented in the quote above. He makes this notion of acceptance of S explicit by extending S to a new theory S^+, which captures the informal notion expressed above. He does this by presenting a theory of *believability*, which extends the theory S that we accept with a fresh predicate $B(x)$ for believability and with axioms that govern its behaviour.

The thought is that when a person reflects on the implicit commitments involved in her acceptance of a theory K, she comes to accept a theory of believability $Bel(K)^-$ over K.[20] Cieśliński explains how this process is structured, and he spells out $Bel[K]^-$ as an axiomatic theory (Cieśliński, 2018, p. 254).

Suppose we start with a theory K, formulated in a language $\mathcal{L}_K$. Let $\mathcal{L}_{K,B} = \mathcal{L}_K \cup \{B\}$. And let KB be the theory which is just like K except that its schemata range over all formulas of $\mathcal{L}_{K,B}$. The theory of believability $Bel[K]^-$ is an extension of KB with the following axioms and rules (Cieśliński, 2018, Definition 13.4.1):[21]

(Ax_1) $\forall \psi \in \mathcal{L}_{K,B}[Prov_{KB}(\psi) \to B(\psi)]$,
(Ax_2) $\forall \varphi, \psi \in \mathcal{L}_{K,B}[(B(\varphi) \wedge B(\varphi \to \psi)) \to B(\psi)]$,

$$(NEC)\ \frac{\vdash \varphi}{\vdash B(\varphi)} \qquad (GEN)\ \frac{\vdash \forall n : B(\varphi(n))}{\vdash B(\forall x \varphi(x))}$$

Let us now apply Cieśliński's general theory to a concrete example. Consider the "weak" typed disquotational truth theory TB^-, which is like the disquotational theory TB except that the truth predicate is not allowed to occur in the induction schema. Suppose that we accept TB^-. Then if we make the acceptance of TB^- explicit via $Bel[\mathsf{TB}^-]^-$, we recover compositional principles for typed truth (Cieśliński, 2018, p. 264):

THEOREM 4.5.1. *Bel* $[\mathsf{TB}^-]^- \vdash B(\mathsf{CT})$,

where $B(\mathsf{CT})$ consists of all sentences $B(\varphi)$ such that φ is an axiom of CT. In particular we obtain the believability of mathematical induction for $\mathcal{L}_T$ from a situation where we only accepted induction for $\mathcal{L}_{\mathsf{PA}}$.

Analogous results hold in type-free settings. Consider the typed disquotational truth theory TFB^-, which is like TFB except that the truth predicate is not allowed to occur in the induction schema. Suppose that we accept TFB^-. Then if we make the acceptance of TFB^- explicit via $Bel[\mathsf{TFB}^{-}]^{-}$, we recover compositional principles for type-free truth (Cieślińskip, 2018, p. 266):

Theorem 4.5.2. $Bel\ [\mathsf{TFB}^{-}]^{-} \vdash B(\mathsf{KF})$.

So, taking stock: if we are committed to typed (type-free) disquotational truth and if this commitment is made explicit via a theory of believability, then this theory proves that the compositional principles for typed (type-free) truth are indeed believable.

The believability theory over the disquotational truth theory does not contain a factivity principle or rule ("B-Out") for the believability predicate B. Indeed, the inference from the believability of a statement to the statement itself is a *defeasible* rule. For this reason, we do not have $Bel[\mathsf{TB}^{-}]^{-} \vdash \mathsf{CT}$. Nonetheless, according to Cieśliński's informal definition of acceptance of a theory, this then means that, in the absence of independent reasons for disbelieving compositional principles of typed (type-free) truth, we should be ready to accept them. In this sense Cieśliński's results provide and argument for the thesis that our commitment to compositional truth principles is not greater than the commitment to disquotational truth principles.

It would take us too far to give a detailed evaluation of Cieśliński's position, so again we confine ourselves to a few cautiously critical remarks. Cieśliński argues that processes of reflection on one's acceptance of a theory K can be described as proofs in a believability theory $Bel[K]^{-}$ for K. But it is not clear that all principles of $Bel[K]^{-}$ are in all circumstances correct. In particular, for the same reasons as why Galinon's Principle of Responsibility might not in all cases be correct, it is not clear that axiom Ax_1 of $Bel[K]^{-}$ is always true. Might there not be circumstances where the agent starts out by accepting K, but by reflecting on K comes to abandon parts of K—perhaps because in the reflective process she comes to realise that K is actually quite strong—rather than to judge that K is believable? It seems to us that a deeper phenomenological analysis of reflection processes than has been given thus far is needed to decide this question.[22]

4.6 Reflective Processes

The reflection principles that we have discussed in the previous sections take the form of conditional statements. These conditional statements express the result of *reflective processes*, which have an argumentative structure. They aim systematically to draw out consequences from hypothetical situations. The resulting formal reflection principles intend to express a necessary connection between the "input" of a reflection process and the "output" of that process.

Because of this, reflection principles have played a role in debates in the foundations of mathematics about the justification of mathematical theories. However, the extent to which proof theoretic reflection principles can play a justificatory role in this context, is contested. On the one hand, Horsten and Leigh (2016) argue that if accepting a theory S is justified, then accepting a proof-theoretic reflection principle for S is also epistemically warranted.[23] On the other hand, Dean (2014) urges caution. He argues that even in a context where accepting a theory S is justified, justification for proof-theoretic reflection principles for S must be obtained before we are warranted to accept them. Getting to the bottom of this requires deeper philosophical reflection on the nature of proof-theoretic reflection than has been carried out so far. Indeed, we believe that reflection processes that underpin formal reflection principles deserve more attention from philosophers of mathematics than has hitherto been accorded to them.

In this chapter we have concentrated on reflection principles that are connected with reflective processes that start from hypothetical facts about provability in a formal system. Some such reflective processes terminate in propositions that attribute truth to statements (Section 4.4); others terminate in propositions about rational believability (Section 4.5). However, there exists a class of reflection principles that are related to reflective processes that do not terminate in, but rather start from, hypothetical propositions that attribute truth to statements. Such principles are called *set theoretic reflection principles*.[24]

It can be argued that proof-theoretic reflection principles are related to set theoretic reflection principles.[25] Consider, for instance, local reflection for a theory S. For theories S that prove the completeness theorem, Rfn_S is equivalent to the scheme

$$\varphi \to \exists M : M \models S + \varphi,$$

which is a set theoretic reflection principle.[26] Of course this principle is so weak that it is hardly mentioned in discussions of set theoretic reflection. Indeed, the weakest set theoretic reflection principle that is widely discussed is Montague-Levy reflection. The Montague-Levy reflection principle is provable in ZFC. Nonetheless, the fact that it has proof-theoretic strength is shown by the fact that over the remaining axioms of ZFC, it is equivalent to the axiom of infinity plus the axiom of replacement.

It is commonly assumed that “set theoretic reflection principles can be very strong, but proof-theoretic reflection principles are always weak”. But in an absolute sense, this is not quite correct, as can be seen as follows.[27] The axiom MC, which expresses that there exists a measurable cardinal, can be expressed as an embedding principle (the existence of a non-trivial embedding from Gödel’s L to L). And such embedding principles are often (but not always) informally described as set theoretic reflection principles. But even though $\mathsf{ZFC} + \mathsf{MC}$ proves the consistency of ZFC, it is easy to see that $\mathsf{ZFC} + \mathsf{MC} \nvdash \mathsf{ZFC} + Rfn_{\mathsf{ZFC}}$. So there is a sense in which even local reflection is strong.

The discussion of set theoretic reflection principles falls outside the scope of this chapter. The same holds for the discussion of the nature of our epistemic warrant for set theoretic reflection principles. We restrict ourselves here to observing that it should not automatically be assumed that our epistemic warrant for even moderately strong set theoretic reflection principles is of the same nature as our warrant for proof theoretic reflection principles. We have seen that our warrant for a proof theoretic reflection principle for a theory *S* is often taken somehow to be implicit in our warrant for *S*. But it is hard to see how something like this might be true for set theoretic reflection principles, since even the modest ones (such as Montague-Levy reflection) make no explicit reference to a background theory.

Acknowledgments

Thanks to Kentaro Fujimoto, Karl-Georg Niebergall, and the editors of this book for valuable comments on an early version of this chapter. The second author also thanks the South, West & Wales Doctoral Training Partnership for (grant reference AH/L503939/1) the financial and ideal support for his research project *Theories of Truth and Foundations for Mathematics: Epistemic Warrants and Reflection Principles*.

Notes

1. For more on the philosophical significance of the use of non-canonical definitions, see Franzén (2004a,b).
2. Feferman's completeness theorem can be strengthened. Using the notion of *smooth progression* developed in Beklemishev (1995) it can be shown that the length of this path can be shortened to ω^{ω^2+1}. For an idea of the proof of this improvement, see Franzén (2004b).
3. It is also known that completeness depends on the choice of the path in $\mathcal{O}$. Feferman and Spector (1962) showed for instance that there are paths *through* $\mathcal{O}$, such that corresponding uniform reflection progression does not even prove every true Π^0_1 sentence.
4. *H* is the extension of first-order *Peano Arithmetic*, PA, with Kreisel's *hyperarithmetic comprehension rule* (*HCR*): see Fererman (1964, p. 17) for Feferman's original formulation of the system *H* and of *HCR*. *R* is a system of *Ramified analysis*: see Feferman (1964, pp. 21–22).
5. See Feferman (1964, p. 23, Theorem 6.10) for Feferman's original formulation of the theorem.
6. The claim has been made in Feferman (1964). Thanks to Kentaro Fujimoto for pointing this out to us.
7. For a discussion of the role of idealisation in the epistemological discussion of transfinite progressions of formal theories, see Antonutti Marfori and Horsten (2019).
8. In TB the induction scheme is extended to allow also formulae that contain the truth predicate.
9. Of course there are also other non-classical logics that one can opt for, such as Strong, Weak Kleene Logic, etc. For background on these non-classical logics, see for instance Priest (2008).

10. See Halbach (2014, Section 19.3).
11. See Leigh (2016, Section 5).
12. See Halbach (2014, chapter 8).
13. See Halbach and Horsten (2006) and Halbach (2014, chapters 15, 16).
14. Although, as we pointed out, in the classical case a restricted version of compositionality is obtained, starting with positive biconditionals.
15. See Fischer et al. (2017, Section 2.2) for more details.
16. See Section 4.4.1 above.
17. No claim of originality for this result is made in this chapter. Indeed, this elementary observation is folklore.
18. Galinon argues for the Principle of Responsibility on the basis of norms of rationality (Galinon, 2014, Section 7), and he argues for the second principle on the basis of a "Gödelian Dutch book argument" (Galinon, 2014, Section 5).
19. This stance is taken in Fischer et al. (forthcoming).
20. Cieśliński also considers a believability theory $Bel(K)$ over K that is stronger than $Bel(K)^{-}$. We do not discuss this stronger theory $Bel(K)$ here.
21. In the interest of readability we are sloppy with the Gödel coding in what follows.
22. An attempt to provide such an analysis is given in (Horsten, forthcoming).
23. In this connection, see also Fischer et al. (forthcoming).
24. In the literature on predicativity, reflection principles are considered that take facts about *definability* as input: see Lorenzen (1958). Discussion of these principles falls outside the scope of this chapter.
25. Kreisel and Levy are undecided whether proof theoretic and set theoretic reflection are related: see Kreisel and Lévy (1968, p. 101).
26. Thanks to Kentaro Fujimoto for putting it this way.
27. Thanks to Karl-Georg Niebergall for pointing this out to us.

References

Antonutti Marfori, M. and Horsten, L. (2019). Human-effective computability. *Philosophia Mathematica*, 27(1): 61–87.

Beklemishev, L. (1995). Iterated local reflection versus iterated consistency. *Annals of Pure and Applied Logic*, 75(1): 25–48.

Cieśliński, C. (2017). *The Epistemic Lightness of Truth. Deflationism and its Logics*. Cambridge University Press.

Cieśliński, C. (2018). Minimalism and the generalisation problem: On Horwich's second solution. *Synthese*, 195: 1077–1101.

Dean, W. (2014). Arithmetical reflection and the provability of soundness. *Philosophia Mathematica*, 23(1): 31–64.

Feferman, S. (1962). Transfinite recursive progressions of axiomatic theories. *The Journal of Symbolic Logic*, 27(3): 259–316.

Feferman, S. (1964). Systems of predicative analysis. *The Journal of Symbolic Logic*, 29(1): 1–30.

Feferman, S. (1991). Reflecting on incompleteness. *The Journal of Symbolic Logic*, 56(1): 1–49.

Feferman, S. and Spector, C. (1962). Incompleteness along paths in progressions of theories. *Journal of Symbolic Logic*, 27(4): 383–390.

Fischer, M., Nicolai, C., and Horsten, L. (2017). Iterated reflection overfull disquotational truth. *Journal of Logic and Computation*, 27(8): 2631–2651.

Fischer, M., Nicolai, C., and Horsten, L. (forthcoming). Hypathia's silence. Truth, justification, and entitlement. *Noûs*.

Franzén, T. (2004a). *Inexhaustibility: A Non-exhaustive Treatment*. Association of Symbolic Logic.

Franzén, T. (2004b). Transfinite progressions: A second look at completeness. *The Bulletin of Symbolic Logic*, 10(3): 367–389.

Galinon, H. (2014). Acceptation, cohérence et responsabilité. In *Liber Amicorum Pascal Engel*. J. Dutant, D. Fassio, and A. Meylan, editors, Université de Genève.

Halbach, V. (2001). Disquotational truth and analyticity. *Journal of Symbolic Logic*, 66(4): 1959–1973.

Halbach, V. (2014). *Axiomatic Theories of Truth*. Cambridge University Press.

Halbach, V. and Horsten, L. (2006). Axiomatizing Kripke's theory of truth. *Journal of Symbolic Logic*, 71(2): 677–712.

Halbach, V. and Nicolai, C. (2018). On the costs of nonclassical logic. *Journal of Philosophical Logic*, 47: 227–257.

Horsten, L. (forthcoming). On reflection. *Philosophical Quarterly*.

Horsten, L. and Leigh, G. E. (2016). Truth is simple. *Mind*, 126(501): 195–232.

Horwich, P. (1990). *Truth*. Clarendon Press.

Kreisel, G. (1958). Ordinal logics and the characterization of informal concepts of proof. In *Proceedings of the International Congress of Mathematicians* (1958), pages 289–299. J. A. Todd, editor, Cambridge University Press, 1960.

Kreisel, G. and Lévy, A. (1968). Reflection principles and their use for establishing the complexity of axiomatic systems. *Mathematical Logic Quarterly*, 14: 97–142.

Leigh, G. E. (2016). Reflecting on truth. *IFCoLog Journal of Logics and their Applications*, 3: 557–593.

Lorenzen, P. (1958). Logical reflection and formalism. *The Journal of Symbolic Logic*, 23(3): 241–249.

Priest, G. (2008). *An Introduction to Non-Classical Logic From If to Is*. Cambridge University Press.

Rathjen, M. (1997). The realm of ordinal analysis. In Cooper, S. B. and Truss, J. K., editors, *Sets and Proofs*, pages 219–279. Cambridge University Press.

Schütte, K. (1964). Eine Grenze für die Beweisbarkeit der transfiniten Induktion in der verzweigten Typenlogik. *Archiv für Mathematische Logik und Grundlagenforschung*, 7: 45–60.

Schütte, K. (1965). Predicative well-orderings. In Crossley, J. and Dummett, M., editors, *Formal Systems and Recursive Functions*, volume 40 of Studies in Logic and the Foundations of Mathematics, pages 280–303. Elsevier.

Turing, A. M. (1939). Systems of logic based on ordinals. *Proceedings of the London Mathematical Society*, s2–45(1): 161–228.

Zicchetti, M. (2020). *Truth, Trustworthiness and Reflection*. Submitted for publication.

5 The Expressive Power of Contextualist Truth

Julien Murzi and Lorenzo Rossi

The truth predicate is often argued to be *naïve*, in the sense a sentence φ and its truth-ascription 'φ is true', in symbols $\mathsf{Tr}(\ulcorner\varphi\urcorner)$, are in some way equivalent (in all non-opaque contexts).[1] Some authors require that φ and $\mathsf{Tr}(\ulcorner\varphi\urcorner)$ be interderivable (from possibly open assumptions), and thus obey the following introduction and elimination rules:[2]

$$\frac{\varphi}{\mathrm{Tr}(\ulcorner\varphi\urcorner)}\,\text{Tr-I} \qquad \frac{\mathrm{Tr}(\ulcorner\varphi\urcorner)}{\varphi}\,\text{Tr-E}$$

Other authors go further and demand that the truth predicate obey the full T-Schema:

$$\mathsf{Tr}(\ulcorner\varphi\urcorner) \leftrightarrow \varphi$$

or the so-called *transparency* requirement, according to which any sentence φ is intersubstitutable *salva veritate* with $\mathsf{Tr}(\ulcorner\varphi\urcorner)$, in all non-opaque contexts.

It has been forcefully argued that naïveté is required in order to account for crucial and non-eliminable uses of the truth predicate, involving *blind ascriptions* (such as 'Everything Lois says is true'), *infinitary generalizations* (such as 'All theorems of Peano Arithmetic are true'), and their combinations (Field, 2008; Beall, 2009; Horsten, 2012). Call this argument the *Argument for Naïveté*. Since all forms of naïveté are incompatible with classical logic (given a modicum of syntax theory, the argument can be taken to establish that, in order to fulfil its inferential role, the truth predicate requires a suitable non-classical logic.

To be sure, the *Argument for Naïveté* does not settle which form of naïveté is best suited to underwrite the expressive role of 'true'. Some authors (see e.g. Field, 2008 and Beall, 2009) argue that naïve rules such as Tr-I and Tr-E do not suffice to model reasonings involving blind ascriptions and infinitary generalizations in certain *embedded contexts*, such as the antecedent of a conditional claim. For this reason, they propose stronger forms of naïveté, such as the T-Schema (in the presence of a suitably strong logic of the conditional), or transparency. Call this the *Argument for Transparency*.

In this chapter, we argue that both the *Argument for Transparency* and the *Argument for Naïveté* are fundamentally misguided. We first show, *contra* Field, that the expressive role of 'true' does not require transparency: principles such as $\mathsf{Tr\text{-}I}$ and $\mathsf{Tr\text{-}E}$ are indeed sufficient. However, we also argue, *contra* an argument offered by Picollo and Schindler (2018) that principles such as $\mathsf{Tr\text{-}I}$ and $\mathsf{Tr\text{-}E}$ are necessary: theories that do not validate similar principles irredeemably cripple the expressive role of 'true' (§5.1). We then point to some fully classical, non-naïve theories of truth featuring versions of $\mathsf{Tr\text{-}I}$ and $\mathsf{Tr\text{-}E}$ that are strong enough to validate arguments involving blind ascriptions and infinitary generalizations and yet weak enough to avoid paradox-driven triviality. In particular, we argue that *contextualist* theories of truth, despite being fully classical, feature expressively adequate, *context-shifting* versions of the naïve truth introduction and elimination rules (§5.2). We conclude that the expressive role of the truth predicate requires neither transparency nor naïveté.

5.1 The Inferential Role of Truth and Non-Classical Logic

We begin by considering simple, unembedded cases of agreement and disagreement (§5.1.1) and then move to more complex, embedded such cases, and to Field's Argument for Transparency (§5.1.2). We suggest that unembedded and embedded cases alike provide strong evidence for the interderivability of φ and 'φ is true', at least for languages, such as English, that are rich enough to express so-called contingent liars. However, we also argue that, pace Field, embedded cases of agreement and disagreement fail to establish that the truth predicate ought to be 'transparent'—i.e. transparency is not necessary for the inferential role of 'true'.

5.1.1 Simple Agreement and Disagreement

It is often argued that simple cases of agreement and disagreement motivate the adoption of naïve introduction and elimination rules for 'true' and, in turn, of a non-classical theory of truth. But this argument is itself naïve, or so the classical theorist might argue.

The Argument for Naïveté

Consider the following argument—call it Agreement:

(1) All the theorems of Peano Arithmetic are true;
(2) If $\ulcorner\varphi\urcorner$ is a theorem of Peano Arithmetic, then $\mathsf{Tr}(\ulcorner\varphi\urcorner)$;

(3) $\ulcorner\varphi\urcorner$ is a theorem of Peano Arithmetic;
(4) $\mathsf{Tr}(\ulcorner\varphi\urcorner)$;
(5) φ.

Here, thanks to the truth-predicate, we move from an expression of agreement with a given set of claims to the actual assertion of one of these claims. The steps from (1) to (4) seem unassailable: (1) and (3) are assumptions, (2) follows from (1) by universal instantiation, and (4) follows from (2) and (3) by *modus ponens*. To be sure, principles such as universal instantiation and *modus ponens* might be called into question (see e.g. McGee, 1985). However, given our focus on truth, we bracket aside any non-truth-theoretic qualms one might have with classical rules. Then, the only potentially suspicious step is the one from (4) to (5), which employs Tr-E.

A converse argument illustrates how one can move from an expression of disagreement with a given set of claims to the actual assertion of the negation of one of those claims—call this DISAGREEMENT:

(6) Everything Lois said yesterday is not true;
(7) If Lois said $\ulcorner\psi\urcorner$ yesterday, then $\neg\mathsf{Tr}(\ulcorner\psi\urcorner)$;
(8) Yesterday Lois said $\ulcorner\psi\urcorner$;
(9) $\neg\mathsf{Tr}(\ulcorner\psi\urcorner)$;
(10) $\neg\psi$.

As above, the steps from (6) to (9) appear completely unproblematic: (6) and (8) are assumptions, (7) follows from (6) by universal instantiation, and (9) follows from (7) and (8) by *modus ponens*. Again, the only potentially controversial step is the one from (9) to (10), which employs Tr-I (given contraposition).

Speeches such as AGREEMENT and DISAGREEMENT *prima facie* suggest that the expressive role of 'true' requires Tr-I and Tr-E. After all, the argument goes, it is difficult to see how these reasonings can be run without these principles. Without Tr-E, one can at best use AGREEMENT to establish $\mathsf{Tr}(\ulcorner\varphi\urcorner)$ from the premises that the theorems of Peano Arithmetic are true and that φ is a theorem of Peano Arithmetic. But this falls short of establishing φ. Similar considerations hold for DISAGREEMENT. In order to fulfil its expressive role, the truth predicate must obey both Tr-I and Tr-E and, for this reason, any expressively adequate theory of truth must be non-classical, or so non-classical theorists argue (Field, 2008; Beall, 2009; Horsten, 2012).

Contingent Liars

It might be objected that the argument at best establishes the validity of *certain instances* of Tr-E and Tr-I—namely, those occurring in compelling instances of Agreement and Disagreement. For instance, consider the following (non-schematic) instance of Agreement:

(1) All the theorems of Peano Arithmetic are true;
$(2^\star)$ If $\ulcorner 2+2=4 \urcorner$ is a theorem of Peano Arithmetic, then $\mathsf{Tr}(\ulcorner 2+2=4 \urcorner)$;
$(3^\star)$ $\ulcorner 2+2=4 \urcorner$ is a theorem of Peano Arithmetic;
$(4^\star)$ $\mathsf{Tr}(\ulcorner 2+2=4 \urcorner)$;
$(5^\star)$ $2+2=4$.

And consider the following (non-schematic) instance of Disagreement:

(6) Everything Lois said yesterday is not true;
$(7^\star)$ If Lois said $\ulcorner$grass is red$\urcorner$ yesterday, then $\neg\mathsf{Tr}(\ulcorner$grass is red$\urcorner)$;
$(8^\star)$ Yesterday Lois said $\ulcorner$grass is red$\urcorner$;
$(9^\star)$ $\neg\mathsf{Tr}(\ulcorner$grass is red$\urcorner)$;
$(10^\star)$ It is not the case that grass is red.

Both of the above arguments arguably provide good evidence for the corresponding instances of Tr-E and Tr-I. However, the classical theorist might insist, such instances fall short of justifying the *schematic* arguments Agreement and Disagreement—and hence fall short of justifying *full* Tr-E and Tr-I and the consequent abandonment of classical logic. There are two related reasons for this.

For one thing, virtually every classical theory of truth will allow one to infer $2+2=4$ from $\mathsf{Tr}(\ulcorner 2+2=4 \urcorner)$, or $\neg(0=1)$ from $\neg\mathsf{Tr}(\ulcorner 0=1 \urcorner)$.[3,4] For another, virtually every classical theory of truth will disallow applications of Tr-I and Tr-E to *paradoxical* sentences, such as Liar and Curry sentences. However, and this is the crucial point, instances of Agreement and Disagreement involving such sentences do not constitute convincing evidence to accept Tr-I and Tr-E, or at least so the classical theorist might argue.[5] To see this, let λ be a Liar sentence (that is, a sentence equivalent to its own negated truth-ascription $\neg\mathsf{Tr}(\ulcorner \lambda \urcorner)$), and consider the following instance of Disagreement:

(6) Everything Lois said yesterday is not true;
$(7^{\star\star})$ If Lois said $\ulcorner \lambda \urcorner$ yesterday, then $\neg\mathsf{Tr}(\ulcorner \lambda \urcorner)$;
$(8^{\star\star})$ Yesterday Lois said $\ulcorner \lambda \urcorner$;
$(9^{\star\star})$ $\neg\mathsf{Tr}(\ulcorner \lambda \urcorner)$;
$(10^{\star\star})$ $\neg\lambda$.

Classical theorists might insist that is not at all clear whether it is desirable to infer $\neg\lambda$, that is $\mathsf{Tr}(\ulcorner \lambda \urcorner)$, from $\neg\mathsf{Tr}(\ulcorner \lambda \urcorner)$. As a result, it is not at all clear whether the corresponding instance of Tr-I (which is required to move from $(9^{\star\star})$ to $(10^{\star\star})$) is justified.

Putting the two halves of her argument together, the classical theorist can claim that she can recover all the uncontroversial instances of Agreement and Disagreement while insisting that the ones involving paradoxical sentences fail to justify *full* $\mathsf{Tr\text{-}I}$ and $\mathsf{Tr\text{-}E}$. For instance, the classical theorist might insist that she only accepts instances of argument forms such as Agreement and Disagreement that are clearly safe—e.g. instances involving sentences involving no semantic vocabulary, or truth iterations of such sentences.

The foregoing rejoinder, though, presupposes that it is always clear whether an instance of $\mathsf{Tr\text{-}E}$ or $\mathsf{Tr\text{-}I}$ is paradoxical, or unsafe. But this is actually not the case. As Kripke (1975) famously pointed out, whether a sentence is paradoxical can depend on contingent facts. For instance, suppose Lois says that what the lady on TV with the red dress says is not true and that, as a matter of fact, *she* is the lady on TV with the red dress. Then, Lois' utterance is paradoxical, but, one can persuasively argue, the corresponding versions of Agreement and Disagreement are perfectly acceptable, as is arguably shown by the following instance of Disagreement:

(6) Everything Lois says is not true;
(7^{***}) If Lois says $\ulcorner$what the lady on TV with the red dress says is not true$\urcorner$, then $\neg\mathsf{Tr}(\ulcorner$what the lady on TV with the red dress says is not true$\urcorner)$;
(8^{***}) Lois says $\ulcorner$what the lady on TV with the red dress says is not true$\urcorner$;
(9^{***}) $\neg\mathsf{Tr}(\ulcorner$what the lady on TV with the red dress says is not true$\urcorner)$;
(10^{***}) It is not the case that what the lady on TV with the red dress says is not true.

Now, whether 'what the lady on TV with the red dress says is not true' is paradoxical depends on the contingent facts, i.e. on whether the speaker *is* the lady on TV with the red dress. But, then, there are countless perfectly harmless instances of (6)–(10^{***}). And, it seems plausible to maintain, it would be unduly restrictive to disallow them *all*. As Field puts it, this would irredeemably 'cripple' ordinary reasoning. It would seem, then, that the inferential role of the truth predicate indeed requires, just like non-classical theorists maintain, that 'true' satisfies principles that validate the interderivability of φ and the claim that φ is true, such as $\mathsf{Tr\text{-}E}$ and $\mathsf{Tr\text{-}I}$.

Field further argues that embedded uses of 'true' (as in the antecedent of conditionals) actually require that truth be *transparent*, i.e. that the truth predicate satisfies the intersubstitutivity of $\mathsf{Tr}(\ulcorner\varphi\urcorner)$ and φ in all non-opaque contexts. We disagree. Before we say why, we first review Field's *Argument for Transparency*.

5.1.2 Embedded Truth-Ascriptions and the Argument for Transparency

We present Field's *Argument for Transparency* and argue that it misses its target: the embedded uses of 'true' Field points to can be adequately modelled by Tr-E and Tr-I. We then consider a recent argument by Lavinia Picollo and Thomas Schindler, which can be interpreted as showing that Tr-E is sufficient to model the expressive uses of 'true', and find it wanting.

Field's Argument

According to Field (2008), more complex arguments involving embedded truth-ascriptions show that the truth predicate must obey full transparency, and not merely Tr-E and Tr-I. He writes:

> Talk of truth isn't just a means of expressing agreement and disagreement, for the same reason that talk of goodness isn't just a means of expressing approval and disapproval: 'true', like 'good', occurs in *embedded contexts* (contexts embedded more deeply than a negation). In particular, 'true' is used inside *conditionals*. And in order for it to serve its purpose, it needs to be well-behaved there: inside conditionals as in unembedded contexts, 'true' needs to serve as a device of infinite conjunction or disjunction.... Suppose I can't remember exactly what was in the Conyers report on the 2004 election, but say.
>
> (11) If everything that the Conyers report says is true, then the 2004 election was stolen.
>
> Suppose that what the Conyers report says is $\varphi_1, \ldots, \varphi_n$. Then relative to this last supposition, (11) better be equivalent to
>
> (12) If $\varphi_1, \ldots, \varphi_n$, then the 2004 election was stolen.
>
> And this requires True($\ulcorner\varphi\urcorner$) to be intersubstitutable with φ even when φ is the antecedent of a conditional.
>
> (Field, 2008, pp. 109–110; the original numbering has been adapted to ours)

However, we argue, *pace* Field, embedded truth-ascriptions do not show that truth has to be fully transparent in order to fulfil its inferential role.

Tr-E *and* Tr-I *Suffice*

Let C($\ulcorner\varphi\urcorner$) and σ be shorthand for, respectively, 'The Conyers report says that φ' and 'The election was stolen'. Let χ be the conjunction of all the claims in the Conyers report. Consider the following two sentences:

(13) $\forall x(\mathsf{C}(x) \to \mathsf{Tr}(x)) \to \sigma$;
(14) $\chi \to \sigma$.

We now show that (13) and (14) are interderivable via Tr-E and Tr-I, together with $\mathsf{C}(\ulcorner\chi\urcorner)$ and suitable assumptions on how to formalise the sentence concerning what the report says.

We first show that (13) follows from (14) and $\mathsf{C}(\ulcorner\chi\urcorner)$, given Tr-E:

$$\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{}{\forall x[\mathsf{C}(x) \to \mathsf{Tr}(x)]}{\scriptstyle 1}}{\mathsf{C}(\ulcorner\chi\urcorner) \to \mathsf{Tr}(\ulcorner\chi\urcorner)}{\scriptstyle \forall\text{-E}} \qquad \mathsf{C}(\ulcorner\chi\urcorner)}{\mathsf{Tr}(\ulcorner\chi\urcorner)}{\scriptstyle \to\text{-E}}}{\chi}{\scriptstyle \mathsf{Tr}\text{-E}} \qquad \chi \to \sigma}{\sigma}{\scriptstyle \to\text{-E}}}{\forall x[\mathsf{C}(x) \to \mathsf{Tr}(x)] \to \sigma}{\scriptstyle \to\text{-I},\ 1}$$

We then establish that (14) follows from (13), given Tr-I and $\forall x(\mathsf{C}(x) \leftrightarrow x = \ulcorner\varphi_1\urcorner \vee \ldots \vee x = \ulcorner\varphi_n\urcorner)$, which expresses the fact that the Conyers report says exactly that χ. Let ζ be a shorthand for $\mathsf{C}(x) \leftrightarrow x = \ulcorner\varphi_1\urcorner \vee \ldots \vee x = \ulcorner\varphi_n\urcorner$. We can then reason thus:

$$\dfrac{\dfrac{\forall x[\mathsf{C}(x) \to \mathsf{Tr}(x)] \to \sigma \qquad \dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\forall x(\zeta)}{\zeta}{\scriptstyle \forall\text{-E}} \quad \dfrac{}{\mathsf{C}(x)}{\scriptstyle 2}}{x = \varphi_1 \vee \cdots \vee x = \varphi_n}{\scriptstyle \leftrightarrow\text{-E}} \qquad \dfrac{\dfrac{}{x = \varphi_i}{\scriptstyle 1} \quad \dfrac{\dfrac{\dfrac{}{\varphi_1 \wedge \ldots \wedge \varphi_n}{\scriptstyle 3}}{\varphi_i}{\scriptstyle \wedge\text{-E}}}{\mathsf{Tr}(\ulcorner\varphi_i\urcorner)}{\scriptstyle \mathsf{Tr}\text{-I}}}{\mathsf{Tr}(x)}{\scriptstyle =\text{-E}}}{\mathsf{Tr}(x)}{\scriptstyle \vee\text{-E},\ 1}}{\mathsf{C}(x) \to \mathsf{Tr}(x)}{\scriptstyle \to\text{-I},\ 2}}{\forall x[\mathsf{C}(x) \to \mathsf{Tr}(x)]}{\scriptstyle \forall\text{-I}}}{\sigma}{\scriptstyle \to\text{-E}}}{(\varphi_1 \wedge \ldots \wedge \varphi_n) \to \sigma}{\scriptstyle \to\text{-I},\ 3}$$

The line labelled '$\vee$-E, 1' is a condensed way to indicate *n*-many uses of the rule of disjunction elimination (and the corresponding assumption discharges). Writing this step in full, we should have first applied disjunction elimination to $x = \ulcorner\varphi_1\urcorner \vee \ldots \vee x = \ulcorner\varphi_n\urcorner$, then to $x = \ulcorner\varphi_2\urcorner \vee \ldots \vee x = \ulcorner\varphi_n\urcorner$, and so on, until we reach $x = \ulcorner\varphi_{n-1}\urcorner \vee x = \ulcorner\varphi_n\urcorner$. We've omitted the full steps and abused notation in the manner indicated for readability's sake.

The situation is exactly parallel for cases of embedded disagreement. Letting υ be the disjunction of all the claims in the Conyers report (so that $\neg\upsilon$ amounts to the conjunction of all the negated claims in the report), consider the two following sentences:[6]

(15) $\forall x(\mathsf{C}(x) \to \neg\mathsf{Tr}(x)) \to \neg\sigma$;
(16) $\neg\upsilon \to \neg\sigma$.

We first show that (15) follows from (16) and $\mathsf{C}(\ulcorner \upsilon \urcorner)$, given Tr-I:[7]

$$
\begin{prooftree}
\AxiomC{}
\RightLabel{3}
\UnaryInfC{$\forall x[\mathsf{C}(x) \to \neg\mathsf{Tr}(x)]$}
\RightLabel{$\forall$-E}
\UnaryInfC{$\mathsf{C}(\ulcorner \upsilon \urcorner) \to \neg\mathsf{Tr}(\ulcorner \upsilon \urcorner)$}
\AxiomC{$\mathsf{C}(\ulcorner \upsilon \urcorner)$}
\RightLabel{$\to$-E}
\BinaryInfC{$\neg\mathsf{Tr}(\ulcorner \upsilon \urcorner)$}
\AxiomC{}
\RightLabel{2}
\UnaryInfC{σ}
\AxiomC{}
\RightLabel{1}
\UnaryInfC{$\neg\upsilon$}
\AxiomC{$\neg\upsilon \to \neg\sigma$}
\RightLabel{$\to$-E}
\BinaryInfC{$\neg\sigma$}
\BinaryInfC{$\bot$}
\RightLabel{CR, 1}
\UnaryInfC{υ}
\RightLabel{Tr-I}
\UnaryInfC{$\mathsf{Tr}(\ulcorner \upsilon \urcorner)$}
\RightLabel{$\neg$-E}
\BinaryInfC{$\bot$}
\RightLabel{$\neg$-I, 2}
\UnaryInfC{$\neg\sigma$}
\RightLabel{$\to$-I, 3}
\UnaryInfC{$\forall x[(\mathsf{C}(x) \to \neg\mathsf{Tr}(x)] \to \neg\sigma$}
\end{prooftree}
$$

We then show that (16) follows from (15) and $\forall x(\zeta)$, given Tr-E. For every $\varphi_i \in \{\varphi_1, \dots, \varphi_n\}$, call $\mathcal{D}_i$ the derivation of $\neg\mathsf{Tr}(\ulcorner \varphi_i \urcorner)$ from the open assumption $\neg\varphi_i$ via Tr-E:[8]

$$
\begin{prooftree}
\AxiomC{}
\RightLabel{i}
\UnaryInfC{$\neg\varphi_i$}
\AxiomC{}
\RightLabel{$(i \times k)$}
\UnaryInfC{$\mathsf{Tr}(\ulcorner \varphi_i \urcorner)$}
\RightLabel{Tr-E}
\UnaryInfC{φ_i}
\RightLabel{$\neg$-E}
\BinaryInfC{$\bot$}
\RightLabel{$\neg$-I, $(i \times k)$}
\UnaryInfC{$\neg\mathsf{Tr}(\ulcorner \varphi_i \urcorner)$}
\end{prooftree}
$$

We can now derive

$$
\begin{prooftree}
\AxiomC{$\forall x[\mathsf{C}(x) \to \neg\mathsf{Tr}(x)] \to \neg\sigma$}
\AxiomC{$\mathcal{D}_1$}
\noLine
\UnaryInfC{$\neg\mathsf{Tr}(\ulcorner \varphi_1 \urcorner)$}
\AxiomC{$\dots$}
\AxiomC{$\mathcal{D}_n$}
\noLine
\UnaryInfC{$\neg\mathsf{Tr}(\ulcorner \varphi_n \urcorner)$}
\AxiomC{$\forall x(\zeta)$}
\RightLabel{logic}
\QuaternaryInfC{$\forall x[\mathsf{C}(x) \to \neg\mathsf{Tr}(x)]$}
\RightLabel{$\to$-E}
\BinaryInfC{$\neg\sigma$}
\RightLabel{$\to$-I, n}
\UnaryInfC{$\neg\varphi_n \to \neg\sigma$}
\RightLabel{$\to$-I, $n-1$}
\UnaryInfC{$\neg\varphi_{n-1} \to (\neg\varphi_n \to \neg\sigma)$}
\noLine
\UnaryInfC{$\vdots$}
\RightLabel{$\to$-I, 1}
\UnaryInfC{$\neg\varphi_1 \to (\dots(\neg\varphi_{n-1} \to (\neg\varphi_n \to \neg\sigma))\dots)$}
\RightLabel{logic}
\UnaryInfC{$\neg\upsilon \to \neg\sigma$}
\end{prooftree}
$$

which completes our proof. The upper inference line labelled 'logic' is established exactly as above, while the lower one corresponds to multiple uses of the import-export and De Morgan laws.

It might be objected that our reconstruction of the interderivability between (13) and (14), plus auxiliary assumpions, is unacceptable because it uses Tr-E and Tr-I together with →-I and →-E, and these four rules, if taken unrestrictedly, yield triviality.[9] On the other hand, the objection continues, showing (13) and (14) to be interderivable 'via transparency' does not risk triviality: there are transparent and non-trivial theories of truth (such as Field's own theory) where (13) and (14) are interderivabile, but where our reconstruction is not available, since one of →-I and →-E is not unrestrictedly valid. Therefore, the objection concludes, the *Argument for Transparency* still stands: we need transparency to non-trivially model the equivalence between (13) and (14).

We find the objection mistaken, for at least two reasons. First, the fact that a natural piece of reasoning is classically inconsistent is not by itself an objection against the principles employed in the reasoning (*a fortiori* for theorists who are in general open to revise classical logic). Consider again Agreement and Disagreement. They, too, are pieces of reasoning which employ classical logic and naïve truth-theoretical principles *in order to motivate* Tr-I and Tr-E, and hence the need for a theory which validates them. Entirely analogously, our reconstruction of the interderivability between (13) and (14) can be taken to motivate Tr-I and Tr-E, or relevantly similar principles, and the need for a suitable theory that validates those principles. Second, as we will see, contextualism does provide a classical theory where versions of Tr-I and Tr-E are available, thus allowing us to recover our reconstruction of the equivalence between (13) and (14).

Summing up, the above arguments show that, *pace* Field, embedded cases of agreement and disagreement such as the above ones don't require truth to be transparent. *Pace* Picollo and Schindler, however, such cases still require that the truth predicate obey *both* Tr-E and Tr-I, or suitably related principles.

Picollo and Schindler on Tr-E

In a recent paper, Picollo and Schindler (2018) offer arguments that can be interpreted as showing that truth-elimination principles such as Tr-E suffice for the truth predicate to serve its expressive purposes. One of their arguments employs a result of Halbach (1999), to the effect that truth-theoretical generalizations (that is sentences of the form $\forall x(\varphi(x) \rightarrow \mathsf{Tr}(x))$) have the same truth-free consequences as the collection of their instances (that is all the instances of the schema $\varphi(\ulcorner \psi \urcorner) \rightarrow \psi$) in (suitably expressive) classical theories closed under Tr-I and Tr-E, where, crucially, the application of the latter rules is restricted to truth-free sentences. Picollo and Schindler observe that Halbach's result can also be proven if one only assumes closure under Tr-E alone (again, this principle needs to be restricted to truth-free sentences).[10] They subsequently turn to Field's argument concerning embedded agreement and disagreement, and argue that truth-introduction principles are not required. As they put it:

> there is an alternative and easy way to deal with the [embedded agreement and disagreement] case that only involves an elimination principle. For we can express (14) with a simple generalisation of the form $\forall x(\varphi(x) \rightarrow \mathsf{Tr}(x))$, instead of (13). Let $\varphi(x)$ be the predicate 'x is the unique sentence obtained by concatenating the conjunction of the $\mathsf{C}(x)$ with $\rightarrow \sigma$'. Then, we can choose $[\forall x(\varphi(x) \rightarrow \mathsf{Tr}(x))]$ to express (14): by [Picollo and Schindler's strengthening of Halbach's result], in any classical theory (that contains enough syntax theory to prove basic facts about concatenation) [closed under Tr-E restricted to sentences not containing the truth predicate] we can derive the

> latter from the former, relative to the assumption that what [is written in the report] is exactly $\varphi_1, \dots, \varphi_n$.
>
> (Picollo and Schindler, 2018, p. 913, numbering and notation have been adapted to ours)

To exemplify, Picollo and Schindler suggest to interpret

(11) If everything that the Conyers report says is true, then the 2004 election was stolen.

as follows. First, consider all the things that the Conyers report says, namely $\varphi_1, \dots, \varphi_n$. Then, consider the sentence $\varphi_1 \wedge \dots \wedge \varphi_n \rightarrow \sigma$, which says that if $\varphi_1 \wedge \dots \wedge \varphi_n$, then the election was stolen. Now consider the predicate 'results from concatenating $\varphi_1 \wedge \dots \wedge \varphi_n$ with $\rightarrow\sigma$'. Of course, there is exactly one sentence satisfying such a predicate, namely $\varphi_1 \wedge \dots \wedge \varphi_n \rightarrow \sigma$. Such a sentence can now be used to express (11) as a truth-theoretical generalisation, namely

(11^+) Every sentence that results from concatenating $\varphi_1 \wedge \dots \wedge \varphi_n$ with $\rightarrow\sigma$ is true.

Picollo and Schindler therefore suggest to express (11), a sentence of the form $\forall x(\varphi(x) \rightarrow \mathsf{Tr}(x)) \rightarrow \psi$, as a sentence of the form $\forall x(\chi(x) \rightarrow \mathsf{Tr}(x))$. Crucially, the interderivability of sentences such as (11) and (11^+) only requires $\mathsf{Tr\text{-}E}$.

Picollo and Schindler's strategy fails to convince, however, for at least two reasons. First, the applicability of their argument is undermined by the restriction of $\mathsf{Tr\text{-}E}$ to truth-free sentences. To see this, notice that while it is consistent for (sufficiently expressive) classical base theories to be closed under the *unrestricted* rule $\mathsf{Tr\text{-}E}$ (for examples, see Friedman and Sheard, 1987 and Halbach, 2011, Ch. 14), no classical base theory can be closed under the unrestricted $\mathsf{Tr\text{-}E}$ together with the assumption $\forall x(\mathsf{C}(x) \leftrightarrow x = \ulcorner\varphi_1\urcorner \vee \dots \vee x = \ulcorner\varphi_n\urcorner)$, for $\mathsf{C}(x)$ an arbitrary monadic formula of the language including the truth predicate. For one can then take the latter formula to be $\forall x(\mathsf{Tr}(x) \leftrightarrow x = \ulcorner\lambda\urcorner)$, for λ a Liar sentence, thus deriving $\mathsf{Tr}(\ulcorner\lambda\urcorner) \leftrightarrow \ulcorner\lambda\urcorner = \ulcorner\lambda\urcorner$ (by universal instantiation), and then $\mathsf{Tr}(\ulcorner\lambda\urcorner)$ (by *modus ponens* and the logic of identity), and finally $\neg\mathsf{Tr}(\ulcorner\lambda\urcorner)$ (by $\mathsf{Tr\text{-}E}$ and the definition of λ), thus proving a contradiction.[11]

Second, it is not clear what, if not a covert use of $\mathsf{Tr\text{-}I}$, can possibly justify the choice of $\varphi(x)$ in the formula $\forall x(\varphi(x) \rightarrow \mathsf{Tr}(x))$ in Picollo and Schindler's argument. They suggest to use $\forall x(\varphi(x) \rightarrow \mathsf{Tr}(x))$ in lieu of $\forall x(\mathsf{C}(x) \rightarrow \mathsf{Tr}(x)) \rightarrow\sigma$, where $\varphi(x)$ actually yields $\chi \rightarrow \sigma$ (relative to the assumption that what is written in the report is exactly $\varphi_1, \dots, \varphi_n$).

More precisely, we need some device for replacing a truth-theoretical generalisation in the antecedent of a conditional with its instances. But what justifies such a replacement, i.e. Picollo and Schindler's use of $\forall x(\varphi(x) \to \mathsf{Tr}(x))$? Picollo and Schindler do not address this question. Yet, derivations such as the derivation of (14) from (13) strongly suggest that the obvious answer is: a use of Tr-I! But such an answer is clearly precluded to Picollo and Schindler, since it would contradict the claim that Tr-E is sufficient for reasoning about embedded truth-ascriptions.[12] It would seem, then, that Picollo and Schindler can maintain that Tr-E suffices to account for cases of embedded agreement and disagreement only if Tr-I is already assumed to hold.[13]

In conclusion, Field's argument overshoots because transparency is not necessary for the expressive role of truth, even in cases of embedded truth-ascriptions, while Picollo and Schindler's argument undershoots because Tr-E alone is not sufficient. We would rather suggest that *in medio stat virtus*: namely, the combination of both Tr-E and Tr-I is exactly what is required by the expressive role of 'true'. To be sure, while this shows that the *Argument for Transparency* misses its target, our conclusion seemingly vindicates a version of the *Argument for Naïveté*—one according to which the expressive role of 'true' requires full Tr-E and Tr-I and, for this reason, is incompatible with classical theories of truth. As a result, it would seem, the classical theorist is still left with an uncomfortable choice between classical logic, on the one hand, and the expressive role of 'true', on the other. In the next section, however, we argue that this is a false dilemma. There exist *contextualist* versions of Tr-E and Tr-I that fully underwrite the expressive role of 'true' and that yet do not require one to adopt a non-classical theory of truth.

5.2 Contextualist Agreement and Disagreement

The problem, for most classical theories, is that they cannot validate both Tr-E and Tr-I. As the Knower Paradox shows (Kaplan and Montague, 1960; Myhill, 1960), they cannot simultaneously validate principles of truth introduction and truth elimination, even if one of the two is severely restricted (i.e. restricted to sentences that are proven from no assumptions). While non-hierarchical classical theories simply give up one between Tr-E and Tr-I, *hierarchical* theories typically still validate a certain version of Tr-I. In particular, *contextualist theories* feature versions of both Tr-E and Tr-I, where, crucially, the introduction rule is *context-shifting*.

Our plan is as follows. We first introduce orthodox contextualist approaches and point to some of their limitations (§5.2.1). We then sketch our own preferred contextualist approach (§5.2.2) and argue that it is expressively adequate, in the sense of validating both unembedded and embedded cases of agreement and disagreement (§5.2.3). We finally close by considering some objection and by offering some replies on the contextualist's behalf (§5.2.4).

5.2.1 An Overview of Orthodox Truth-Theoretic Contextualism

Contextualist proposals (first advanced by Charles Parsons, 1974 and extensively developed by Michael Glanzberg, 2001, 2004a, 2015) offer a more sophisticated, and nuanced, interpretation of truth-theoretical statements. Contextualists typically assume that *propositions* are the primary bearers of truth and falsity. Accordingly, a sentence φ is true in a context α if and only if it expresses a true proposition in α (Kaplan, 1989, p. 522). Following Glanzberg (2001, 2004a), we formalise the right-hand side of this biconditional as follows:

$$\exists_\alpha p(\mathsf{Exp}(\ulcorner\varphi\urcorner, p, \alpha) \land \mathbf{Tr}(p)),$$

where $\mathsf{Exp}(\ulcorner\varphi\urcorner, p, \alpha)$ reads 'φ expresses p in α', $\exists_\alpha p$ expresses existential quantification over a domain of propositions determined by the context α, and **Tr** expresses propositional truth. To further simplify notation, we write $\mathsf{Tr}_\alpha(\ulcorner\varphi\urcorner)$ for $\exists_\alpha p(\mathsf{Exp}(\ulcorner\varphi\urcorner, p, \alpha) \land \mathbf{Tr}(p))$ (operating under the assumption that, in each given context, every sentence expresses at most one proposition). Still following Glanzberg, we assume that any adequate (naïve) contextualist theory contains all the instances of the following schema:

$$\text{(CTS)}\ \forall_\alpha p[\mathsf{Exp}(\ulcorner\varphi\urcorner, p, \alpha) \to (\varphi \leftrightarrow \mathbf{Tr}(p))].$$

In contextualist approaches, a Liar sentence λ is interpreted as the claim that λ doesn't express a true proposition in a given context α. In these theories, λ doesn't express a true proposition in α, but expresses a true proposition in a new, richer context β. On the further assumption that λ, thus understood, involves propositional quantifiers whose domains vary with context, the Liar reasoning is blocked. Since the Liar reasoning is now interpreted as showing that λ is not true in α but true in β, it is no longer paradoxical and its conclusion no longer contradictory.

Here is, in more detail, the contextualist construal of the Liar reasoning (we informally follow Glanzberg, 2004a, pp. 33–34). Let γ be the initial context of reasoning and let λ be a sentence equivalent to 'λ doesn't express a true proposition in γ':

$$\lambda \leftrightarrow \neg\exists_\gamma p(\mathsf{Exp}(\ulcorner\lambda\urcorner, p, \gamma) \land \mathbf{Tr}(p))$$
$$\text{(the latter can also be written as } \neg\mathsf{Tr}_\gamma(\ulcorner\lambda\urcorner)).$$

One assumes that λ expresses a proposition in γ, and reasons (using CTS) that any such proposition p is true if and only if it is not true—a contradiction. Hence, one must negate and discharge the initial supposition that λ expresses a proposition, and conclude that

(17) λ does not express a proposition in γ.

One can then reason as follows:

(18) But, then, λ does not express a *true* proposition in γ [17, logic]
(19) Then, λ [Definition of λ]
(20) Thus, λ expresses a true proposition in γ [19, Tr-I]

Clearly, (18) and (20) contradict each other.

Contextualists maintain that the foregoing argument is not valid.[14] In particular, a *context shift* takes place between (18) and (20)—more specifically between (19) and (20) (Murzi and Rossi, 2018)—so that (20) should be interpreted as:

(20*) λ expresses a true proposition in γ' (for a context γ' different from γ).

Since (18) and (20*) are consistent, the Liar reasoning is blocked.[15]

Contextualist theories have several advantages. To name but a few: they retain classical logic; they model paradoxical reasonings as intuitively correct arguments, making them consistent by uncovering a context-shift in them; finally, they improve on some of our best theories of truth. For example, contextualist theories *à la* Glanzberg validate the truth-theoretical axioms of KF + Cons—that is, Feferman's classical axiomatization (Feferman, 1991) of Kripke's theory of truth (Kripke, 1975), plus the consistency axiom, ensuring that no sentence is both true and false.[16] However, while KF + Cons is incompatible with Tr-In, its contextualist version can be consistently closed under a context-shifting version of Tr-In, while retaining the virtues of the original theory.

On the other hand, on the orthodox contextualist view, quantification is *always* restricted to less than absolutely comprehensive domains. The reason, once again, is semantic paradox. In order to consistently interpret the Liar reasoning, λ is interpreted *twice*: first, in the initial context γ in which it does not express a true proposition, then in γ' where it does (see (18) and (20*) above). More specifically, the truth-introduction rule triggers a shift in the domain for the propositional quantifier—from a starting domain α to a domain β, where the propositions available for expressions in β strictly include those of α.[17] And since one can run paradoxical reasonings in any given context, and paradoxical reasonings extend—on the contextualist construal—any given domain of quantification, it seemingly follows that quantifiers can never range over absolutely everything. However, this feature of the approach clearly overshoots. Consider the following sentence:

(21) Everything is self-identical.

On an orthodox contextualist construal, even in sentences such as (21), the quantifiers range over less than absolutely everything. But this is unacceptable. Intuitively, the quantifier in (21) ranges over *absolutely everything that there is*, and a restricted reading of (21) would simply misconstrue it.[18]

In order to overcome this problem, we adopt the *bicontextualist* theory developed in Rossi (2021). The theory provides an absolutely general interpretation of unproblematic sentences such as (21), while preserving the context-shifting interpretation of paradoxical sentences and the positive aspects of orthodox contextualism *à la* Glanzberg more generally.[19]

5.2.2 Bicontextualism

The basic bicontextualist idea is that whether quantification can be absolutely general depends on the *sentence* in which the quantifiers appear: in some sentences (call them 'unparadoxical'), quantifiers can be interpreted as absolutely unrestricted; in others (call them 'paradoxical'), they cannot. Having distinguished between unparadoxical and paradoxical sentences, bicontextualist semantics provides an *absolutist semantics* for the former and a *relativist* semantics for the latter. In a slogan, bicontextualist semantics *recaptures absolute generality* whenever possible, just like standard non-classical approaches recapture classical logic whenever possible.[20]

In a bicontextualist semantics, we model context shift by moving from an interpretation for a given language to an interpretation for another, more expressive language. For this reason, we employ a proper class of first-order languages

$$\mathfrak{L} := \mathcal{L}_0, \mathcal{L}_1, \ldots, \mathcal{L}_\alpha, \ldots$$

where every $\mathcal{L}_\alpha \in \mathfrak{L}$ is expressive enough to admit the encoding of syntactic notions and $\cup_{\alpha<\beta}\mathcal{L}_\alpha \subseteq \mathcal{L}_\beta$, for $\alpha < \beta$. We write $\mathsf{Tr}_\alpha(\ulcorner\varphi\urcorner)$ for 'φ is true in α'. For simplicity, we take sentences as truth-bearers: as pointed out by Parsons (1974, p. 391 and following), this reformulation is harmless, and the talk about propositions can always be recovered by replacing '$\ulcorner\varphi\urcorner$ is true in α' with '$\ulcorner\varphi\urcorner$ expresses a true proposition in α'.

Absolutist Semantics

In order to construct second-order interpretations for unrestricted first-order languages, we follow Rayo and Uzquiano (1999) and Rayo and

Williamson (2003), adapting their approach to the case in which such languages contain a self-applicable truth predicate. We informally work in second-order Zermelo-Frankel set theory ($\mathsf{ZFC}2$). First, we define a predicate 'X is a model' in $\mathcal{L}_{\mathsf{ZFC}2}$, $\mathbb{M}(X)$. We skip the (lengthy) definition of $\mathbb{M}(X)$, but it can be paraphrased as follows: there is at least one x *s.t.* Xx holds (models are non-empty), and Xx holds whenever x specifies a domain, or the denotation of an individual constant, or the extension of a relation. The notions of 'domain', 'denotation', and 'extension' are given second-order paraphrases in the official definition of $\mathbb{M}(X)$.

We similarly define in $\mathcal{L}_{\mathsf{ZFC}2}$ a second-order predicate $\mathbb{A}(s, X)$, for 's is a variable assignment relative to X', and $\mathbb{V}(s, t, v_i, X)$ for 't is a v_i-variant of s relative to X'. The three second-order predicates $\mathbb{M}(X)$, $\mathbb{A}(s, X)$, and $\mathbb{V}(s, t, v_i, X)$ are now used to define in $\mathcal{L}_{\mathsf{ZFC}2}$ an absolutist notion of satisfaction the object-languages in $\mathfrak{L}$, patterned after the construction of the least fixed point of Kripke's theory of truth (1975).

Let x be a first-order variable, and X, Y, and Z be second-order variables. $\mathsf{KSat}_\alpha(x, X, Y, Z)$ holds just in case x is (the code of) a sentence that is satisfied in the (second-order) model X whose (second-order analogue of) domain is Y, relative to the accepted Z (note that, for simplicity reasons, the definition combines the second-order notions of model and variable assignment into one). The variable Z stands for the (codes of) sentences that are assumed to be true in the construction of the Kripkean model. We skip the (lenghty definition) of KSat for space reasons—the interested reader can consult Rossi (2021).[21]

We can now use KSat to define the set of *absolutely general truths* of each language $\mathcal{L}_\alpha$.

Definition 5.5.2. For every $\mathcal{L}_\alpha \in \mathfrak{L}$, the set of *absolutely general truths* of $\mathcal{L}_\alpha$ is defined as follows:

$$\mathsf{Abs}_\alpha := \{\varphi \in \mathcal{L}_\alpha \,|\, \mathsf{KSat}_\alpha(\varphi, X, Y, \varnothing)\}.$$

Abs_α contains the sentences that are in the least fixed point for the language $\mathcal{L}_\alpha$ whose quantifiers are interpreted as ranging over the domain encoded by X, possibly absolutely everything.

Relativist Semantics

In order to obtain a relativist semantics for paradoxical sentences, we adapt Glanzberg's original treatment to the present setting. In Glanzberg's approach, different (set-sized) 'closed-off' Kripkean fixed points are used to interpret paradoxical sentences in different contexts (the notion of 'closing-off' is sketched in what follows). However, in a bicontextualist semantics, the relativist interpretation has to apply *only* to the paradoxical sentences. Therefore, the paradoxical sentences of $\mathcal{L}_\alpha$ given by the closing off of a fixed point, minus that fixed point.

We now outline how to construct a succession of Kripkean closed-off fixed points, and then explicitly define the relatively general truths, for each $\mathcal{L}_\alpha$. Let $\varnothing$ be a predicate with an empty extension, $\mathcal{M}_\alpha$ be a model of the non-semantic fragment of $\mathcal{L}_\alpha$, and M_α be its support. The relativistic closing off for $\mathcal{L}_0$ is:

$$\mathsf{C\text{-}Off}_0 := \{\varphi \in \mathsf{Sent}_{\mathcal{L}_0} \mid \langle \mathcal{M}_0, \{\psi \in \mathcal{L}_0 \mid \mathsf{KSat}_0(\ulcorner \varphi \urcorner, \mathcal{M}_0, \mathsf{M}_0, \varnothing)\}\rangle \models \varphi\}.$$

For $\alpha > 0$, the closing-off of $\mathcal{L}_\alpha$ is:

$$\begin{aligned}&\mathsf{C\text{-}Off}_\alpha := \{\varphi \in \mathsf{Sent}_{\mathcal{L}_\alpha} \mid \\ &\quad \langle \mathcal{M}_\alpha, \{\psi \in \mathcal{L}_\alpha \mid \mathsf{KSat}_\alpha(\ulcorner \psi \urcorner, \mathcal{M}_\alpha, \mathsf{M}_\alpha \cup \bigcup_{\beta<\alpha} \mathsf{C\text{-}Off}_\beta), \bigcup_{\beta<\alpha} \mathsf{C\text{-}Off}_\beta)\rangle \models \varphi\}.\end{aligned}$$

This succession of closed-off fixed points treats paradoxical sentences exactly as in Glanzberg's original approach. Consider a Liar in $\mathcal{L}_0$, call it λ_0. λ_0 is not in the extension of the fixed point defined by $\mathsf{KSat}_0(\ulcorner \varphi \urcorner, \mathcal{M}_0, \mathsf{M}_0, \varnothing)$. Since $\mathsf{KSat}_0(\ulcorner \varphi \urcorner, \mathcal{M}_0, \mathsf{M}_0, \varnothing)$ yields the extension of Tr_0, λ_0 is not in it. Therefore:

$$\begin{aligned}\langle \mathcal{M}_0, \{\psi \in \mathcal{L}_0 \mid \mathsf{KSat}_0(\ulcorner \varphi \urcorner, \mathcal{M}_0, \mathsf{M}_0, \varnothing)\}\rangle &\not\models \mathsf{Tr}_0(\ulcorner \lambda_0 \urcorner)\\ &\models \neg \mathsf{Tr}_0(\ulcorner \lambda_0 \urcorner)\\ &\models \lambda_0\end{aligned}$$

However, since $\lambda_0 \in \mathsf{C\text{-}Off}_0$, when one builds a fixed point for $\mathcal{L}_1$ over $\mathsf{C\text{-}Off}_0$, λ_0 goes into the extension of Tr_1. Therefore:

$$\langle \mathcal{M}_1, \mathsf{KSat}_1(\ulcorner \varphi \urcorner, \mathcal{M}_{1,}\, \mathsf{M}_1 \cup \mathsf{C\text{-}Off}_0, \mathsf{C\text{-}Off}_0)\rangle \models \mathsf{Tr}_1(\ulcorner \lambda_0 \urcorner).$$

That is, λ_0 does not express a true proposition of $\mathcal{L}_0$, but it expresses a true proposition of $\mathcal{L}_1$, as sketched in the reasoning (18)–(20*). Other paradoxical sentences are treated in the very same way.

Now that we have defined the closed-off fixed points, we can easily define the paradoxical sentences.

DEFINITION 5.2.3. For every $\mathcal{L}_\alpha \in \mathfrak{L}$, the set of *relatively general truths* of $\mathcal{L}_\alpha$, in symbols Rel_α, is defined as the set of sentences in $\mathsf{C\text{-}Off}_\alpha$, minus the sentences in Abs_α and the sentences φ *s.t.* $\varphi \in \mathsf{C\text{-}Off}_\alpha$ but $\neg\ \varphi \in \mathsf{Abs}_\alpha$ or vice versa.

The relatively general truths of $\mathcal{L}_\alpha$ are exactly the truths of $\mathcal{L}_\alpha$'s closed-off fixed point minus $\mathcal{L}_\alpha$'s absolutely general truths.

Bicontextualism in Full

We can now finally define a proper bicontextualist theory of truth.

DEFINITION 5.2.4. For every $\mathcal{L}_\alpha \in \mathfrak{L}$, and every $\{\Gamma, \varphi\} \subseteq \mathcal{L}_\alpha$, the argument from Γ to φ is *bicontextually valid*, in symbols $\Gamma \models^{\mathsf{bc}}_{\alpha} \varphi$, if and only if:

> all the sentences in Γ are in $\mathsf{Abs}_\alpha \cup \mathsf{Rel}_\alpha$, so is φ.

For every $\mathcal{L}_\alpha \in \mathfrak{L}$, the bicontextualist semantics for $\mathcal{L}_\alpha$ interprets all the unparadoxical sentences over a possibly absolutely unrestricted domain, and the paradoxical ones over a restricted, set-sized domain.

Bicontextualism delivers a strong theory of truth. Among other things, it validates all the axioms of $\mathsf{KF} + \mathsf{Cons}$ for every language $\mathcal{L}_\alpha$. More importantly for our present purposes, however, it validates forms of truth-introduction and elimination which can be used to recover talk about AGREEMENT and DISAGREEMENT, to which we now turn.

5.2.3 Bicontextualism Agreement and Disagreement

Both orthodox contextualist semantics and bicontextualism validate the following introduction rule for the the truth predicate (we formulate the rules as inferences from sequents to sequents to make fully clear the context in which each sentence occurs):

$$(\mathsf{C\text{-}Tr\text{-}I})\ \text{If}\ \Gamma \models^{\mathsf{bc}}_{\alpha} \varphi,\ \text{then}\ \Gamma \models^{\mathsf{bc}}_{\beta} \mathsf{Tr}_\beta(\ulcorner \varphi \urcorner).$$

Its contrapositive governs the elimination of the negated truth predicate:

$$(\mathsf{C\text{-}\neg Tr\text{-}E})\ \text{If}\ \Gamma \models^{\mathsf{bc}}_{\beta} \neg\mathsf{Tr}_\beta(\ulcorner \varphi \urcorner),\ \text{then}\ \Gamma \models^{\mathsf{bc}}_{\alpha} \neg\varphi,$$

where β is strictly greater than α. While the latter two rules may require a shift of context, the corresponding converse rules are not context-shifting:

$$(\mathsf{C\text{-}Tr\text{-}E})\ \text{If}\ \Gamma \models^{\mathsf{bc}}_{\alpha} \mathsf{Tr}_\alpha(\ulcorner \varphi \urcorner),\ \text{then}\ \Gamma \models^{\mathsf{bc}}_{\alpha} \varphi;$$
$$(\mathsf{C\text{-}\neg Tr\text{-}I})\ \text{If}\ \Gamma \models^{\mathsf{bc}}_{\alpha} \neg\varphi,\ \text{then}\ \Gamma \models^{\mathsf{bc}}_{\alpha} \neg\mathsf{Tr}_\alpha(\ulcorner \varphi \urcorner).$$

As we have seen in §5.1, naïve principles such as $\mathsf{Tr\text{-}E}$ and $\mathsf{Tr\text{-}I}$ seem to be required in order to model speeches such as AGREEMENT and DISAGREEMENT. However, AGREEMENT and DISAGREEMENT can be equally modelled by means of a contextualist truth predicate. In particular, speeches requiring a truth introduction principle can be modelled using the contextualist rules $\mathsf{C\text{-}Tr\text{-}I}$ and $\mathsf{C\text{-}\neg Tr\text{-}E}$. For instance, contextualists may interpret Disagreement as follows (letting ‘$\varsigma(x,y)$’ formalise the claim ‘x said y’):

$$\dfrac{\dfrac{\dfrac{\forall x[\varsigma(l,x)\rightarrow\neg\mathsf{Tr}_{\beta}(x)]}{\varsigma(l,\ulcorner\varphi\urcorner)\rightarrow\neg\mathsf{Tr}_{\beta}(\ulcorner\varphi\urcorner)}\,\forall\text{-E}\qquad\varsigma(l,\ulcorner\varphi\urcorner)}{\neg\mathsf{Tr}_{\beta}(\ulcorner\varphi\urcorner)}\,\rightarrow\text{-E}}{\neg\varphi}\,\text{C-}\neg\text{-Tr-E}$$

where the first three lines occur in β, the fourth line occurs in α, and β is strictly greater than α.

Contextualist approaches, then, can adequately model key uses of the truth predicate such as Agreement and Disagreement, *modulo* the possible occurrence of context shifts. In addition—as shown in §5.1.2—cases of embedded agreement and disagreement can also be modelled by means of contextualist principles of truth introduction and elimination. However, orthodox contextualist theories and the bicontextualist theory we favour differ in one important respect. While the former theories always interpret C-¬Tr-I and C-¬Tr-E as context-shifting, in a bicontextualist framework these rules are *context-shifting only when applied to paradoxical sentences*. By contrast, non-paradoxical sentences do not trigger context-shifts and are interpreted by the absolutist fragment of the semantics. Therefore, cases of (embedded or non-embedded) agreement and disagreement involving non-paradoxical sentences are modelled as if the truth predicate were naïve, i.e. by means of perfectly symmetrical introduction and elimination rules.

Let us now consider again the classical rejoinder given in §5.1.1. There, we argued that the classical truth-theorist might remain unconvinced by the intuitive validity of schematic arguments such as Agreement and Disagreement because they seem to have instances—e.g. those involving paradoxical sentences—that are not clearly compelling. Bicontextualist semantics seems to do justice to this intuition: instances of Agreement and Disagreement that do not involve paradoxical sentences can be modelled naïvely (as in most classical approaches). In addition, however, instances involving paradoxical sentences can also be modelled. However, in this case the introduction of the truth predicate and the elimination of the negated truth predicate semantically correspond to the articulation of a truth-theory for the sentences of a given language, and therefore require a stronger context in which they can be given.

We now turn to some potential objections—in particular, objections concerning whether it is always possible to model agreement and disagreement in a bipartite and hierarchical manner.[22]

5.2.4 Objections and Replies

We consider three different kinds of objections: objections from semantic blindness, Dean-Nixon cases, and objections from ineffability, to the effect that systematic domain restrictions cannot be coherently stated.

Semantic Blindness

Consider a blind ascription, such as

(6) Everything Lois said yesterday is not true,

and suppose we don't know what Lois said yesterday. Kripke (1975, p. 695 and ff.) famously argued, against Tarskian hierarchical approaches, that one may not be in a position to assign a 'level' to the truth predicate occurring in (6).[23] In a contextualist framework, though, Kripke's point is not well taken. The objection assumes that, in order to successfully use sentences such as (6), a speaker must know the interpretation of 'true' in a given context. In our framework, (6) is interpreted as

(6*) Everything Lois said yesterday expresses a true proposition in some context α.

More formally:

(6**) $\forall x(\varsigma(Lois, x) \rightarrow \mathsf{Tr}_{\alpha}(\ulcorner \varphi \urcorner))$.

Thus, in our framework, the objection postulates that a speaker must know in which context the truth ascription in (6**) must be interpreted. However, in general speakers need not know the features of context that are relevant for the interpretation of context-sensitive expressions. Consider the sentence

(22) It's cloudy now.

Speakers utter sentences like these even if they don't exactly know what time it is. Granted, in order to *know what* (22) *says*, speakers must first find out what time it is. But a speaker doesn't need to know the time in order to successfully *use* sentences like (16). Similarly, even if we don't know what Lois said yesterday, we are still able to use (6) to successfully attribute truth or untruth to what she says, even though we are not always in a position to know the content of what we have just agreed or disagreed on. All is required from a contextualist semantics for 'true' is that *there be* a context that provides a suitable interpretation of the context-sensitive expressions in (6). And contextualist approaches standardly satisfy this requirement.

Nixon-Dean Cases

In less liberal hierarchical approaches, such as Tarski's, certain everyday speeches cannot be coherently interpreted. Suppose Trump and Cohen only utter, respectively, the following sentences (cf. Kripke, 1975, pp. 695–696):

(23) Everything Cohen says about the hush payment is true;
(24) Everything Trump says about the hush payment is not true.

In a standard Tarskian framework, (23) and (24) cannot be interpreted, since each sentence would need to involve a truth predicate of higher level than that of the truth predicate occurring in the other sentence—which is of course impossible. However, as Glanzberg (2015, p. 233) points out, the objection doesn't apply to more liberal hierarchical theories. In particular, contextualist theories *à la* Glanzberg use iterations of Kripkean fixed points to interpret paradoxical sentences (and only those). Therefore, (23) and (24) can be interpreted exactly as in Kripke's theory (and closed-off versions of the theory) in every such fixed point.[24]

It might be insisted that if the speakers explicitly talk about the context they're in, a version of Kripke's objection can still be made to work (Field, 2008, pp. 217–218). For suppose Trump and Cohen only utter, respectively, in contexts δ and γ:

(23*) Everything Cohen says in context γ about the hush payment is true;
(24*) Everything Trump says in context δ about the hush payment is not true.

Then, as Field (2008) put it,

> for one guy to succeed in saying what he intended he must pick a strictly higher subscript than the other. Because of this, there is pressure for [Trump] and [Cohen] to get into a subscript contest.
> (p. 217)

However, our treatment of (23*) and (24*) generalises to (23) and (24). The latter two utterances can be given a standard contextualist treatment, interpreting them as a simple variant of the Liar reasoning. More precisely, since both sentences are paradoxical, they are interpreted by some Kripke fixed point in which neither of them expresses a true proposition. Such Kripke fixed points corresponds to interpreting (23*) and (24*) in some suitable context α that extends the contexts over which both (23*) and (24*) quantify. As with the original Liar sentence λ, both (23*) and (24*) can then be shown to express propositions in a context β that extends α.

Stating the View

Field (2008, pp. 220–221) considers a further objection, to the effect that contextualist theories cannot coherently interpret a theorist's disagreement with the *contextualist theory as a whole*. For suppose one wishes to assert that the contextualist theory is *not true*. Then, it might seem that one needs to quantify over all the propositions that are expressed in all the contexts, which is something that contextualist theories cannot do (with the intended absolutely unrestricted interpretation). This is an instance of a more general objection, to the effect that hierarchical approaches to the paradoxes lack the expressive resources to talk about the whole hierarchy (in our case: the whole hierarchy of propositions in context).[25] We now provide a quick sketch of how this objection can be addressed in a contextualist framework.

Bicontextualists place some limits on absolutely unrestricted quantification: the quantifiers of paradoxical sentences are necessarily restricted to less than absolutely general domains. Yet, in many contexts, they are able to quantify over absolutely everything. How to characterise, more precisely, such a view? This is an instance of a more general problem afflicting the two main views about generality: *absolutism*, according to which it is possible to quantify over absolutely everything, and *relativism*, according to which quantification is never absolutely general. On one hand, relativists face the objection that they cannot express their own view, if it is expressed as the thought that we cannot quantify over *absolutely everything*. After all, the objection goes, this very thought presupposes absolute generality.[26] On the other, absolutists are accused of only being able to offer essentially meta-theoretic and hence, one might argue, unintelligible statements of their view. What is the status of bicontextualism in this landscape? The answer, in a nutshell, is that bicontextualism is essentially an absolutist view, with some restrictions. But it is neither incoherent nor unintelligible.

The meta-theory of bicontextualism is absolutist, i.e. it has the resources to express absolutely general quantification. More specifically, some open formulae are satisfied by *everything*, i.e. by a collection of things whose size exceeds that of any set, however large. Accordingly, the bicontextualist can express her view as a conjunction of two claims:

(25) There are open formulae that are satisfied by absolutely everything;

(26) There are open formulae that are only satisfied by set-many things.

Here 'absolutely everything' is shorthand for the satisfaction clause for universally quantified sentences, formulated in a higher-order language. Hence, both (25) and (26) are essentially meta-linguistic claims.

It might be objected that this is deeply unsatisfactory, on the grounds that (25) and (26) look like perfectly ordinary object-language sentences that ought to be expressed in the object-language. Indeed, it might be insisted that English is not divided into English and meta-English and that, in principle, everything that is expressible in the meta-language ought to also be expressible in the object-language—including (25) and (26). After all, what is the point of having a type-free truth predicate, if not to formulate the interpretation (i.e. a traditionally meta-theoretical notion) for a theory *S in S*?

One first point to notice here is that (25) and (26)'s meta-linguistic character isn't specific to bicontextualism. For instance, (25) could be just as well taken to express the claim that it is possible to quantify over absolutely everything—a standard statement of the absolutist view (Williamson, 2003). If stating one's view about the interpretation of quantifiers in the meta-language is problematic, then it is problematic for the absolutist and the relativist just as well.

More importantly, the objection overshoots: it is false in general that everything that is the meta-language ought to be expressible in the object-language; only *what is already expressible in English* should. And, in this respect, there is an asymmetry between truth and higher-order quantification. The reason why a type-free truth predicate ought to be part of the object-language is the simple existence of English speeches that make an essential use of it—speeches such as AGREEMENT, DISAGREEMENT, and the like. However, no parallel motivation for expressing higher-order quantification in the object-language is available. To see this, notice that a genuine higher-order quantification amounts to *quantifying into predicate position*, i.e. quantifying over something that also figures as a predicate in the expression that follows the quantifier. For instance, a faithful reading of an expression of the form $\forall X \forall x(Xx)$ would amount to something like 'Every X Xs every x'. That is, X would need to simultaneously be the syntactic object to which the quantifier applies *and* the predicate in the expression following the quantifier. Yet this doesn't seem to be possible in languages such as English. To be sure, plenty of English paraphrases of $\forall X \forall x(Xx)$ are available. For instance, 'For every predicate, everything satisfies it'. These translations are systematically misleading, though: they effectively treat X as a *first-order* variable, because, in the paraphrases, X does not serve as a predicate in the expression following the quantifier.

5.3 Concluding Remarks

It is usually thought that Tarski's Theorem forces on us an uncomfortable dilemma: either adopt a non-classical theory of truth, or restrict some of the naïve truth principles and 'seriously cripple our ability to make generalizations' (Field, 2008, p. 349). In turn, this has led many to adopt a non-classical theory of truth. However, there are good reasons for thinking that non-classical approaches are fundamentally misguided (Murzi and Rossi, 2020): they inevitably give rise to revenge paradoxes they are either unable to block, or they can only block at unacceptable costs. We hope to have shown that this is a false dilemma. The expressive role of truth doesn't require naïveté: contextualist principles of truth (and untruth) introduction and elimination suffice. There are several possible reasons for adopting a non-classical theory of truth. But the legitimate desire to preserve, on the face of paradox, our ability to make generalisations is not one of them.

Acknowledgements

We are grateful to Carlo Nicolai and Johannes Stern for very helpful comments on a previous draft of this paper, as well as to the FWF (Stand-Alone Grant P29716-G24), for generous financial support during the time this article was written.

Notes

1. Even if the truth-predicate is naïve, certain instances of the naïve principles for truth can plausibly fail in opaque contexts, quite independently of the paradoxes. For instance, 'Anne believes that every even number greater than two is the sum of two primes' may not imply 'Anne believes that "every even number greater than two is the sum of two primes" is true', for the simple reason that Anne may lack the concept of truth, or that she may systematically refuse to apply it to arithmetical sentences.
2. Typically, we leave the formal system unspecified and assume that the arguments we model are formalised in a theory featuring logical rules, a *modicum* of syntax theory (in order to have a well-behaved name-forming device $\ulcorner \cdot \urcorner$), and semantic rules (governing the truth predicate).
3. Even if they don't contain all instances of $\mathsf{Tr\text{-}I}$ and $\mathsf{Tr\text{-}E}$, classical theories allow applications of $\mathsf{Tr\text{-}I}$ and $\mathsf{Tr\text{-}E}$ to (more or less large) classes of sentences. These typically include non-semantic sentences, or sentences involving iterated applications of the truth predicate to non-semantic sentences. For instance, even classical typed theories of truth allow instances of $\mathsf{Tr\text{-}I}$ and $\mathsf{Tr\text{-}E}$ for sentences not containing the truth predicate (see for example the theories TB and

UTB Halbach, 2011, Ch. 7), while strong classical theories allow instances of the naïve principles for possibly very complex iterated applications of the truth predicate to non-semantic sentences (see for example Feferman, 1991, 2008).

4. Note that the foregoing considerations also apply to instances of Agreement and Disagreement involving obviously false sentences, such as 'Donald Trump is a theorem of Peano Arithmetic'.
5. Classical theorists typically restrict naïve principles to non-paradoxical sentences, for some suitable understanding of the notion of paradoxicality. Similarly, non-classical theorists restrict *classical principles* to non-paradoxical sentences. For a discussion of whether this in turn leads to certain revenge problems, see Bacon (2015), for the classical case, and Murzi and Rossi (2020), for the non-classical one.
6. We consider a scenario in which one wishes to disagree with everything that was written in the report. The case of an embedded disagreement involving an antecedent of the form 'something in the Conyers report is untrue' is dealt with similarly.
7. The line labelled 'CR' stands for the rule of classical reductio. Also, we use the assumption $\mathsf{C}(\ulcorner\upsilon\urcorner)$, rather than $\mathsf{C}(\ulcorner\chi\urcorner)$, for simplicity; however, if υ is the disjunction of the claims in the Conyers report and χ is their conjunction, then the former clearly follows from the latter, so we could have equally assumed $\mathsf{C}(\ulcorner\chi\urcorner)$.
8. We use the index $(i \times k)$ to make sure that it is sufficiently high not to clash with the indices of the open assumptions φ_1, which are to be closed and discarded later.
9. There might be exceptions in some substructural settings.
10. It should be noted that Picollo and Schindler do not endorse the claim that results such as the one above show that truth-theoretical generalisations 'express' all their instances.
11. It should be noted that a (sufficiently expressive) classical base theory S that features $\forall x(\mathsf{C}(x) \leftrightarrow x = \ulcorner\varphi_1\urcorner \vee \ldots \vee x = \ulcorner\varphi_n\urcorner)$ amongst its axioms can be consistently closed under all the instances of the inference from $\forall x(\mathsf{C}(x) \rightarrow \mathsf{Tr}(x)) \rightarrow \psi$ to $\varphi_1 \wedge \ldots \wedge \varphi_n \rightarrow \psi$. To see that the first inference goes through it is sufficient to consider any model of the form $\mathcal{A} := \langle \mathcal{M}, A \rangle$, where $\mathcal{M}$ is a model of the truth-free part of the language of S and A is the set $\{x \in M \mid x = \ulcorner\varphi_1\urcorner \text{ or } \ldots \text{ or } x = \ulcorner\varphi_n\urcorner\}$ which serves as the extension of Tr in M (the support of $\mathcal{M}$). The converse inference, i.e. from $\varphi_1 \wedge \ldots \wedge \varphi_n \rightarrow \psi$ to $\forall x(\mathsf{C}(x) \rightarrow \mathsf{Tr}(x)) \rightarrow \psi$, impossibly requires closure under unrestricted Tr-E.
12. We are indebted to Carlo Nicolai for bringing this point to our attention.
13. Picollo and Schindler (2019) further discuss the expressive role of 'true'. In this more recent paper, they too argue that both Tr-I and Tr-E are essential for the expressive role of 'true'.
14. See e.g. Parsons (1974) and Glanzberg (2001, 2004a).
15. On *why* exactly context shifts in the course of the Liar reasoning, see Glanzberg (2004b, 2015), Gauker (2006), and Murzi and Rossi (2018).
16. For more details on KF and its extensions, see Field (2008, Chapters 7 and 13) and Halbach (2011, Chapter 15).
17. To be sure, the question arises why truth-introduction principles induce a context shift. For instance, Glanzberg (2004b, 2015) argues that when we explicitly articulate our acceptance of a given theory S, we are committed to principles that determine the truth-conditions of S's sentences, that is, we are committed to a suitable theory of truth S' for S. Since S''s truth predicate cannot be defined in the theory S itself, on pain

of triviality, a more expressive theory is required, whence the context-shifting properties of truth-theoretical reflection (see also Murzi and Rossi, 2018).

18. Williamson (2003) further argues that absolutely general quantification is crucial for philosophical and scientific theorizing and that, for this reason, any view on which quantification is necessarily restricted is at odds with the level of generality required by scientific reasoning.
19. For reasons of space, we only offer a brief, informal outline of the theory. The following subsection draws extensively from the more detailed presentation in Rossi (2021). We also assume some familiarity with the models constructed in Kripke (1975). The reader uninterested in the outline of the theory can easily skip the next subsection.
20. For simplicity, we present a theory in which the paradoxical sentences are exactly the sentences that are in the gap of a suitable minimal Kripkean fixed point (see Kripke, 1975). See Rossi (2019, 2021) for details on how to extend the theory to languages with richer vocabularies, in which further semantic notions, including revenge-breeding notions, can be formulated.
21. The extension of Z is empty in the case of the least fixed point, and non-empty for a non-minimal fixed point.
22. In addressing the main objections faced by a bicontextualist treatment of agreement and disagreement, we also touch on more general issues regarding the theory of quantification adopted by bicontextualist semantics. For reasons of space, however, our discussion will be necessarily brief.
23. Here we are not yet concerned with the possibility that (6) may be paradoxical. We consider paradoxical sentences when discussing Nixon-Dean cases below.
24. We adopt in essence Glanzberg's reply to the objection from Nixon-Dean cases. The crucial difference between Glanzberg's theory and ours is that we limit the use of iterations of Kripke fixed-point models to paradoxical sentences.
25. The general objection is familiar and has been pressed by a number of authors, including e.g. Priest (2006, §§1.6–7) and Linnebo (2006, §4).
26. See e.g. Lewis (1991, pp. 61–68) and Williamson (2003, pp. 427–428). Relativists typically reject this objection, because it presupposes an absolutist interpretation of 'everything'. They insist that, once 'everything' is interpreted relativistically, the sentence is simply false, and hence inadequate to express their view.

References

Bacon, A. (2015). Can the classical logician avoid the revenge paradoxes? *Philosophical Review*, 124(3): 299–352.

Beall, J. (2009). *Spandrels of Truth*. Oxford University Press.

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

Feferman, S. (2008). Axioms for determinateness and truth. *Review of Symbolic Logic*, 1(2): 204–217.

Field, H. (2008). *Saving Truth from Paradox*. Oxford University Press.

Friedman, H. and Sheard, M. (1987). An axiomatic approach to self-referential truth. *Annals of Pure and Applied Logic*, 33: 1–21.

Gauker, C. (2006). Against stepping back: A critique of contextualist approaches to the semantic paradoxes. *Journal of Philosophical Logic*, 35 (4): 393–422.

Glanzberg, M. (2001). The liar in context. *Philosophical Studies*, 103(3): 217–251.

Glanzberg, M. (2004a). A contextual-hierarchical approach to truth and the liar paradox. *Journal of Philosophical Logic*, 33: 27–88.

Glanzberg, M. (2004b). Truth, reflection, and hierarchies. *Synthese*, 142(3): 289–315.

Glanzberg, M. (2015). Complexity and Hierarchy in Truth Predicates. In Achourioti, T., Galinon, H., Martinez Fernández, J., and Fujimoto, K., editors, *Unifying the Philosophy of Truth*, volume 36 of Logic, Epistemology, and the Unity of Science. Springer.

Halbach, V. (1999). Disquotationalism and infinite conjunctions. *Mind*, 108 (429): 1–22.

Halbach, V. (2011). *Axiomatic Theories of Truth*. Cambridge University Press.

Horsten, L. (2012). *The Tarskian Turn. Deflationism and Axiomatic Truth*. MIT Press.

Kaplan, D. (1989). Demonstratives. In Almog, J., Perry, J., and Wettstein, H., editors, *Themes from Kaplan*. Oxford University Press.

Kaplan, D. and Montague, R. (1960). A paradox regained. *Notre Dame Journal of Formal Logic*, 1: 79–90.

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

Lewis, D. (1991). *Parts of Classes*. Basil Blackwell.

Linnebo, O. (2006). Sets, properties, and unrestricted quantification. In Rayo, A. and Uzquiano, G., editors, *Absolute Generality*, pages 149–178. Oxford University Press.

McGee, V. (1985). A counterexample to modus ponens. *The Journal of Philosophy*, 82: 462–471.

Murzi, J. and Rossi, L. (2018). Reflection principles and the Liar in context. *Philosophers' Imprint*, 18(15): 1–18.

Murzi, J. and Rossi, L. (2020). Generalised revenge. *Australasian Journal of Philosophy*, 98(1): 153–177.

Myhill, J. (1960). Some remarks on the notion of proof. *Journal of Philosophy*, 57(14): 461–471.

Parsons, C. (1974). The Liar Paradox. *Journal of Philosophical Logic*, 3(4): 381–412.

Picollo, L. and Schindler, T. (2018). Disquotation and infinite conjunctions. *Erkenntnis*, 83(5): 899–928.

Picollo, L. and Schindler, T. (2019). Deflationism and the function of truth. *Philosophical Perspectives*, 32: 326–351.

Priest, G. (2006). *In Contradiction*. Oxford University Press, Expanded edition (first published 1987 Kluwer-Dordrecht).

Rayo, A. and Uzquiano, G. (1999). Toward a theory of second-order consequence. *Notre Dame Journal of Formal Logic*, 40(3): 315–325.

Rayo, A. and Williamson, T. (2003). A completeness theorem for unrestricted first-order languages. In Beall, J., editor, *Liars and Heaps*, pages 331–356. Oxford University Press.

Rossi, L. (2019). A unified theory of truth and paradox. *The Review of Symbolic Logic*, 12(2): 209–254.

Rossi, L. (2021). *Bicontextualism*. Unpublished manuscript.

Williamson, T. (2003). Everything. *Philosophical Perspectives*, 17(1): 415–465.

6 Disquotationalism and the Compositional Principles

Richard Kimberly Heck

> [S]emantics ... is a sober and modest discipline which has no pretensions of being a universal patent-medicine for all the diseases of mankind, whether imaginary or real. You will not find in semantics any remedy for decayed teeth or illusions of grandeur or class conflicts. Nor is semantics a device for establishing that everyone except the speaker and his friends is speaking nonsense.
>
> (Tarski, 1944, p. 345)

In their paper "The Use of Force Against Deflationism", Dorit Bar-On and Keith Simmons (2007, p. 61) helpfully distinguish three sorts of deflationary theses about truth. *Metaphysical* deflationism is a thesis about the property of truth, namely, that it is insubstantial, or that it has no essential nature, so that a theory of truth—a correspondence or coherence theory of truth, say—is both unnecessary and impossible. *Linguistic* deflationism is a thesis about the word "true", namely, that its meaning is adequately explained by Alfred Tarski's convention (T) or something along the same lines. *Conceptual* deflationism is a thesis about the role that the notion of truth may legitimately play in our theorizing, namely, that there are no interesting connections between truth and other concepts, such as meaning or belief. Rather, the notion of truth serves only an 'expressive' function, allowing us to formulate certain claims that we could not state without it, but playing no essential explanatory role (see e.g. Field, 1994, §5; Williams, 1999, p. 547).

Most deflationists seem to regard the linguistic thesis as fundamental. Exactly how the metaphysical thesis is supposed to follow from it has never been clear to me, probably because I do not understand what it is supposed to mean that truth is not a 'substantial property' (cf. Field, 1994, p. 265, n. 19). But the real issue, in any event, concerns the conceptual thesis. Deflationists have generally regarded it as following from the linguistic thesis: Surely an expression that is explained in terms that make it all but redundant cannot play any essential explanatory role. In practice, the argument for this claim takes a form made familiar by Paul Horwich (1990). The work the notion of truth appears to do in various settings, it is claimed, can in fact be done entirely by a notion of truth that is stipulatively introduced in accordance with the tenets of linguistic deflationism. So the dialectic consists of the anti-deflationist's identifying some theoretical setting in which the notion of truth seems to be doing important explanatory work and the deflationist's attempting to show that the role the truth-predicate is playing in that context is, in fact, purely expressive. What it is for a use of the truth-predicate to be 'purely expressive' is a question to which we'll return.

Bar-On and Simmons's main purpose in their paper is to show that conceptual deflationism does not follow from linguistic and metaphysical deflationism. In particular, they argue that Gottlob Frege, though he was a linguistic deflationist, is not a conceptual deflationist (Bar-On and Simmons, 2007, §II) and that Robert Brandom, though he is both a linguistic and a metaphysical deflationist, is not a conceptual deflationist, either (Bar-On and Simmons, 2007, §III). More precisely, Bar-On and Simmons argue that the notion of truth plays an essential role both in Frege's account of assertion and in Brandom's account of 'commitment', and that in neither case is truth's role merely expressive. They do not, however, actually defend Frege's claim that assertion is the presentation of a thought as true,[1] nor Brandom's account of commitment in terms of taking to be true, so it is open to a deflationist simply to reject those accounts. Deflationists hold that truth's only *legitimate* role is expressive. They need not deny that less enlightened philosophers have tried to make other uses of it.

Frege's own attitude toward truth is nonetheless instructive. On the one hand, Frege famously insists that "the sentence 'The thought that 5 is a prime number is true' contains ... the same thought as the simple '5 is a prime number'" (Frege, 1984c, op. 34). On the other hand, truth plays an absolutely central role in Frege's thought about language and, in particular, in the semantics that he develops for his formal language in Part I of *Grundgesetze der Arithmetik* (Frege, 2013). For Frege, the most fundamental linguistic unit, from a logical point of view, is the sentence; the most basic semantic fact about a sentence is its being true or false; and the sense of a sentence—the thought it expresses—is its truth-condition. How can Frege hold all these views? What explains the apparent tension is the fact that Frege's deflationary remarks always concern ascriptions of truth to what he called 'thoughts' (that is, to propositions, more or less).[2] There is no reason to think that Frege was a linguistic deflationist about *sentential* truth. Frege never expresses a view about the meanings of sentences like "'Snow is white' is true", probably because, as Sir Peter Strawson (1950, pp. 129–131) makes clear, attributions of truth to sentences (let alone to utterances) are extremely uncommon in ordinary language. In so far as Frege was a deflationist at all, then, he was a deflationist about *propositional* truth, not about sentential truth. And, in the context of Frege's semantics, the notion of truth that is in play is one that applies not to thoughts but to sentences. So Frege is no kind of deflationist about the notion of truth that plays a role in his semantic theory, that is, in the theory of truth he develops for his formal language.

A semantic theory of the sort Frege was the first to develop is a theory of truth, however, only in Tarski's sense, not in the sense in which the coherence theory is a theory of truth. But one need not think it is possible, or even desirable, to have a theory of truth in that sense—the sense relevant to metaphysical deflationism—to think that the notion of truth might do interesting and useful work (see e.g. Davidson, 1990).[3] Truth-conditional semantic theories are *about* truth just as they are about sentences and other linguistic items, and if the role truth plays in such theories cannot be revealed as purely expressive, then conceptual deflationism is false. Now, questions about the role that truth plays in semantic theory are, as we have already said, questions about the truth and falsity of *sentences*,[4] so what we need to ask is whether linguistic deflationism about *sentential* truth provides us with the resources to unmask the use of truth in semantics as purely expressive. To put it differently, the question is whether deflationism about sentential truth is consistent with taking semantics seriously.[5]

I shall henceforth use the now common term 'disquotationalism' for deflationism about sentential truth. The term comes from W. V. O. Quine:

> By calling ["snow is white"] true, we call snow white. The truth predicate is a device of disquotation.... We need it to restore the effect of objective reference when for the sake of some generalization we have resorted to semantic ascent.
>
> (Quine, 1986, p. 12)

The linguistic part of the disquotationalist thesis is thus that the two sentences

(1) Snow is white.
(2) "Snow is white" is true.

are not just materially equivalent but equivalent in some much stronger sense that makes the truth-predicate "dispensible when attributed to sentences that are explicitly before us" (Quine, 1987, p. 214).[6] Quine would never call such pairs of sentences 'synonymous', of course, and he denies that (1) and (2) are necessarily equivalent (Quine, 1956, p. 187). What he says instead is that "[a]scription of truth just cancels the quotation marks" (Quine, 1990, p. 80). Hartry Field (1994, p. 250) expresses a similar idea when he says that (1) and (2) are "fully cognitively equivalent", and, unlike Quine, Field (1994, p. 258) explicitly regards their equivalence as a "conceptual necessity".[7]

This is an extremely strong claim, and one that has some very odd consequences. As Field (1994, §9) both notes and emphasizes, for example, the following is true if the truth-predicate is read disquotationally:[8]

(3) Even if "snow" had meant *grass*, the sentence "snow is white" would still have been true.

That is because it is equivalent to:

(4) Even if "snow" had meant *grass*, snow would still have been white.

And that is because, as Field (1994, p. 266) remarks, echoing Quine, "to call 'Snow is white' disquotationally true is simply to call snow white," whether or not we are inside a modal context. But surely (3) is false.[9]

This is not a dispensible feature of the truth-predicate as a disquotationalist understands it. It is, in fact, central to conceptual disquotationalism, that is, to how disquotationalists understand the 'expressive' function of the truth-predicate. Suppose, for example, that I were to say:

(5) The axioms of Euclidean geometry are not all true, but they might have been.

One might think that (5) is true for the boring reason that the sentences that express the axioms of Euclidean geometry might have meant something else. But, or so Field (1994, p. 265) argues, if "true" as it occurs in (5) is understood in accord with the tenets of disquotationalism, then (5) says that Euclidean geometry is contingently false. In particular, (5) is supposed to express exactly what:

(6) It is not the case that EG, but it might have been the case that EG.

expresses, where "EG" abbreviates the conjunction of the axioms of Euclidean geometry. As it happens, there are infinitely many such axioms,[10] so we cannot actually write that conjunction down. But that, say disquotationalists, is precisely what makes the disquotational truth-predicate so useful: It allows us to express what (6) does without having to write out an infinite conjunction (see e.g. Quine, 1970, pp. 11–13).

Similarly, if I were to say:

(7) It is sometimes possible to see objects behind the sun because the axioms of Euclidean geometry are not all true.

then that is supposed to be a way for me to affirm that the non-Euclidean character of space is responsible for the somewhat surprising behavior of photons, not to make the absurd claim that optics is beholden to the semantics of English. But if that is to be so, then "The axioms of Euclidean geometry are all true" must simply express what the axioms of Euclidean geometry jointly do and, in particular, must not have any extra, 'semantic' content (Field, 1994, p. 266). If ⌜"*A*" is true⌝ *did* have some content beyond that of *A* itself, then that extra content might figure in causal explanations that invoked the notion of truth. It would not yet follow that truth *did* have some explanatory role to play (i.e., that conceptual disquotationalism was false), but the usual strategy for showing that truth doesn't play such a role would fail.[11]

Consider, for example, the standard response to the so-called success argument. We often explain people's ability to satisfy their desires in terms of the truth of their beliefs, so it looks as if truth is playing an explanatory role here. For example, we might explain how Alex managed to satisfy their desire for a beer in terms of the truth of their belief that there were beers in the cooler. Deflationists regard the mention of truth here as gratuitous: If what allowed Alex to satisfy their desire was the fact that their belief that there were beers in the cooler was true, then what allowed them to satisfy their desire was really just the fact there were beers in the cooler; the truth-involving explanation reduces to an object-level explanation. But if *A* and ⌜"*A*" is true⌝ are not intersubstitutable inside the scope of "because", then this reduction fails.[12] Of course, there might be other ways of resisting the success argument. But the most common strategy for doing so will have failed.

To summarize, disquotationalists have generally regarded *A* and ⌜"*A*" is true⌝ (where "true" is read disquotationally) as equivalent in some sense strong enough to license intersubstitution in modal and causal contexts. That view flows from their commitment to conceptual disquotationalism, since, if ⌜"*A*" is true⌝ had some content beyond that of *A* itself,[13] that would open up the possibility that truth might have some important theoretical role to play. In that sense, disquotationalism is an heir to the redundancy theory: As we have already seen Quine put it, "true" would be *dispensible* were it not for our need to make certain kinds of generalizations.

Now, one might have wanted to say instead that the reason (5) expresses the contingency of what the axioms express is because the 'axioms' are not sentences but what those sentences express. That is, truth is being predicated not of sentences but of propositions (Heck, 2004, §2). But this is not a line that a disquotationalist can take, since invoking propositions threatens to commit us to a substantial notion of representational content, which is part of what disquotationalism opposes (Field, 1994, pp. 266–267). A disquotationalist precisely does not want to understand attributions of truth to sentences in terms of the truth of the proposition the sentence expresses. Propositions (in anything but a pleonastic sense) are anathema to disquotationalism. The disquotationalist view, rather, is that attributions of truth to sentences are primitive, and they are to be understood in terms of disquotation.[14]

Still, it is often useful, when one is trying to understand what a disquotational truth-predicate is supposed to be, to compare it to a propositional truth-predicate. The claim that *A* and ⌜It is true that *A*⌝ are equivalent in some very strong sense seems reasonable.[15] For that reason, this sentence:

(8) Even if "snow" had meant *grass*, it would still have been true that snow was white.

is unproblematically true. But the disquotationalist's (3) is intended to be equivalent to (8), though it uses a sentential truth-predicate rather than a propositional one, so as to avoid the commitment to propositions. When "true" is read disquotationally, then, ⌜"*A*" is true⌝ is supposed to be just as obviously, and just as strongly, equivalent to *A* as ⌜It is true that *A*⌝ is.[16]

As noted above, the underlying point is that, if ⌜"*A*" is true⌝ had some content beyond that of *A* itself, then that content could well be essential to putative explanations in which attributions of truth appeared. It would not follow that conceptual disquotationalism was false, but the typical strategy for establishing it would no longer be available. Uses of the truth-predicate could not simply be eliminated in the way disquotationalists propose, even when truth was ascribed to a single, explicitly specified sentence, let alone when it was used in generalizations, as it is in semantics.

A simple example of such a generalization is:

(9) ⌜*A* and *B*⌝ is true iff *A* is true and *B* is true.

Following Field (2005), I shall call such generalizations 'compositional principles'. And the main question I want to discuss in the remainder of this chapter is how disquotationalists should understand the use of the truth-predicate in such principles, which are central to certain sorts of semantic theories.[17] Even in (9), the use of the truth-predicate is supposed to be 'purely expressive'. The first question we shall consider below is: How so exactly?

I shall argue in §6.1 that the truth-predicate is not being used in (9) to express an infinite conjunction, as is often suggested. I shall then turn, in §6.2, to a prior question, namely, what right disquotationalists even have to compositional principles. As we shall see, Field (2005) has offered an answer, one that also yields an answer to the question what expressive role a disquotationalist should regard truth as playing in those principles. I shall show in §6.3, however, that Field's method for generating compositional principles *over*-generates and then argue, in §6.4, that this reveals a deeper problem, which is that what Field's method yields simply are not compositional principles as they are understood in semantic theory. In particular, part of what (9) is typically understood to express is that conjunction is truth-functional, and Field's account would apply whether or not "and" was truth-functional.

Let me acknowledge something in advance. I shall be arguing that disquotationalism is committed to claims that may seem so absurd that no one could possibly accept them. This is worrying. But they are no more absurd than (3), in the end, and they flow from exactly the same source: The insistence that the primary function of "is true" is simply to erase quotation marks. That, obviously, is the very core of disquotationalism.

6.1 What Expressive Role Does "True" Play in Compositional Principles?

As mentioned above, disquotationalists regard the truth-predicate as merely an 'expressive' device. People sometimes put this point by saying that the truth-predicate allows us to make generalizations we could not make without it. But the slogan can't just mean that. *Every* predicate allows us to make generalizations we could not make without it. For example, the predicate "blue" allows us to express the generalization "All pigs are blue", which we could not express if we did not have the word "blue" in our language (or a synonym). So the disquotationalist slogan must mean something else. Which, of course, it does.

When disquotationalists characterize the truth-predicate as a 'device of generalization', what they mean is that it gives us a way to express generalizations all of whose instances we can already assert. It's just the generalization that we can't assert. That distinguishes the cases involving "true" from the case of "All pigs are blue". Not even the instances of "All pigs are blue" can be asserted without the use of "blue". More precisely still, the truth-predicate is supposed to act as a 'device of infinite conjunction' (Field, 1994, §5; Halbach, 1999). In the cases of interest, we shall be able to assert the various instances of some generalization, but, because there are infinitely many such instances, we cannot actually assert them all, absent some mechanism for forming infinite conjunctions, which is exactly what the truth-predicate is supposed to give us—much as in the cases of (5) and (7) above.

Consider, for example, the law of excluded middle:

(10) Every sentence of the form $A \vee \neg A$ is true.

The disquotationalist's suggestion is that (10) means no more and no less than that either Bill smokes or Bill does not smoke, and either Fred runs or Fred does not run, and so forth. The individual instances—"Either Bill smokes or Bill does not smoke", "Either Fred runs or Fred does not run"—have nothing to do with truth. To express the generalization, of course, what we need to do is quantify over these instances. What we seem to want to say is thus something like:

(11) For all S, if S is of the form $A \vee \neg A$, then S.

But that, familiarly, is ill-formed, since S is occupying both term and sentence positions. What the truth-predicate does, according to disquotationalists, is to solve this syntactic problem by converting the position occupied by a sentence into one occupied by a term: "Fred runs or Fred does not run" is replaced by "'Fred runs or Fred does not run' is true", and our generalization can then expressed as:

(12) For all S, if S is of the form $A \vee \neg A$, then S is true.

All we're trying to do here, or so disquotationalists say, is to wrap the instances of $A \vee \neg A$ into a neat package and affirm them, all at once. The role the truth-predicate is playing is thus purely grammatical, and (12) means no more than what (11) was supposed to mean: It simply expresses the infinite conjunction that either Bill smokes or Bill does not smoke, and either Fred runs or Fred does not run, and so forth. In particular, (12) has no more to do with truth, semantics, or logic than would that infinite conjunction, if only we could write it down.

There are several difficulties with this idea. First of all, it is a delicate matter just how we should understand the claim that such sentences as (10) 'express' infinite conjunctions. As Anil Gupta (1993, §III) argues, "express" here has to be understood in a very strong sense—which makes it surprising that so little effort has been made to articulate what that sense is.[18] The only serious attempt to do so, which is due to Volker Halbach (1999), cannot be regarded as successful (Heck, 2004, §3). Worse, as Gupta also notes, this sort of proposal appears to conflate the truth of a generalization with the joint truth of its instances. The statement that all sentences of the form $A \vee \neg A$ are true has nothing to do with which instances of that form happen to be present in the language. On the contrary, excluded middle is supposed to be a law: It is supposed to be, in the usual sense, 'projectible'. Even if new sentences are added to our language, so that $A \vee \neg A$ comes to have new instances, those too are required to be true.

I shall not pursue these complaints further here, however, as sympathetic with them as I may be. There is a more important point in the vicinity: Compositional principles simply are not statements in which the truth-predicate is plausibly being used as a device of infinite conjunction.[19]

The problem is very simple. The only 'infinite conjunction' we might plausibly take (9) to express is far too weak to have any chance of being what it actually does express. Formalize (9) as:

(13) $\forall x \forall y[\mathsf{T}(\ulcorner x \wedge y \urcorner) \equiv \mathsf{T}(x) \wedge \mathsf{T}(y)]$

Which sentences not involving the truth-predicate are supposed to count as instances of the infinite conjunction allegedly expressed by (13)?[20] It is hard to see what they might be if not sentences of the form: $A \wedge B \equiv A \wedge B$, so that (13) expresses the infinite conjunction of such claims as "Fred runs and Bill walks iff Fred runs and Bill walks", that is, the infinite conjunction of a bunch of instances of $p \equiv p$. That might sound like music to the disquotationalist's ears. But it should not.

Consider these two generalizations:[21]

(14) $\forall x \forall y[\mathsf{T}(\ulcorner x \wedge y \urcorner) \equiv \mathsf{T}(\ulcorner x \wedge y \urcorner)]$
(15) $\forall x \forall y[\mathsf{T}(\ulcorner x \wedge y \equiv x \wedge y \urcorner)]$

If (13) is supposed to express an infinite conjunction, then presumably these do, as well. And there appears to be no option but to take them, too, to express the conjunction of all sentences of the form: $A \wedge B \equiv A \wedge B$. But these are all very different. For example, (14) is logically valid. And, while (15) is not itself logically valid, it follows from any set of principles entailing that all instances of a truth-functionally valid schema are true (for example, that very principle).[22] But (13), together with similar principles, at least, has significant logical strength (Heck, 2015, 2018a).

The point, then, is that (13), (14), and (15) have very different logical properties. It follows, or so it seems to me, that they cannot all express the same infinite conjunction. On the contrary, at most one of them can express the infinite conjunction of sentences of the form: $A \wedge B \equiv A \wedge B$. But if any of these expresses that infinite conjunction, then surely it is (15), which does so in precisely the sense in which $\forall x[\mathsf{T}(\ulcorner x \vee \neg x \urcorner)]$ expresses the conjunction of instances of the law of excluded middle. But then (13) does *not* express that infinite conjunction, and there is no other infinite conjunction that it plausibly does express.

The point applies to other compositional principles as well, such as:

(16) $\forall x[\mathsf{T}(\ulcorner \neg x \urcorner) \equiv \neg \mathsf{T}(x)]$

If (16) expresses an infinite conjunction, it can only be the conjunction of all sentences of the form: $\neg A \equiv \neg A$. But, if so, then that infinite conjunction seems equally to be what is expressed by these two generalizations:

(17) $\forall x[\mathsf{T}(\ulcorner \neg x \urcorner) \equiv \mathsf{T}(\ulcorner \neg x \urcorner)]$
(18) $\forall x[\mathsf{T}(\ulcorner \neg x \equiv \neg x \urcorner)]$

Once again, (16), (17), and (18) have very different logical properties, and at most one of them can express the infinite conjunction of all sentences of the form: $\neg A \equiv \neg A$. But (18) expresses that conjunction if anything does, so (16) does not express it.

I have heard it said in conversation that disquotationalists do not really mean that "true" is *always* used to express infinite conjunctions, but that this is its 'purpose' or 'role' in our language. But I have no idea what to make of such teleological claims. Most words do not have 'purposes', except to allow us to express whatever concepts they are used to express, and I see no reason to regard "true" as different from other words in this respect.[23] Moreover, even if the truth-predicate had been explicitly introduced so as to allow us to express infinite conjunctions, it would not follow that, once we had it, we could not use it for quite different purposes, including ones that involved truth's playing a significant explanatory role.[24] The disquotationalist therefore owes us an answer to the question what 'purely expressive' role the truth-predicate plays in compositional principles, if it is not to allow us to formulate infinite conjunctions.

6.2 Disquotationalist Derivations of Compositional Principles

Disquotationalists often seem to be of two minds about compositional principles. On the one hand, for example, Field (1994, p. 269) insists that "compositional principles have no interest in their own right". On the other hand, however, Field is of course aware that there are those who find such principles as:

(9) $\ulcorner A$ and $B \urcorner$ is true iff A is true and B is true.

to be of substantial interest. So he wants to be able to explain both why such principles are true, when they are, and why they are nonetheless insubstantial and can do no explanatory work.

There is a history here. In his book *Truth*, which helped launch contemporary deflationism, Horwich (1990) suggested that we can get everything we need to know about truth from the 'minimal' theory that consists just of the T-sentences.[25] It is obvious, however, that a theory of truth containing just the T-sentences is very, very weak.[26] Without such principles as (9), it's hard to see how such a truth-predicate could be of much use at all. It certainly could not be used for the sorts of purposes for which the truth-predicate is typically used in logic.[27]

One might well think, however, that the compositional principles can be derived from the T-sentences—or, better, from the 'T-scheme', thought of as an axiom scheme having the T-sentences themselves as instances. The argument proceeds as follows:[28]

(i) "*A* and *B*" is true iff *A* and *B*.
(ii) "*A*" is true iff *A*.
(iii) "*B*" is true iff *B*.
(iv) "*A* and *B*" is true iff "*A*" is true and "*B*" is true.

The first three steps are delivered by the T-scheme; the last then follows by simple propositional reasoning.

Prima facie, this argument has two serious problems. First, it does not appear to be an argument at all. An argument consists of a sequence of claims allegedly related in some relevant way (e.g., deductively). But this 'argument' does not consist of a sequence of claims, and it is not at all obvious how to interpret it. This leads to the second problem: In so far as one does have some idea how to interpret this argument, one wants to regard "*A*" and "*B*" as variables. But then these variables appear both inside and outside quotation marks, something that is usually regarded as problematic.

One charitable way to interpret the 'argument' is to regard it not as an argument but as an argument schema. So understood, however, the argument fails to show what its proponents claim it shows. What the schematic argument shows is that we can prove every instance of (9). That, as Gupta (1993, p. 67) emphasizes, is not at all the same thing as being able to prove (9) itself. Indeed, if our background logic is first-order logic, or some other logic for which the compactness theorem holds, the generalization (9) cannot follow from its infinitely many instances. Otherwise, it would have to follow from finitely many of them, which it obviously does not (Shapiro, 1998, p. 496).

One might think the disquotationalist can simply concede this point.[29] What the disquotationalist needs to show is that an 'insubstantial' theory of truth can do the work for which a 'substantial' theory of truth is supposed to be needed. The suggestion, then, in response to the foregoing, would be that what needs amending is just the proposed content of that 'insubstantial' theory: It should not contain just the T-sentences, or something of the sort, but also the sorts of generalizations we've been discussing. That is: The disquotationalist should just claim the compositional principles as their own.

This reply begs the question whether the disquotationalist actually has a right to the compositional principles. Worse, it is utterly *ad hoc*. The suggestion is that we should add generalizations like (9) to the 'minimal' theory containing just the T-sentences. But which generalizations are 'like' (9)? Discussion of these issues tends to idealize by taking the language to which the truth-predicate applies to be a formal (usually first-order) language. And, in that case, we know well enough which compositional principles will be needed to allow us to make the generalizations the truth-predicate typically allows us to make.[30] But disquotationalism isn't a view about truth as applied only to the sentences of formal languages.[31] Truth, at least as it is used in semantics, applies to sentences of natural language. But then there is no clear limit to the sorts of compositional principles we will need to add. Indeed, if we take the case of first-order languages as our model, then, in that case, what we need to add to the minimal theory, to make it do the work we need it to do, is a full-blown Tarski-style theory of truth, as Field (1999, pp. 534–535) himself has noted. That makes me suspect that, in the case of a natural language, what would be needed is what, from a different point of view, would be regarded as constituting a full semantic theory for the language in question. That makes the question what right a disquotationalist has to compositional principles pressing once again.[32] Compositional principles are the very principles that semantic theories articulate, and truth appears to play a substantial role in such theories. If so, then "Disquotationalism doesn't work without the compositional principles, so we'd better add them" looks worryingly like another way of saying "Disquotationalism doesn't work, so we shouldn't be disquotationalists".

If something like the schematic argument rehearsed above could be resuscitated, however, then the problems we have been discussing would vanish. In light of our discussion in §6.1, however, the disquotationalist must abandon the claim that the compositional principles express infinite conjunctions; in light of Gupta's criticisms, they must also abandon the claim that compositional principles simply follow from their instances. But the possibility is still open, at least in principle, that the compositional principles should follow from certain general principles about truth that the disquotationalist anyway accepts. In particular, compositional principles might follow from the T-scheme, itself understood as having a certain kind of generality, rather than simply as a convenient way of summarizing the infinite list of its instances. And a form of this proposal has been developed by Field (2005).

The contrast to which I have just alluded, between two ways of understanding the T-scheme, is one that has been of significant interest to philosophers of logic and mathematics. Consider, for example, the induction scheme of Peano arithmetic:

$$(19)\quad A(0) \wedge \forall x(A(x) \rightarrow A(Sx)) \rightarrow \forall x A(x)$$

There are two ways of understanding this scheme. One is to take it as simply a convenient shorthand for the infinite list of 'induction axioms' of PA. The other is to regard it as having a kind of generality, so that the scheme itself is, in some sense, the real axiom.

The contrast emerges when we consider what ought to happen when we expand the language of PA, say, by adding a function-symbol for exponentiation.[33] If we understand the induction scheme in the first way, so that it is just a compact way of summarizing the real axioms, of which there are infinitely many, then it is of no significance that the expansion of the language introduces new sentences of the same form as the induction axioms we already accept. That the axioms we accept have a common form is of no interest beyond the fact that it permits such a compact summary of them. We might call such a conception of the induction scheme *static*.

On the other way of understanding the scheme, (19) is not just a compact way of listing a bunch of axioms. The common form of those axioms is precisely what is of interest. The fact that the expanded language contains new sentences of that form then *does* provide us with new axioms. The scheme is thus *dynamic* or, as it is more often put, 'open-ended'.

It is the static conception that is usually regarded as the 'official' one in mathematical logic, but Solomon Feferman (1996) has shown that the dynamic conception can be made to do mathematical work.[34] The dynamic conception has also been put to philosophical work by Vann McGee (1997) and to joint philosophical and technical work by me (Heck, 2011, 2018b). It is probably fair to say that the dynamic conception remains controversial. But it is also fair to say, or so it seems to me, that it is actually the more natural of the two. And it is, I think, pretty clearly what the founders of modern logic (e.g., Zermelo) had in mind.

Moreover, it is the dynamic conception that articulates how a disquotationalist should, and most disquotationalists do, understand the role of the T-scheme. The T-scheme is not supposed to be a static summary of a bunch of principles that apply only to our language as we now have it, so that the introduction of a new expression would give us no reason to accept the new instances of the T-scheme that then arise. On the contrary, disquotationalists understand the T-scheme as a general principle—indeed, as *the* general principle—that governs the use of the truth-predicate (Field, 1994, p. 266, n. 20). To put the point in terms of conceptual role semantics, which many disquotationalists seem happy to adopt,[35] the T-scheme summarizes a disposition that competent users of the truth-predicate must have, namely, to infer a sentence from an attribution of truth to that sentence, and conversely (or, more strongly, to substitute one for the other). So the T-scheme, as disquotationalists understand it, is 'open-ended', generating new instances of itself as the language evolves.[36]

Given this understanding of the T-scheme itself as a general principle, it is natural to wonder if there isn't some way of reasoning with it *as* a general principle so as to establish other general principles from it. Such reasoning is precisely what the schematic argument on behalf of (9) was attempting. As we saw, there are several problems one might have with that reasoning. But something like this reasoning is, in fact, quite common. I've come to realize, in fact, that I have sometimes given arguments of the same sort myself.[37]

The argument I have in mind is a well-known argument for the claim that the T-scheme fixes the extension of the truth-predicate. Here is the argument. Suppose that $\mathsf{T}(x)$ and $\tau(x)$ both satisfy the T-scheme. Then:

(i) $T(\ulcorner A \urcorner) \equiv A$, since $\mathsf{T}(x)$ satisfies the T-scheme.
(ii) $\tau(\ulcorner A \urcorner) \equiv A$, since $\tau(x)$ satisfies the T-scheme.
(iii) $T(\ulcorner A \urcorner) \equiv \tau(\ulcorner A \urcorner)$, by propositional logic.

Since this holds for any sentence A, $\mathsf{T}(x)$ and $\tau(x)$ have the same extension (on their common range).

It's a nice question how such arguments should be understood. One might suspect that the argument tacitly uses the notion of truth. I would not be unsympathetic. But, dialectically, I doubt this sort of worry will get much traction. The argument just rehearsed does not seem to employ the notion of truth, but to make perfectly good sense in its own right. And Field (2005, §3) offers a detailed account of the sorts of principles that might govern such arguments, principles that suffice, he claims, to allow for proofs of all the compositional principles comprising a full truth-theory for any first-order language. In the case of the schematic argument for (9), the thought is that the various steps of the argument are, as Field puts it, "part of the language", in perfectly good order as they are. They just contain free schematic variables. And once we have reached the conclusion:

(iv) "*A* and *B*" is true iff "*A*" is true and "*B*" is true.

we may infer

(v) For all sentences *S* and *T*, $\ulcorner S \text{and} T \urcorner$ is true iff *S* is true and *T* is true.

by a principle allowing for the replacement of schematic letters that are everywhere within quotation marks by objectual variables ranging over sentences. The details are a little messy, but not that bad.[38]

For our purposes, the more important point is that Field's interpretation of such arguments also yields an answer to the question how disquotationalists should understand the expressive role played by the truth-predicate in compositional principles: It is 'syntactic sugar' that allows what is really substitutional quantification to appear as objectual quantification.[39] Indeed, Field (1994, p. 259) suggests that the logic of schematic arguments "corresponds to a very weak fragment of a substitutional quantifier language".

The problem, then, is not that 'schematic reasoning' cannot be used to establish principles like (9). The problem, as we shall see in the next section, is that this kind of argument works too well.

6.3 Schematic Reasoning Over-Generates

There is another argument that Field might have given for (9):

(i) "*A* and *B*" is true iff *A* and *B*.
(ii) "*A* and *B*" is true iff "*A*" is true and "*B*" is true.
(iii) For all sentences *S* and *T*, ⌜*S* and *T*⌝ is true iff *S* is true and *T* is true.

The first two steps are justified by the disquotational character of the truth-predicate; the last, by principles governing schematic reasoning.

Once one notices this sort of argument, however, it becomes apparent that it can equally well be used to prove all sorts of other things, e.g.:

(20) For all sentences *S* and *T*, ⌜*S* because *T*⌝ is true iff *S* is true because *T* is true.

Thus:

(i) "*A* because *B*" is true iff *A* because *B*.
(ii) "*A* because *B*" is true iff "*A*" is true because "*B*" is true.
(iii) For all sentences *S* and *T*, ⌜*S* because *T*⌝ is true iff *S* is true because *T* is true.

Here again, the first two steps are justified by the disquotational character of the truth-predicate; the last, by principles governing schematic reasoning.

One might think it obvious that this last argument should fail. Consider the adaptation of the original schematic argument to the case of "because":

(i) "*A* because *B*" is true iff *A* because *B*.
(ii) "*A*" is true iff *A*.
(iii) "*B*" is true iff *B*.
(iv) "*A* because *B*" is true iff "*A*" is true because "*B*" is true.

That certainly does fail, since substitution of material equivalents is not permitted inside intensional contexts. Of course, for the disquotationalist, the first three biconditionals all hold "of conceptual necessity ... in virtue of the cognitive equivalence of the left and right hand sides" (Field, 1994, p. 258). But conceptual necessity, by itself, would not necessarily justify the inference to (iv). It is not even clear whether cognitive equivalence would justify it.[40] But, as we saw earlier, the whole point of a disquotational truth-predicate is to allow us to make certain sorts of generalizations we could not make without it. This includes such cases as

(7) It is sometimes possible to see objects behind the sun because the axioms of Euclidean geometry are not all true.

and related sentences that occur in the success argument. But (7) will not mean what the disquotationalist wants it to mean unless *A* and ⌜"*A*" is true⌝ are intersubstitutable within the scope of "because". This intersubstutability principle is what is driving the argument for (20), so it ought to be acceptable to a disquotationalist.

This will no doubt seem odd, but it is a familiar oddity. One might equally well have thought that substitution of ⌜"*A*" is true⌝ for and by *A* itself should not be permitted in modal contexts. But, as we saw earlier, Field (1994, p. 265) is explicit that it must be, and for good reason. So we can also prove a compositional principle for necessity:

(21) For all sentences *S*, ⌜Necessarily, *S*⌝ is true iff, necessarily, *S* is true.

Thus:

(i) "Necessarily, *A*" is true iff, necessarily, *A*.
(ii) "Necessarily, *A*" is true iff, necessarily, "*A*" is true.
(iii) For all sentences *S*, ⌜Necessarily, *S*⌝ is true iff, necessarily, *S* is true.

Again, that the first step of this argument is legitimate, if "true" is read disquotationally, simply follows from what Field himself argues is required if "true" is to play the generalizing role he thinks it plays: ⌜"*A*" is true⌝ must be substitutable for and by *A*, even when it occurs in a modal context, lest

(5) The axioms of Euclidean geometry are not all true, but they might have been.

not mean what it is supposed to mean.

The same point emerges if one reflects on the fact, mentioned at the end of the last section, that, for a disquotationalist, the truth-predicate functions essentially as syntactic sugar sprinkled over an underlying substitutional quantifier. On that interpretation, (9) amounts to

$$\Pi S\,\Pi T\,\Pi U[\text{‘}U\text{’} = \text{‘}S \text{ and } T\text{’} \rightarrow U \equiv (S \text{ and } T)]$$

and (20) amounts to:

$$\Pi S\,\Pi T\,\Pi U[\text{‘}U\text{’} = \text{‘}S \text{ because } T\text{’} \rightarrow U \equiv (S \text{ because } T)]$$

There is nothing whatsoever wrong with either of these—other than that they are trivialites, facts not of semantics but of orthography.[41]

Field (2005, pp. 23–24) discusses a closely related point in connection with the question whether his schematic treatment extends to belief attributions. The particular question at issue there is whether schematic reasoning can be used to prove:

(22) For all sentences S and names N, ⌜N believes that S⌝ is true iff N believes that S is true.

Field offers various reasons to doubt that (22) can in fact be proven schematically and expresses some doubt about (22) itself. But, for the reasons already given, I find Field's discussion hard to align with his disquotationalist commitments.[42] From a semantic point of view, (22) is extremely dubious. But from the disquotational point of view, surely (22) ought to be correct. It has to be correct if the truth-predicate is to play the expressive role disquotationalists think it plays. Suppose I want to affirm John's belief in the Euclidean character of space. Then I might say, "John believes that the axioms of Euclidean geometry are all true". If that is not to be a comment on John's beliefs about the semantics of English, then ⌜"A" is true⌝ has to be equivalent to A,[43] even inside hyperintensional contexts (cf. Field, 1994, pp. 265–266). So (22) seems to be provable in the usual way:

(i) "S believes that A" is true iff S believes that A.
(ii) "S believes that A" is true iff S believes that "A" is true.
(iii) For all sentences S and names N, ⌜N believes that S⌝ is true iff N believes that S is true.

Obviously, this strategy is going to generalize.

Field (2005, p. 24) also expresses concern about the last step of the argument, since it "seems to depend for its plausibility on the assumption that we can unproblematically quantify into" the appropriate sort of context. But there is nothing wrong with quantifying into intensional or even hyperintensional contexts. We do so all the time in ordinary language.[44] If 'quantifying-in' seems puzzling, it is because of other theoretical commitments we have. Specifically, the problems connected with quantifying-in arise *because of our commitments regarding the semantics of quantification*: We want to regard the quantified variable as ranging in such cases over ordinary objects (people and planets) rather than intensions or modes of presentation. Even more fundamentally, the reason that quantifying-in is a problem is simply that we think of variables as having values. If one thought of the truth of a quantified statement in terms of the truth of its instances, then, as Ruth Barcan Marcus (1972) was fond of pointing out, there would be no problem. Field cannot have it both ways. He cannot both suggest that we regard questions about the semantics of attitude attributions as "misguided" and then invoke the very problems that motivate such questions when trying to avoid unwelcome consequences of his own view.

But whatever the status of (22), it should be clear that few of Field's worries about it carry over to (20) and (21). There is no obstacle whatsoever to quantifying into causal or modal statements. And the suggestion—which we did not discuss—that (22) might fail because N "may have peculiar beliefs about truth" (Field, 2005, pp. 23–24) has no analogue in those cases. While my own view, then, is that disquotationalists are committed to (22) as well as to (20) and (21), it is enough for what follows if they are committed only to the latter two. Indeed, it is enough if they are committed just to (21), though I shall focus in what follows on (20).

6.4 Truth-Functionality

6.4.1 Compositional Principles and Truth-Functionality

Friends of semantics will have greeted

(20) For all sentences S and T, $\ulcorner S$ because $T\urcorner$ is true iff S is true because T is true.

with bemusement. But what exactly is supposed to be wrong with it? Why does it feel so different from

(9) $\ulcorner A$ and $B\urcorner$ is true iff A is true and B is true.

and related principles? As (9) is generally understood by friends of semantics, what it says, in part, is that the truth-value of a conjunction is entirely determined by the truth-values of its parts, i.e., that conjunction is truth-functional. It is the truth-value *of the conjuncts* that matters, but what the conjuncts contribute is just their truth-value and nothing else. Similarly, then, (20), as understood by friends of semantics, would tell us that the truth-value of "A because B" is wholly determined by the truth-values of A and B and the causal relationship between those truth-values (e.g., by whether the True because the False). That is barely coherent, but never mind. What the provability of (20) really shows us is that the proof of (9) goes through whether or not "and" is truth-functional. So, if (9) is understood in the sense in which it is schematically provable, it does not affirm the truth-functionality of conjunction; so understood, then, it is not the compositional principle for "and" as friends of semantics would understand it; *that* principle is *not* schematically provable.

That went by pretty quickly. Let me fill in some details.

Strawson (1952, pp. 79–82) famously denied that “and” is truth-functional, on the ground that there is a temporal aspect to its meaning.[45] A sentence like

(23) They got married, and they had a baby.

could be true, Strawson claimed, even though

(24) They had a baby, and they got married.

was false. But how this issue is decided has no effect whatsoever on whether (9) is true when the truth-predicate is read disquotationally. If truth is disquotational, then ⌜“*A*” is true and “*B*” is true⌝ just means what ⌜*A* and *B*⌝ means which is just what ⌜“*A* and *B*” is true⌝ means, and that is the end of it, no matter what “and” actually means. If that were not so, then the truth-predicate could not be used, in the context of a conjunction, for the expressive purposes for which the disquotationalist thinks we need it. It follows that the sorts of arguments we have been examining do not allow the disquotationalist to prove the compositional principles in the sense that friends of semantics understand them. In particular, this sort of reasoning cannot be used to demonstrate the truth-functionality of conjunction, which I take to be one of the central semantic facts about it—if, indeed, it is a fact about it.

I am assuming, of course, that whether “and” is truth-functional is an important semantic issue, one that will affect the form of the compositional principle we accept for it. Foes of semantics might not agree, but my goal here is not to convince them of the virtues of semantics. My goal is to argue that the notion of truth, *as it appears in semantics*, is not playing the merely expressive role that a disquotational truth-predicate plays: One cannot understand the use of the truth-predicate, *in semantics*, as purely disquotational, as just a tool of ‘semantic ascent’. Rather, the notion of truth is doing serious theoretical work, and that work is nowhere more visible than in disputes over truth-functionality. The fact that “and” is truth-functional (if, again, it is a fact) is supposed to explain certain aspects of the behavior of that word, and the fact that “because” is not truth-functional is supposed to explain some of the ways in which it is unlike “and”. So truth has a robust, explanatory role to play in semantics.

A disquotationalist might respond that truth-functionality is, to be sure, an important phenomenon,[46] but that it is not a semantic but a logical phenomenon, one that has to do with what sorts of inferences are valid, not with whether (9) is true. Presumably, the inferences in question would be something like:

(i) *A* and *B*.

(ii) *B* if, and only if, *C*.
(iii) Hence, *A* and *C*.

Call these 'inferences constitutive of truth-functionality'.

Even at first sight, one ought to be suspicious of this sort of inferential characterization of truth-functionality. If the biconditional is not itself truth-functional, then it may license such substitutions in cases where the connective under discussion is not truth-functional.[47] For example, if the biconditional expresses necessary equivalence, then it will license substitutions inside modal contexts. It is rarely held nowadays that the natural language conditional is truth-functional, so this sort of inference, formulated in natural language, almost certainly does not capture truth-functionality. But, even if we set that worry aside, the offered condition is easily seen to be both too weak and too strong.

To see that the condition is too weak, consider the conditional itself. Then the inference in question takes the form:

(i) If *A*, then *B*.
(ii) *B* if, and only if, *C*.
(iii) Hence, if *A*, then *C*.

This will be valid so long as the conditional is transitive, whether or not it is truth-functional, and the same goes for the case in which we substitute in the antecedent. The inferential conception would thus count any transitive conditional as truth-functional.

One might suggest, in response, that the inference should not be stated in terms of "if and only if", but in some other terms. One idea, for example, would be to take the second premise to be:

(ii′) Either *B* and *C*, or it is not the case that *B* and it is not the case that *C*.

which is disquotationally equivalent to:

(ii″) Either "*B*" is true and "*C*" is true, or "*B*" is false and "*C*" is false.

But this proposal has similar flaws. As mentioned earlier, it is controversial whether "and" is truth-functional. If it is not, then (ii′) is not truth-functional, either, and the possibility will again arise that it will support inferences it should not.[48]

But however the inference is formulated, requiring it to be valid if a connective is to be truth-functional is too strong a condition. I have mentioned several times now that it is controversial whether "and" is truth-functional. That question is *not*, however, decided simply by considering the truth-values of such sentences as (23) and (24). Even if those two sentences can have different truth-values, it does not follow that “and” is not truth-functional. The reason for the difference in truth-value might lie not in the meaning of “and” but in the interaction of the tenses on the verbs with each other and with other elements of the syntactic structure of the sentence. Such an account was developed by Barbara Partee (1984, §IV) in the context of Discourse Representation Theory, but similar accounts can be developed in other frameworks (see e.g. King and Stanley, 2004, §V).

Ultimately, of course, it is an empirical question what accounts for the temporal reading of conjunctions; that is, it is an empirical question whether “and” is truth-functional. But the point here is conceptual: One cannot read off from the truth-values of sentences in which “and” occurs whether it is truth-functional. There is too much else going on in such sentences for such an inference to be legitimate. If so, however, no inferential test of the sort we are considering can work. To steal from Hilary Putnam: Truth-functionality just ain’t a matter of inference.[49]

One might respond that, if Partee is right, then (23) and (24) aren’t, on the readings in question, actually conjunctions. They have some more complex structure. So, if we consider sentences that really are conjunctions, then the inferential test will work. But then the question is how we are supposed to tell which sentences ‘really are’ conjunctions. The point is not just that it is hard to know. The point is that, in practice, claims about what the structure of a sentence ‘really is’ are evaluated by considering how that sentence’s having a given structure would affect its meaning. One can’t even discuss this sort of question unless one has some sort of idea how structure-facts affect meaning-facts.[50] But how structure affects meaning is what semantics is about—one of the things it is about, anyway. So the question whether (23) and (24) are ‘really conjunctions’ is one that makes no clear sense outside the semantic project.

This sort of point is perhaps easiest to understand in connection with quantification. Consider, for example:

(25) Most professors know some student who hates every class.

There are several possible readings for this sentence, but it seems to me that (25) cannot mean that every class is hated by some student that most professors know.[51] Why not? The familiar answer is that, while the three quantifiers in (25) can take different scopes, not every possible ordering of the scopes is available. And surely something along those lines must be right.[52] But it is only a partial explanation, and it would be no explanation at all if we did not understand how scope affects meaning: (25) has the readings it does, and not others, because certain scope relationships are possible, and a sentence in which the quantifiers take *these* scopes has *this* meaning; one in which they take *those* scopes has *that* meaning; and to get *this other* (unavailable) meaning, the quantifiers would have to take *these other* scopes which, for some reason, they can't.

6.4.2 Schematic Reasoning and Compositional Principles

Toward the end of his paper on schematic reasoning, Field makes two suggestions:

> First, it may simply be misguided to look for compositional truth or meaning principles for attitude constructions. Second … , the fact that compositional principles of truth or meaning are straightforward for some constructions but not others is not fundamentally a fact about the application of the notion of truth or meaning to different constructions, but is simply a fact about the underlying logic of those constructions. Facts about the logic of these constructions explain the facts about how the notions of truth and *meaning that* apply to them, rather than the other way around.
>
> (Field, 2005, p. 24)

The logic of "and" and "or" permits a simple derivation of the compositional principles for them, but the logic of attitude constructions, Field thinks, does not. Now, as I have said, if the truth-predicate is disquotational, then that is wrong. But set that aside. Suppose we accept, as almost anyone would,[53] that *A* and ⌜"*A*" is true⌝ are materially equivalent. Then it looks as if schematic reasoning will permit the derivation of the compositional principle for "and", if it is true (i.e., if "and" is truth-functional), though it will not then permit the derviation of the compositional principles for "because" and "necessarily". That is, it looks as if (9) will be *provable* by schematic reasoning (if it is true), just given the very weak assumption just mentioned. How then can (9) be regarded, as friends of semantics want to regard it, as an empirical hypothesis?

I do not mean to attribute this line of thought to Field, but it is naturally suggested by his paper, and it took me a while to formulate an answer to it. So it seems worth considering, at least briefly.

To see the response, consider:

(26) "Alex swims and Tony runs" is true iff Alex swims and Tony runs.

Friends of semantics would regard (26) as one of the empirical claims in whose explanation (9) figures. But (26) is just an instance of the T-scheme. How can it need explaining at all? There seems to be a way of knowing that (26) is true that simply involves reflecting on the meaning of the word "true". But even if (26) can, in this way, be known by reflection, or even *a priori*, it does not follow that (26) is not empirical, nor that it does not stand in need of explanation. I can know by reflection that I am here now, but it is an empirical fact nonetheless, and one that can be explained.

Moreover, my own view is that the full story about how one might come to know (26) on the basis of reflection is significantly more complicated than: 'true' disquotes (Heck, 2004, §5). The complications matter. In particular, one cannot come to know (26) by this sort of reflection unless one already understands the embedded sentence "Alex swims and Tony runs". That understanding, on my view (see Heck, 2007), partially consists in knowledge of (26). That need not make the reflective knowledge circular. It just means that, once one has come to know (26) in one way—by learning to speak the fragment of English of which the sentence it mentions is a part—and one has also come to understand the word "true", then another, more 'reflective' way to know (26) also becomes available (Heck, 2004, §3).

Something similar is true of compositional principles. The semantical view is that principles like (9) are more fundamental than such T-sentences as (26): Part of the reason (26) is true is because a conjunction is true just in case both its conjuncts are, but it is no part of why (9) is true that (26) is true.[54] It is no threat to this position if, once such T-sentences are in place, one can do 'reverse semantics' and derive the more fundamental compositional principle for conjunction from the less fundamental T-sentences it partly explains by reflecting on the pattern exhibited by such T-sentences generally.[55]

It is no doubt an interesting question why such schematic derivations seem to be available in some cases but not in others. But there is an obvious sense in which Field's schematic arguments simply piggyback on genuine semantics: Field introduces deflated versions of whatever machinery is required by genuine semantics and then treats it schematically.[56] I suspect that such mimicry is also possible in other sorts of cases. The difference is simply that, in those other cases, more sophisticated sorts of semantic machinery will be involved. In many cases, there is, as yet, no semantic theory to mimic. If there were, then I'm sure Field could mimic it, too.

6.4.3 A Brief History of Truth-Functionality

I argued in §6.4.1 that truth-functionality cannot be characterized in terms of the validity of inferences. It is, rather, a semantic phenomenon, one that can only be characterized in terms of the notion of truth, which thus plays a role in semantics that is not simply expressive. The history of the notion of truth-functionality teaches us the same lesson.[57]

I used to think that the notion of truth-functionality was due to George Boole, but I was wrong. Boole does think of (what we would call) sentence-letters as having 'values', and he thinks of conjunction and the like as corresponding to operations on those values, with the values and the operations together forming (what we now call) a Boolean algebra. So Boole does think of conjunction as a *function*. But he does not think of it as a *truth*-function, because he does not think of the values of sentence-letters as truth-values. They are, rather, classes, subsets of the 'universe of discourse', which he originally regards, in *The Mathematical Analysis of Logic,* as comprised of 'cases' or 'circumstances': The value of a sentence-letter is the set of circumstances in which it is true (Boole, 1847, pp. 48ff).

By *The Laws of Thought*, Boole (1854, ch. XI, §16) had become dissatisfied with this view, because it requires "a definition of what is meant by a 'case'", which he thinks will involve us in matters beyond the bounds of logic. In this later book, then, he regards the universe of discourse as consisting of times—apparently, these are within the bounds of logic—and the value of a sentence-letter becomes the set of the times at which it is true. Other Booleans made yet other choices. But almost all the Booleans take the values of sentence-letters to be subsets of some universe of discourse.[58] The crucial advantage of this view, as the early Booleans saw it, is that hypothetical judgements can thereby be unmasked as universal affirmative propositions, relations between classes. Thus, we find Boole writing:

> Let us take, as an instance for examination, the conditional proposition "If the proposition *X* is true, the proposition *Y* is true". An undoubted meaning of this proposition is, that the *time* in which the proposition *X* is true, is *time* in which the proposition *Y* is true.
> (Boole, 1854, ch. XI, §5, emphasis original)

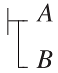
More generally, the 'calculus of judgements' (sentential logic) can, in this way, be reduced to the 'calculus of classes' (Aristotelian logic, more or less), thus unifying what might otherwise have looked like unrelated parts of logic.

The question what comprises the universe of discourse has proved not to be the crucial point. Boole's great insight was precisely that, no matter what we take the universe to comprise, if we treat the sentential connectives as expressing set-theoretic operations on its power set, then the (Boolean) algebra so determined will validate the laws of classical logic. And the flexibility inherent in Boole's approach has proven a great advantage. His original view, that the universe comprises 'cases', inspired some of the earliest work on modal logic. His later view, that it comprises times, had a similar influence on tense logic.

Boole does regard the case in which the universe contains just one element as special. Then we have a two-element Boolean algebra, with elements Boole would have denoted "1" and "0". But, even in this case, Boole does not interpret 1 and 0 as truth and falsity: They are the universe and the empty set, as they always are in his work. Boole regards this case as especially important because it makes the calculations in which he is interested especially easy.[59] But—and this is the point—none of this affects Boole's *theory* of those calculations, that is, his attempt to axiomatize the structure of a Boolean algebra, i.e., his attempt to formalize sentential logic. What I earlier called 'inferences constitutive of truth-functionality' were well-known to Boole. Identity, which functions like the biconditional, features prominently in Boole's calcuations, as do the substitutions that it licenses. But the notion of truth-functionality simply is not present, because Boole does not, as we have seen, think in terms of truth-values.

The notion of truth-functionality is not present in Frege's *Begriffsschrift*, either, though Frege does there present a complete formalization of sentential logic, and 'inferences constitutive of truth-functionality' are frequently made. Indeed, one of the rules governing the sign for 'identity of content', which acts much like the biconditional (and is written: $\equiv$), is proposition (52), which is a form of Leibniz's Law and which permits precisely the substitutions embodied in the 'inferences constitutive of truth-functionality'.

The striking fact, though, is that, in *Begriffsschrift*, Frege simply does not explain the conditional in terms of truth and falsity. His explanation reads, rather, as follows:

> If *A* and *B* stand for contents that can become judgements . . . , there are the following four possibilities:
>
> (i) *A* is affirmed and *B* is affirmed;
> (ii) *A* is affirmed and *B* is denied;
> (iii) *A* is denied and *B* is affirmed;
> (iv) *A* is denied and *B* is denied.
>
> Now
>
> 
>
>
> stands for the judgement that the third of these possibilities does not take place, but one of the other three does.
>
> (Frege, 1967, §5, emphasis removed)

Frege does not think of the conditional as expressing any kind of function in *Begriffsschrift* (Linnebo, 2003), let alone a truth-function. He seems to get the idea that it expresses a function a few years later, from Boole. The idea that it expresses a *truth*-function, however, is Frege's own, and it does not appear until 1891, in his lecture *Function and Concept* (Frege, 1984b, opp. 20ff). The crucial innovation Frege has made at that point is to introduce the notion of a truth-value—and with it the idea that sentences denote their truth-values. Then the way is open to regarding the conditional as, quite literally, expressing a truth-function: a function whose arguments and values are truth-values.

The notion of truth-functionality did not emerge, then, either from the algebraic manipulations we find in Boole nor from the codification of inference presented in *Begriffsschrift*. Though these earlier efforts no doubt help prepare the way, truth-functionality appears only within, and as a central part of, the semantic perspective that Frege embraces in his later writings.

I would suggest, in fact, not only that truth-functionality cannot be explained in terms of inference, but that even compositional principles like (9) do not really capture it. Properly to capture it, we need to follow Frege and think of sentences as having truth-values as their 'semantic values' and of connectives like "and" as operating on those values.[60] That, in fact, is often how things are done in developed presentations of truth-theoretic semantics for natural languages (e.g., Larson and Segal, 1995). The possible semantic values for sentences, truth and falsity, form a two element Boolean algebra, and the semantic values of conjunction, disjunction, and the like are the operations of that algebra. In such a treatment, it is not just the particular clause for conjunction that expresses its truth-functionality, but the semantic framework in which that clause is stated. Semantic notions then play an even more fundamental role than they do if the theory is formulated simply using compositional principles.[61]

6.5 Closing

Donald Davidson (1984) famously takes the task of semantic theory to be the construction of a compositional theory of truth that delivers such theorems as:

(27) "Snow is white" is true iff snow is white.

Ironically, though, given how frequently (27) is used to illustrate the goals of semantic theory, we do not actually know how to formulate a theory that will generate it without getting a great deal else wrong. That is, we do not have a (widely accepted) semantics for mass terms. Field (1994, p. 269) thinks there may well be none to be had and so that it is a virtue of disquotationalism that it relieves us of the need to look for one. And Field (2005, p. 24) would further deny that any explanation of the truth of (27), of the sort that semantic theory purports to provide, is required, going so far as to suggest that "it may simply be misguided to look for compositional truth or meaning principles" in such cases. After all, if disquotationalism is true, then (27) is just a verbose way of writing:

(28) Snow is white iff snow is white.

and surely no deep explanation is needed of the truth of (28). So, ultimately, it seems unsurprising that disquotationalism cannot make good sense of such compositional principles as (9), or of semantic theory more generally.

It is of course open to a disquotationalist simply to insist that the use that semantic theory makes of the concept of truth, since it is not 'merely expressive', is illegitimate. And it is not my purpose here to convince anyone of the interest of natural language semantics, nor of linguistic theory more generally. My goal has simply been to force a choice between conceptual disquotationalism and semantics by showing that the work that the notion of truth does in semantic theory cannot be regarded as 'merely expressive', even in the simplest cases. As it happens, my own view is that the insights gained over the last few decades more than suffice to demonstrate the fruitfulness of the semantic enterprise and of the central role that the thesis of compositionality has played in shaping it. To be sure, there is no *a priori* guarantee that there are compositional principles to be found for mass terms, attitude constructions, and the like. But that is simply a reflection of the fact that the compositionality of natural language is an empirical hypothesis, and a strong one (which is part of why it has proven so fruitful).[62] Quite generally, though, the difficulty of formulating a semantic theory for mass terms, or attitude constructions, or generics, or what have you, should be no more surprising than is the difficulty of formulating conditions on the use of resultatives, in syntax.

I choose, then, to embrace natural language semantics and to reject conceptual disquotationalism. But I have not argued for that choice here.

Acknowledgments

Thanks to the editors, Carlo Nicolai and Johannes Stern, for their comments on the chapter, which helped me to clarify one of its central arguments. Thanks also to John P. Burgess for conversations early in the history of the chapter that helped me to clarify some of its central ideas and to J. C. Beall for some very helpful remarks over lunch. A talk based upon this chapter was presented at a workshop on semantics held at Rutgers University in September 2012. Thanks to my commentator, Karen Lewis, whose remarks significantly improved the chapter, and to everyone present for their input, especially David Beaver, Josh Dever, Cian Dorr, Michael Glanzberg, Ernie Lepore, Zoltan Gendler Szabó, and Rich Thomason. Thanks to Ernie again for the invitation.

Notes

1. For a nice historical discussion of this aspect of Frege's position, see Textor (2010). My own view is that there is something profoundly right about that position but that, to understand it properly, we need to see it as expressed through Frege's thesis that the reference of a sentence is its truth-value and, therefore, that the sense of a sentence—the thought it expresses—is its truth-condition (Heck and May, 2018, 2020). Which brings us to the next paragraph.
2. It's also important to appreciate the true purpose of Frege's deflationary remarks: to undermine the view that the relationship between a thought and its truth-value is that of subject to predicate rather than, as Frege thinks, that of sense to reference. For discussion, see Heck (2010), Heck (2012, Part I), and Heck and May (2018, esp. §5).
3. The opponent of deflationism also does not need an alternative to linguistic deflationism. If there is no simple explanation of what the word "true" means, then that simply shows that it is like most other words. That said, from a Davidsonian perspective, the correct axiom for sentential "true" would seem to be something along the lines of (see Heck, 2004, §4):

   "*x* is true" is true of *S* iff *S* is true.

   The case of propositional "true" is more complicated only because of the presence of intensional language. But that is a separate problem.
4. Or, in a more developed setting, sentences and contexts, or utterances, or something of the sort. This complication will not be relevant here, so I will continue to speak of sentences. Indeed, disquotationalism famously has serious problems with context-dependence (Heck, 2004, §4), so ignoring it can only help my opponent.
5. It seems clear that deflationism about *propositional* truth is consistent with taking semantics seriously. Soames (1988, 1999) holds precisely such a combination of views. Semantics, as he sees it, assigns propositions to sentences (relative to contexts). Truth simply does not enter the picture. It's a more interesting question whether propositional deflationism is compatible with *truth-conditional* semantics.
6. Gupta (1993, esp. §IIII) spends a good deal of time emphasizing just how strong this equivalence needs to be. Much of what follows simply reinforces that point.
7. Strictly speaking, as Field (1994, pp. 250–251) notes, the latter sentence seems committed to the existence of the sentence "Snow is white", whereas the former sentence does not. So, officially, Field's view is that they are fully cognitively equivalent modulo that commitment. But Field himself tends to disregard this aspect of the view, and I will tend to do so as well. That said, Marian David (2005, §IV) argues that it is a more serious problem than is usually acknowledged.
8. If one is worried about the use of "means" in the antecedent, replace it by: Even if "snow" had been used the way "grass" is used, and conversely.
9. Some philosophers think sentences have their meanings essentially (see e.g. Simchen, 2012). I find the conception of 'sentence' on which such views are based to be incompatible with any plausible theory of human language comprehension. But we can set that issue aside here and simply reformulate (3) in terms of a particular uttertance made by me at some fixed time of a sentence specified purely syntactially: The claim is then that *that* utterance would have been false had "snow" meant *grass*.

[10] Assuming we are working in a first-order language. But, in the second-order case, the logic itself is not finitely axiomatizable.

[11] Here again, this point is essentially due to Gupta (1993).

[12] Gamester (2018) discusses this sort of strategy at length and argues that it fails even if the substitution is granted.

[13] As mentioned in note 7, ⌜“*A*” is true⌝ may have *some* additional content, but, if anything, that fact already poses problems for the disquotationalist’s favored reading of (5) and (7). But I’m setting that issue aside here.

[14] I do not wish to argue about labels here. If anyone does, then what I am arguing in this section is that the sort of view I am describing is particularly important, and it seems obvious that it is reasonably called ‘disquotationalism’. Why that view that is my focus here is part of what I am in the process of explaining.

[15] Modulo, as David (2005, p. 387) points out, following Moore (1953, p. 276), the commitment to the existence of the proposition that *A*. Compare note 7, again. I’ll now stop pointing out this kind of caveat. Consider it included throughout in what follows.

[16] I hope it is clear that whether the verbal string “Snow is white is true” (or the written one, with or without extra quotation marks) can be understood, in colloquial English, as equivalent to “That snow is white is true” is wholly irrelevant. We are not doing ordinary language philosophy here but are discussing the status of “true” as a predicate specifically of sentences.

[17] I will focus here on truth-theoretic semantics, but it should be clear, I hope, that nothing depends upon this restriction. The disquotationalist’s task would only be harder if we were discussing a semantic theory that made use of more complex sorts of semantic values.

[18] Of course, if such generalizations as (10) are supposed to ‘express’ claims that do not really involve truth, then attributions of truth to single sentences must do so as well: I.e., ⌜“*A*” is true⌝ must really ‘express’ a claim that does not involve truth. The only candidate is what is expressed by *A* itself. Hence, again, the redundancy of disquotational truth.

[19] I have made this point before (Heck, 2004, pp. 331–332), but in a somewhat different context, and without sufficient emphasis. Halbach (2001, p. 192, fn. 26) seems to agree with it.

[20] One might think that (13) should express an infinite conjunction of infinite conjunctions, but this does not evade the problem, since the terms of that conjunction are the same as in the cases we shall discuss, and the difference is only one of grouping.

[21] A missing bracket in “Truth and Disquotation” may have obscured what formula (12) on p. 332 was meant to be. It was what is (15) here.

[22] The notion of a tautology is definable in PA, so we can interpret a theory containing that principle in one with, say, just the T-sentences by reinterpreting $\mathsf{T}(x)$ to mean: $\mathsf{T}(x)$, in the old sense, or x is a tautology.

[23] J. L. Austin (1950, p. 122) famously regarded “true” as an “extraordinary word”. But it has never been clear to me on what ground. Words are words.

[24] I owe this point to Jamie Tappenden, who remarked in 1995 or so that, even if the extension of the truth-predicate is fixed by something like convention (T), it does not follow that we cannot go on to theorize about the set of true sentences and formulate possibly significant generalizations about it. McGee (2005, p. 144) has since made essentially the same point.

[25] Horwich’s discussion proceeds in terms of propositional truth, and so his position in *Truth* is not obviously disquotationalist. But Horwich is, in fact, committed to disquotationalism, since he is also committed to a deflationist view of meaning (Horwich, 1998).

26. Formally speaking, if we take a theory $\mathcal{T}$ and add to it all T-sentences for sentences in the language of $\mathcal{T}$, then the resulting theory is locally interpretable in $\mathcal{T}$ (Heck, 2018a, Theorem 2.1).
27. Here's an example. Let $\mathcal{T}$ be a theory. Suppose we have a theory of truth for $\mathcal{T}$ that allows us to prove the T-sentences. Then, if $\mathcal{T}$ is finitely axiomatized, we will also be able to prove, pretty trivially, that all of $\mathcal{T}$'s axioms are true. But what if $\mathcal{T}$ is not finitely axiomatized or not even finitely axiomatiz*able*? In that case, we cannot, in general, prove that *all* of $\mathcal{T}$'s axioms are true, though we can prove that *each* of them is. To prove the stronger claim, we typically need to appeal to compositional principles (Heck, 2015, §3.4).
28. This sort of argument is outlined by Field (1994, pp. 258–259), who actually discusses the case of disjunction. I'll discuss conjunction, for reasons that will become clear below. There are similar arguments to be found in the writings of many others. I learned of such arguments as a student from Sir Michael Dummett, who discusses something similar in *The Logical Basis of Metaphysics* (Dummett, 1991, pp. 56ff).
29. This sort of suggestion was made during the discussion period when I presented related material at Princeton in 2009. Cieśliński (2010, p. 412) comes close to making it in print in his discussion of the conservativeness argument against deflationism. He seems simply to assume, however, that the deflationist is entitled to the compositional principles when he adopts what he calls $\mathsf{PA}(S)^-$ as his base theory. That entitlement is precisely what I am questioning here.
30. Actually, it is not entirely clear that generalizations 'like' (9) would suffice, anyway. See, for example, the wide variety of truth-involving principles discussed by Friedman and Sheard (1987, 1988).
31. It isn't always clear whether disquotationalism is supposed to be a 'revolutionary' or 'hermeneutic' view. The point to be made next makes that moot in the present context. Still, my sense is that many disquotationalists have overlooked the importance of the issues I am discussing here because they have tended to focus on formal languages.
32. It is of course open to a disquotationalist to reply that semantics should be done in other terms, e.g., in terms of inference-rules. But surely it is an empirical question how semantics should be done. It would be odd if disquotationalism, a view defended entirely on a priori grounds, had such empirical implications.
33. The usual language of PA includes only function symbols for succession, addition, and multiplication, and exponentiation is then defined as a *relation* which can be proven (in PA, or in $\mathsf{I}\Sigma_1$, but not in $\mathsf{I}\Delta_0$) to satisfy existence and uniqueness conditions.
34. Feferman introduces the notion as a 'more natural' way of developing ideas which with he was, in one form or another, concerned throughout his career (Feferman, 1962, 1991).
35. Just to be clear: I have no interest in conceptual role semantics myself.
36. Gupta (1993, §5) expresses some doubt about whether disquotationalists can understand the T-scheme as any kind of generalization. I think this is at best a stand-off and so shall not pursue the issue.
37. There's another interesting question in this same vicinity, namely, whether schematic reasoning might help explain what justification we have for regarding all the axioms of PA as true. There is no problem about why we regard *each* of the axioms as true: We accept the axiom, we accept the T-sentence for it, and we make a simple inference. But it is much less obvious with what right we regard *all* the axioms as true. In the context of an axiomatic

theory of truth for the language of arithmetic, the proof is by induction, and the instance of induction we need necessarily involves semantic vocabulary. One might wonder if there is a different story to be told, however, along the lines we are discussing.

38. Long after this paper was completed, Leon Horsten and Graham Leigh (2017) showed that the compositional principles can be derived from so-called reflection principles. Unfortunately, I cannot consider their paper in any detail here. But many of the points to be made below apply *mutatis mutandis*. From the present point of view, what they observe is simply that reflection principles allow one to move from the observation that all instances of a scheme are provable to an assertion of that very scheme. (So their reflection principles are, in essence, an ω-rule for sentences, as the editors remarked to me.) The arguments below largely concern restrictions on and presuppositions of that sort of move.
39. This is a fairly common idea in deflationist writing (see e.g. Hill, 2002), so it is no surprise that it should surface here.
40. As Schnieder (2011, pp. 445–446) notes, if 'cognitive equivalence' means that "a speaker who understands [two sentences] normally has to adopt the same epistemic stance towards them", then it may not support substitution within the scope of "because". One might think, in particular, that "It is true that snow is white because snow is white" is true, but the converse false, even though "Snow is white" and "It is true that show is white" are cognitively equivalent. But we have already seen that disquotational truth-predicates behave in sometimes surprising ways. We are simply seeing that again. (Special thanks here to Johannes Stern.)
41. Much the same point can be made about Horwich's attempt to deflate compositionality for meaning (Heck, 2013).
42. I am independently puzzled by this remark:

    > [I]n order for [(22)] to be usable in a full compositional semantics, we'd also need other applications of substitutivity that are likewise dubious; e.g., we'd need that *S* believes that '*p* or *q*' is true if and only if *S* believes that '*p*' is true or *S* believes that '*q*' is true (Field, 2005, p. 24).

    I'm not sure what Field is thinking here—he doesn't explain further—and I cannot think of any reason myself that one would need such a principle in a compositional semantics, whether it was based upon (22) or not. In any event, the issue is whether Field is committed to (22), not what work (if any) it might do. (Possibly, what Field meant to write was: … we'd need that *S* believes that '*p* or *q*' is true if and only if *S* believes that '*p*' is true or '*q*' is true. But that is unproblematic.)
43. As I mentioned in note 7, these are not entirely equivalent: The former commits one to the existence of the sentence *A*. But it is not obvious that this commitment has to be the believer's, as opposed to the attributor's. To assume it did would be to make strong assumptions about how the content of the complement clause has to be related to the content of the belief attributed.
44. There is someone Superman knows can fly that Lois does not, because Superman knows that Clark Kent can fly, and Lois does not. But there may well not be anyone that Lois knows can fly that Jimmy does not, even though Lois knows that Superman can fly, and Jimmy does not know that Clark can fly.

45. The issue remains controversial. Relevance theorists, in particular, often deny that "and" is truth-functional (see e.g. Carston, 1988).
46. More radically, one might deny that truth-functionality is an important phenomenon at all. But, again, my claim, at present, is just that disquotationalism cannot make sense of compositional principles *as friends of semantics understand them*. So we are faced with a choice between semantic theory and disquotationalism.
47. As is often noted, this sort of inference is valid in many logics in which the connectives are not truth-functional. For example, such inferences are valid in intuitionistic logic, in supervaluational systems, and so forth, and not just for conjunction but for the other connectives, as well, even though those connectives are not truth-functional in such systems.
48. Even an inference with premises *B* and *C* runs into similar problems, since "They had a baby. They got married." can seem relevantly similar to (23).
49. Note that the same point disposes of the suggestion that a connective * is truth functional just in case whether *A* * *B* is completely determined by whether *A* and whether *B*. Even if whether they got married and whether they had a baby does not completely determine whether they had a baby and they got married, it could yet be that "and" is truth-functional.
50. This is true even in simple cases like "He saw John in the mirror". The usual claim is that "He" cannot be bound by "John". But that claim makes little sense absent the background assumption that facts about binding imply facts about meaning, in this case, that such binding implies *de jure* co-reference. As is often pointed out, "He" can perfectly well be *de facto* co-referential with "John".
51. Whereas it can mean, I think, that every class is such that most professors know some student who hates it. (This reading is more natural if "*every*" is stressed.) Even the question why that reading is less natural wants answering, and the explanation also adverts to structure.
52. Variable-free semantic theories would explain the phenomenon differently (see e.g. Jacobsen, 2014). But the remarks to follow also apply to them. Perhaps even more so, since syntax does so little work in such frameworks.
53. Modulo concerns about the paradoxes, of course, though those will affect this entire discussion and so may be set aside for the moment.
54. Indeed, (9) could be true even if (26) was not true: "Alex swims and Tony runs" might be an idiom.
55. There is a program in the foundations of mathematics known as 'reverse mathematics' (Simpson, 2009). It is sometimes said to involve deriving axioms from theorems.
56. See, for example, Field's deflationary treatment of quantification (Field, 2005, §5).
57. For defense of the interpretive claims made here, and some caveats, see Heck and May (2018, 2020).
58. Ernst Schröder (1972, p. 224) expresses this sort of view in his review of Frege's *Begriffsschrift* (1967). Hugh MacColl (1877, pp. 9–10) comes closest to the modern conception, but his official view is that the sentence-letters denote 'statements'.
59. These are the calculations that would now be done with truth-tables. Boole does not have those, however. His calculations are, instead, manipulations of algebraic formulae.
60. This point is argued in detail by Dummett in the early chapters of *The Logical Basis of Metaphysics*. He expresses it by saying that a "meaning-theory must ... incorporate a semantic theory" (Dummett, 1991, p. 63).

61. The history of quantification theory can be used to illustrate these sorts of points, as well. But, for reasons of space, I'll have to defer that discussion to another time.
62. Indeed, the precise formulation of the principle of compositionality is controversial, and it is easy to find non-compositional treatments of various constructions in the literature (cf. Dever, 2006; Szabó, 2012). But these still make serious use of notions like reference and truth.

References

Austin, J. (1950). Truth. *Proceedings of the Aristotelian Society*, 24: 111–128.

Bar-On, D. and Simmons, K. (2007). The use of force against deflationism: Assertion and truth. In Greimann, D. and Siegwart, G., editors, *Truth and Speech Acts: Studies in the Philosophy of Language*, pages 61–89. Routledge.

Boole, G. (1847). *The Mathematical Analysis of Logic, Being an Essay Towards a Calculus of Deductive Reasoning*. Macmillan, Barclay, & Macmillan.

Boole, G. (1854). *An Investigation Into the Laws of Thought*. Walton and Maberly.

Carston, R. (1988). Implicature, explicature, and truth-theoretic semantics. In R. M. Kempson, editor, *Mental Representations: The Interface Between Language and Reality*, pages 155–181. Cambridge University Press.

Cieśliński, C. (2010). Truth, conservativeness, and provability. *Mind*, 119: 409–422.

David, M. (2005). Some T-biconditionals. In AmourGarb, B. and Beall, J., editors, *Deflationary Truth*, pages 382–419. Chicago, Open Court.

Davidson, D. (1984). Truth and meaning. In *Inquiries Into Truth and Interpretation*, pages 17–36. Clarendon Press.

Davidson, D. (1990). The structure and content of truth. *Journal of Philosophy*, 87: 279–328.

Dever, J. (2006). Compositionality. In Lepore, E. and Smith, B. C., editors, *The Oxford Handbook of Philosophy of Language*. Oxford University Press.

Dummett, M. (1991). *The Logical Basis of Metaphysics*. Harvard University Press.

Feferman, S. (1962). Transfinite recursive progressions of axiomatic theories. *Journal of Symbolic Logic*, 27: 259–312.

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

Feferman, S. (1996). Gödel's program for new axioms: Why, where, how and what? In P. Hájek, editor, *Gödel' 96*, volume 6 of Lecture Notes in Logic. Springer.

Field, H. (1994). Deflationist views of meaning and content. *Mind*, 103: 249–285. Reprinted in Field, 2001, Ch. 4.

Field, H. (1999). Deflating the conservativeness requirement. *Journal of Philosophy*, 96: 533–540.

Field, H. (2001). *Truth and the Absence of Fact*. Clarendon Press.

Field, H. (2005). Compositional priniciples vs. schematic reasoning. *The Monist*, 89: 9–27.

Frege, G. (1967). Begriffsschrift: A formula language modeled upon that of arithmetic, for pure thought. Translated by S. Bauer-Mengelberg, in J. van

Heijenoort, editor, *From Frege to Gödel: A Sourcebook in Mathematical Logic 1879–1931*, pages 5–82. Harvard University Press.

Frege, G. (1984a). *Collected Papers on Mathematics, Logic, and Philosophy*, McGuiness, B., editor. Basil Blackwell.

Frege, G. (1984b). Function and concept. Translated by P. Geach, in Frege 1984a, pages 137–156. Also in Frege 1997, pages 130–148.

Frege, G. (1984c). On sense and meaning. Translated by M. Black, in Frege 1984a, pages 157–177. Also in Frege 1997, pages 151–171.

Frege, G. (1997). *The Frege Reader*, Beaney, M., editor. Oxford, Blackwell.

Frege, G. (2013). *The Basic Laws of Arithmetic*. Translated by P. A. Ebert and M. Rossberg. Oxford University Press.

Friedman, H. and Sheard, M. (1987). An axiomatic approach to self-referential truth. *Annals of Pure and Applied Logic*, 33: 1–21.

Friedman, H. and Sheard, M. (1988). The disjunction and existence properties for axiomatic systems of truth. *Annals of Pure and Applied Logic*, 40: 1–10.

Gamester, W. (2018). Truth: Explanation, success, and coincidence. *Philosophical Studies*, 175: 1243–1265.

Gupta, A. (1993). A critique of deflationism. *Philosophical Topics*, 21: 57–81.

Halbach, V. (1999). Disquotationalism and infinite conjunction. *Mind*, 108: 1–22.

Halbach, V. (2001). How innocent is deflationism? *Synthese*, 126: 167–194.

Heck, R. K. (2004). Truth and disquotation. *Synthese*, 142: 317–352. Originally published under the name "Richard G. Heck, Jr".

Heck, R. K. (2007). Meaning and truth-conditions. In Greimann, D. and Siegwart, G., editors, *Truth and Speech Acts: Studies in the Philosophy of Language*, pages 349–376. Routledge. Originally published under the name "Richard G. Heck, Jr".

Heck, R. K. (2010). Frege and semantics. In Potter, M. and Ricketts, T., editors, *The Cambridge Companion to Frege*, pages 342–378. Cambridge University Press. Originally published under the name "Richard G. Heck, Jr".

Heck, R. K. (2011). A logic for Frege's Theorem. In *Frege's Theorem*, pages 267–296. Oxford, Clarendon Press. Originally published under the name "Richard G. Heck, Jr".

Heck, R. K. (2012). *Reading Frege's Grundgesetze*. Clarendon Press. Originally published under the name "Richard G. Heck, Jr".

Heck, R. K. (2013). Is compositionality a trivial principle? *Frontiers of Philosophy in China*, 8: 140–155. Originally published under the name "Richard G. Heck, Jr".

Heck, R. K. (2015). Consistency and the theory of truth. *Review of Symbolic Logic*, 8: 424–466. Originally published under the name "Richard G. Heck, Jr".

Heck, R. K. (2018a). The logical strength of compositional principles. *Notre Dame Journal of Formal Logic*, 59: 1–33. Originally published under the name "Richard G. Heck, Jr".

Heck, R. K. (2018b). Logicism, ontology, and the epistemology of second-order logic. In Fred, I. and Leech, J., editors, *Being Necessary: Themes of Ontology and Modality from the Work of Bob Hale*, pages 140–169. Oxford University Press.

Heck, R. K. and May, R. (2018). Truth in Frege. In M. Glanzberg, editor, *The Oxford Handbook of Truth*, pages 193–215. Oxford University Press.

Heck, R. K. and May, R. (2020). The birth of semantics. *Journal for the History of Analytic Philosophy*, 8, no 6: 1–31.

Hill, C. (2002). *Thought and World: An Austere Portrayal of Truth, Reference, and Semantic Correspondence*. Cambridge University Press.

Horsten, L. and Leigh, G. E. (2017). Truth is simple. *Mind*, 126: 195–232.

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

Horwich, P. (1998). *Meaning*. Clarendon Press.

Jacobsen, P. (2014). *Compositonal Semantics: An Introduction to the Syntax/Semantics Interface*. Oxford University Press.

King, J. C. and Stanley, J. (2004). Semantics, pragmatics, and the role of semantic content. In Z. G. Szabó, editor, *Semantics versus Pragmatics*, pages 111–164. Oxford University Press.

Larson, R. and Segal, G. (1995). *Knowledge of Meaning*. MIT Press.

Linnebo, Ø. (2003). Frege's conception of logic: From Kant to Grundgesetze. *Manuscrito*, 16: 235–252.

MacColl, H. (1877). The calculus of equivalent statements and integration limits. *Proceedings of the London Mathematical Society*, IX: 9–20.

Marcus, R. B. (1972). Quantification and ontology. *Noûs*, 6: 240–250.

McGee, V. (1997). How we learn mathematical language. *Philosophical Review*, 106: 35–68.

McGee, V. (2005). Afterword: Trying (with limited success) to demarcate the disquotational–correspondence distinction. In AmourGarb, B. and Beall, J., editors, *Deflationary Truth*, pages 143–152. Open Court.

Moore, G. E. (1953). *Some Main Problems of Philosophy*. Allen and Unwin.

Partee, B. (1984). Nominal and temporal anaphora. *Linguistics and Philosophy*, 7: 243–286.

Quine, W. V. O. (1956). Quantifiers and propositional attitudes. *The Journal of Philosophy*, 53: 177–187. Reprinted in Quine, 1976, Ch. 17.

Quine, W. V. O. (1970). *Philosophy of Logic*. Prentice Hall.

Quine, W. V. O. (1976). *The Ways of Paradox and Other Essays*. Harvard University Press.

Quine, W. V. O. (1986). *Philosophy of Logic*. Harvard University Press, 2nd edition.

Quine, W. V. O. (1987). *Quiddities: An Intermittently Philosophical Dictionary*. Belknap Press.

Quine, W. V. O. (1990). *Pursuit of Truth*. Harvard University Press.

Schnieder, B. (2011). A logic for "because". *Review of Symbolic Logic*, 4: 445–465.

Schröder, E. (1972). Review of Frege's Conceptual Notation. Translated by T. W. Bynum, in T. W. Bynum, editor, *Conceptual Notation and Related Articles*, pages 218–232. Oxford University Press.

Shapiro, S. (1998). Proof and truth: Through thick and thin. *Journal of Philosophy*, 95: 493–521.

Simchen, O. (2012). *Necessary Intentionality: A Study in the Metaphysics of Aboutness*. Oxford University Press.

Simpson, S. (2009). *Subsystems of Second Order Arithmetic*. Cambridge University Press, 2nd edition.

Soames, S. (1988). Semantics and semantic competence. In Schiffer, S. and Steele, S., editors, *Cognition and Representation*, pages 185–207. Westview Press.

Soames, S. (1999). *Understanding Truth*. Clarendon Press.

Strawson, P. F. (1950). Truth. *Proceedings of the Aristotelian Society*, 24: 129–156. Reprinted in Strawson, 1971, Ch. 10.

Strawson, P. F. (1952). *Introduction to Logical Theory*. Methuen.

Strawson, P. F. (1971). *Logico-Linguistic Papers*. Methuen.

Szabó, Z. G. (2012). The case for compositionality. In Hinzen, W., Machery, E. and Werning, M., editors, *The Oxford Handbook of Compositionality*, pages 64–80. Oxford University Press.

Tarski, A. (1944). The semantic conception of truth and the foundations of semantics. *Philosophy and Phenomenological Research*, 4: 341–375.

Textor, M. (2010). Frege on judging as acknowledging the truth. *Mind*, 119: 615–655.

Williams, M. (1999). Meaning and deflationary truth. *Journal of Philosophy*, 96: 545–564.

7 Belief, Truth, and Ways of Believing

Johannes Stern

7.1 Introduction

Philosophy without truth, knowledge, and belief would be a fairly boring discipline—there would only be the good and the beautiful left to discuss. Fortunately, philosophy is exciting and truth, knowledge, and belief are notions at the center of the discipline responsible for many important philosophical questions and puzzles. The three notions are intimately connected and, as a consequence, so will be the philosophical questions and puzzles of the respective notions. For example, knowledge guarantees truth, i.e., it is factive and, indeed, it is arguably at least in parts this characteristic that distinguishes knowledge from mere belief. Whether this means that knowledge can be defined on the basis of knowledge, truth and, possibly, some further condition has been the question shaping much of the recent debate in epistemology. With this observation in mind one would expect that most formal philosophizing is conducted in a formal framework in which truth, knowledge, and belief are treated simultaneously, so the formal and philosophical views regarding the connection of the different notions can be tested for their consequences. Surprisingly, no satisfactory such framework has—to our knowledge—been developed to date. Of course, starting with Hintikka (1962) there has been a lot of work on formal semantics and logics of knowledge and belief but unfortunately very little work on how to construct an adequate theory of truth in these contexts.[1] The aim of this chapter is to take first steps in developing a satisfactory formal framework in which contemporary debates in epistemology can be aptly represented. To this end, we focus on the notion of truth in belief contexts—although a number of observations would also apply to the interaction of the notions of truth and knowledge—and start by examining a major hurdle or puzzle in way of a satisfactory semantics for truth in doxastic contexts. We then analyze the philosophical underpinnings of the puzzle and develop a semantics for the notion of truth in doxastic contexts, which is based on our analysis. We discuss some of its consequences and, before concluding, point to some limitations of the semantics and outline some alternative strategies for developing adequate semantics for truth in doxastic contexts.

7.2 Semantics for Truth and Belief: Overgeneration

As mentioned, Hintikka's seminal *Knowledge and Belief* (Hintikka, 1962) lays the foundation for formal philosophizing about knowledge and belief. Hintikka proposed a formal interpretation of belief and, respectively, knowledge within the framework of possible world semantics (henceforth PWS). According to Hintikka an agent believes (knows) that φ if and only if 'φ' is true at all of the agent's doxastic (epistemic) alternatives—worlds that are accessible via the doxastic (epistemic) accessibility relation—where the truth predicate is understood in a metalinguistic, that is, model-theoretic sense. Most subsequent work on formal semantics for belief has followed Hintikka's footsteps in analyzing belief in some form of possible world semantics broadly conceived, that is, as some form of quantifier over worlds, states, or situations.[2] It seems only reasonable then to take the possible world analysis as a starting point for a combined formal framework for truth and belief. What we are after is a framework in which the notions of truth and belief figure in the object-language, that is, we want to formulate claims such as

(1) Not everything Boris believes is true.

As a consequence, standard PWS for epistemic notions will not be sufficient because, as mentioned, the truth predicate at play in the semantics is the metalinguistic one.[3]

If an object-linguistic truth predicate is introduced to the framework of PWS, its semantic interpretation needs to be specified, that is, the interpretation of the truth predicate at every possible world has to be determined. To this end, it is not sufficient to determine the interpretation of the object-linguistic truth predicate at a given world by fiat. Rather, if, following the outlines of a commonly accepted view on truth and paradox, semantic states of affairs supervene on non-semantic states of affairs (cf., e.g., Tarski, 1944; Kripke, 1975; Yablo, 1982; Leitgeb, 2005), the interpretation of the truth predicate at a given world should arguably depend on the interpretation of the non-semantic expressions at that world: a sentence φ will be in the interpretation of the truth predicate at a world only if the possible world models satisfies φ at a world w. If this idea is taken seriously, then an interpretation of the truth predicate f is adequate relative to a possible world model M and world w only if, where t_φ is a name of φ,

$$M, w \models^{f} \mathrm{T}t_\varphi \Leftrightarrow M, w \models^{f} \varphi.^{4} \tag{TSW}$$

Fortunately, finding adequate interpretations of the truth predicate in PWS does not pose a major technical obstacle: one can simply relativize one's favorite theory of truth to the possible world framework and simultaneously construct the interpretations of the truth predicate relative to every possible world of the modal frame (cf., e.g., Kripke,1975; Asher and Kamp, 1989; Gupta and Belnap, 1993; Halbach and Welch, 2009; Stern, 2014a,b, 2016).

The foregoing suggests that a semantics for truth in doxastic contexts can be obtained by supplementing standard doxastic PWS by an interpretation of the object-linguistic truth predicate relative to each possible world following well-rehearsed strategies discussed in the relevant literature. Unfortunately, it turns out that things are not quite as simple as that: while combining possible world semantics with standard truth-theoretic tools yields a powerful semantics for truth in belief contexts, the semantics turns out to be too powerful and to validate principles and inferences that ought not to be taken for granted. In particular, (TSW) implies that in every belief model *M* and world *w* whatever an agent believes at *w*, they also believe to be true and vice versa. Let's call this the Original Sin (OS) of PWS:

$$M, w \models^{f} \mathrm{B}\varphi \Leftrightarrow M,\ w \models^{f} \mathrm{BT}t_{\varphi}. \tag{OS}$$

(OS) will hold independently of whether we consider worlds, states, or situations, as long as φ and $\mathrm{T}t_{\varphi}$ receive the same semantic value at these points of evaluation, that is, if (TSW) holds at every point of evaluation and the believe operator B is conceived of as a quantifier ranging over points of evaluation.[5] Notice that abandoning (TSW) ought not to be taken lightly, since, at least prima facie, this would undermine the idea that semantic states of affairs ought to supervene on non-semantic states of affairs. In sum, (OS) is a consequence of the two fundamental assumptions underlying PWS for the belief operator and the semantic interpretation of the truth predicate, respectively.

Let us now reflect on why we should be reluctant in accepting (OS), that is, why believing and believing-true ought to be semantically differentiated. To this end, we shall present a number of cases, which, at least at the outset present counterexamples to (OS). One such case is based on the idea that the truth predicate may not be part of an agent's conceptual resources. Meet Xaver:

> **Xaver** believes Bavaria is beautiful. But because his conceptual resources lack the truth predicate Xaver simply cannot form the belief that 'Bavaria is beautiful' is true.

It seems hard to deny that it is impossible for Xaver to form an attitude towards that 'Bavaria is beautiful' is true, however, it is another question altogether whether Xaver's particular disposition amounts to a compelling counterexample to (OS). First, we may simply stipulate PWS for truth in doxastic contexts to be concerned with a theory of the doxastic attitudes of agents that have the necessary conceptual resources, i.e., conceptual resources that comprise the truth predicate. Perhaps, one might think that this condition is overly demanding or restrictive, i.e., even rational agents should not be expected to have a truth predicate at their disposition. But the aim of the semantics is not to give a general theory of attitude reports.[6] Rather the aim is to provide a semantics for truth in doxastic contexts and from this perspective it seems perfectly acceptable to focus on a semantics for agents with the necessary conceptual resources. After all, a similar kind of argument could be applied against the plausibility of any general inference involving higher-order beliefs, e.g., introspection principles—an agent may simply lack the conceptual resources to form an higher-order belief: we would be forced to conclude that the wealth of research on the plausibility of such principles is an idle exercise.

Second, even if Xaver's conceptual resources were not to include a truth predicate, this does not imply that we cannot introduce the truth predicate to the language we employ for theorizing about, or reporting, the agent's attitude. For example, meet Anne:

> **Anne** believes Euclidean geometry to be incorrect and by modus tollens infers that one of the axioms must be incorrect without settling on one specific axiom (she may not even know all the axioms).

In this case it seems—or at least a disquotationalist would argue—that Anne's belief is correctly reported by

(2) Anne believes that not all axioms of Euclidean geometry are true;

independently of whether Anne's conceptual resources comprise the truth predicate. More importantly, at first glance it seems as if we require the truth predicate in our language to describe Anne's belief correctly. Admittedly, the view comes with important theoretical costs, namely, that the truth predicate is transparent even in highly opaque contexts but the point still stands: the absence of the truth predicate from an agent's conceptual resources is not sufficient to argue against (OS).

In sum, we take it that the charge against (OS) based on the idea that an agent's conceptual resources may lack the truth predicate to be unconvincing and will dismiss it for the purpose of our chapter. But there is more damning evidence against (OS). In particular, there are more convincing cases to the effect that an agent can believe something without believing it true. Meet Clara:

> **Clara** believes that Clark Kent is strong. But she would never express her belief in this way because she only believes that 'Superman is strong' is true. She does not believe that 'Clark Kent is strong' is true.

Clara's beliefs are in plain contradiction with (OS): she believes something without believing it true. Moreover, an argument to the effect that despite appearances Clara does believe that 'Clark Kent is strong' is true and that our intuitions contradicting this assessment are down to pragmatic effects rather than a semantic distinction would seem hardly convincing in this case: at least prima facie by reporting that Clara believes that 'Clark Kent is strong' is true, we assert that Clara believes something true relative to a particular syntactic representation. The syntactic representation at stake is made explicit in the belief report and should therefore be part of the semantic content of the belief report.

Admittedly, in reporting Clara's belief we have assumed that the belief relation is merely a relation between an agent and semantic content where names are conceived as rigid designators, i.e., the syntactic or cognitive representations of the belief are not relevant for the semantic evaluation of the belief report. On alternative accounts of attitude reports, it would be incorrect to say that Clara believes that Clark Kent is strong. We take it that by constructing a semantics for truth in belief contexts, one should ideally remain neutral with respect to the particular theory of attitude reports assumed and, hence, not dismiss counterexamples to (OS) because they depend on a particular—rather popular—account of belief reports. Moreover, in general arguments against (OS) do not rely on a particular theory of attitude reports. Meet Max:

> **Max** believes that Goldbach's conjecture is true. His friend Philip told him so and Philip is a mathematical genius. Max has absolute faith in Philip and believes him even though he has no idea what Goldbach's conjecture asserts. In fact, he does not believe that every even number > 2 is the sum of two prime numbers.

Max believes Goldbach's conjecture true without believing it. It seems undeniable that Max has not formed an attitude towards Goldbach's conjecture; he does not believe it. It also seems clear that Max believes Goldbach's conjecture is true. Perhaps one might be tempted to argue that one can only believe that Goldbach's conjecture is true if one is aware of what Goldbach's conjecture asserts. But this imposes too strict and indeed incorrect conditions on believing. We frequently believe claims, theories, etc. true without being fully aware what they assert. Moreover, we often form such beliefs simply due to (hopefully) expert testimony. In sum, we think that Max's beliefs are a clear counterexample against (OS) and that, more generally, the evidence against the semantic equivalence of believing and believing-true is damning: believing and believing-true need to be semantically differentiated.

Having corroborated the claim that the combination of possible world semantics for belief and basic desiderata regarding the interpretation of the truth predicate, when combined, yield unintended results for truth in belief contexts, the question arises whether the unintended results of the semantics, i.e. (OS), are merely a case of a formal semantics having unintended consequences or whether these results point to a deeper, philosophical problem pertaining to the notion of truth in belief contexts. In the latter case, we may yield invaluable insights for developing an adequate semantics by addressing the philosophical problem. Indeed, it turns our that the purported semantic equivalence of believing and believing-true is rooted in a philosophical puzzle about belief: if a disquotational view of truth à la Field (1994) is assumed, then the semantic equivalence of believing and believing-true is but another Fregean puzzle about belief.

7.3 Believing, Believing-True, and a Puzzle About Belief

Traditionally, Fregean puzzles about belief employed the idea that if two names refer to the same object, they should be intersubstitutable salva veritate. But it is well known that the substitution of coreferential terms in belief contexts leads to counterintuitive consequences—indeed it were these counterintuitive consequences that led Frege (1892) to conclude that the referent of a name in oblique contexts such as belief, was not the actual referent but the sense associated with the name—and one might therefore be wary of appealing to the substitution of coreferential terms when reasoning about belief contexts. Kripke (1979) argued that the appeal to the intersubstitutivity of coreferential terms was inessential in formulating Frege-style belief puzzles. Rather Kripke based the formulation of such puzzles on two so-called disquotational principles:[7]

(DQ) If an agent A sincerely, reflectively, and competently accepts a sentence s (under circumstances properly related to a context c), then A believes, at the time of c, what s expresses in c.

(CDQ) If an agent A sincerely, reflectively, and competently denies or withholds acceptance from a sentence s (in a context c), then A does not believe, at the time of c, what s expresses in c.

At least, if agents are competent speakers of the language at stake, (DQ) and (CDQ) are prima facie plausible assumptions linking the acceptance of a sentence by an agent to the agent's belief in the semantic content expressed by the sentence. But if (DQ) and (CDQ) are granted, this raises two puzzles about Clara's beliefs: since, on the one hand, Clara will accept the sentence 'Superman is strong', we can infer by (DQ) that

(3) Clara believes that Clark Kent is strong.

On the other hand, since Clara will withhold acceptance to 'Clark Kent is strong', we can infer

(4) Clara does not believe that Clark Kent is strong.

by (CDQ). Moreover, (DQ) does not only imply (3) but also

(5) Clara believes that Clark Kent is not strong.

since Clara would arguably accept the sentence 'Clark Kent is not strong'. We are left with a dilemma, that is, a Fregean puzzle about belief: not only does Clara hold, in virtue of (3) and (5), mutually incompatible beliefs, but we also face the question whether Clara believes that Clark Kent is strong, as suggested by (3), or not, as claimed by (4).

However, the disquotational principles (DQ) and (CDQ) do not only generate Frege-style puzzles about belief, they also immediately imply that believing and believing-true are semantically equivalent, if a disquotational view of the truth predicate along the lines of Field (1994) is assumed. On such a disquotational perspective the sentence/utterance φ and the sentence/utterance Tt_{φ} are not only thought to be semantically equivalent but cognitively equivalent.[8] But if the sentences φ and Tt_{φ} are cognitively equivalent, it seems that if a rational agent accepts the sentence φ they will also accept the sentence Tt_{φ}, and vice versa, that is, from the disquotational perspective we seem justified to assume the following principle:

> **(TDQ)** An agent A sincerely, reflectively, and competently accepts a sentence s (under circumstances properly related to a context c), if and only if, A sincerely, reflectively, and competently accepts the sentence $T\bar{s}$ (under circumstances properly related to a context c).[9]

Together (DQ), (CDQ), and (TDQ) imply (OS), i.e., the claim that believing and believing-true are semantically equivalent. This suggests that if Field's disquotational perspective on truth is assumed, then the only way to resist (OS) is to reject either (DQ) or (CDQ).

7.3.1 Rejecting Disquotational Belief Principles?

In way of answering traditional puzzles about belief (CDQ) is frequently rejected. Unfortunately, whilst this may help with answering these puzzles it does not really get us out of the fire in the present case. Although, strictly speaking, we can no longer derive (OS) without assuming (CDQ), (OS) will still be a consequence of (DQ) and (TDQ) for instances φ whenever we accept a sentence *s* expressing φ or accept that s is true. Now, Max clearly accepts 'Goldbach's conjecture is true', and hence by appeal to (DQ) and (TDQ) we obtain that Max believes that Goldbach's conjecture is true if and only if Max believes Goldbach's conjecture, which, we have argued, is intuitively wrong.[10] The moral to draw, it seems, is that if the disquotational perspective is accepted in an unqualified way, then one ought to reject both (DQ) *and* (CDQ), if one wants to resist (OS). However, rejecting (DQ) would be at odds with standard semantics of attitude reports, as the principle is widely accepted in the literature on belief reports. Accordingly, we will refrain from explicitly ruling out (DQ) as a plausible principle. For one, in developing our semantics we wish to remain as neutral as possible with respect to the various theories of attitude reports discussed in the literature: it is not the job of a general semantics to be the arbiter between different philosophical or semantic theories. Rather, such a semantics should make the semantic consequences of the different theories precise. For another, there is a more general reason why one should be wary of rejecting (DQ) in reaction to the derivation of (OS): the principal example we used to argue against (OS) seem to also yield a straightforward argument against (TDQ). Recall the case of Max; for all we know Max would accept 'Goldbach's conjecture is true' but reject 'Every even number > 2 is the sum of two prime numbers', that is, Max's acceptance patterns would not be conform with (TDQ). This suggests that the right course of action is to rethink (TDQ) rather than to reject the disquotational belief principles.

7.3.2 (TDQ) and the Disquotational Case for (OS)

The disquotationalist seems to be left with two options. They can either reject (TDQ) or accept (OS) as a valid principle governing the interaction of truth and belief. Indeed we think that the disquotationalist who holds (TDQ) dear should accept (OS). Of course, this is at odds with Clara's and Max's beliefs, but disquotationalists frequently suggest that it is not their aim to capture all reasonable uses of the truth predicate in natural language but only the theoretically useful ones, that is, the disquotational uses of truth.[11] They wish to characterize a truth predicate that can fulfill its theoretical role, i.e. its disquotational function, and to this end it seems required that φ and $\mathrm{T}t_{\varphi}$ are intersubstitutable salva veritate in contexts like (2) for otherwise, it seems, that (2) would not correctly report Anne's belief.[12] On this view, Max's use of the truth predicate would simply not qualify as a use of the disquotational notion of truth since, according to Field (1994), "*a person can meaningfully apply* *'true' in the pure disquotational sense only to utterances that he has some understanding of*" (p. 250). A disquotationalist will hence simply dismiss cases like Max's as irrelevant for their project.[13] They are not legitimate counterexamples to (OS) save (TDQ). On this view, disquotationalist should not flinch and accept (OS) ... alas few do.[14]

However, the disquotationalist's dismissal of the counterexamples against (TDQ) and (OS) points to a different possible course of action. In contrast to the diquotationalist position sketched above, one may be more liberal and allow for non-disquotational uses of the truth predicate, that is, uses of the truth predicate for which (TDQ) and (OS) may fail. Arguably, this failure need not concern the disquotationalist, since it is limited to non-disquotational uses of 'true'. The idea would conceive of (TDQ) and (OS) as principles pertaining only to "ideal" or "disquotational" circumstances, that is, circumstances in which—according to the disquotationalist—an agent "has some understanding of" the utterance they deem true. More precisely, if an agent is aware, or understands, which belief is represented, directly or indirectly, by a term t_φ, then a rational agent holds that particular belief, if and only if, they hold t_φ true, that is, in this case they will believe $\mathrm{T}t_\varphi$ if and only if they believe φ. Another way of putting this idea is that (OS) should hold for an instance t_φ, if and only if, the agent thinks about φ and $\mathrm{T}t_\varphi$ in the same way, that is, in this particular case the agent treats the truth predicate transparently. A semantics based on this idea will attribute independent truth-conditions to $\mathrm{B}\varphi$ and $\mathrm{BT}t_\varphi$ that only coincide in case of ideal, disquotational circumstances. Such a semantics should be acceptable to both disquotationalists, and truth theorists that are neither disquotationalists nor deflationists:[15] after all, $\mathrm{B}\varphi$ and $\mathrm{BT}t_\varphi$ are treated as semantically equivalent if the truth predicate is used disquotationally, but the semantics also allows for non-disquotational uses of the truth predicate in which the semantic equivalence breaks down.

7.3.3 Ways of Believing

In the literature on attitude reports appealing to the way an agent thinks of a given belief, i.e. the way they believe, is a common strategy for explaining allegedly counterintuitive consequences of, roughly, Russellian theories of attitudes. For example, it has been argued that it is possible for a rational agent to believe that φ, while at the same time to believe that $\neg\varphi$ as long as they do not believe φ and $\neg\varphi$ in the same way. There is some disagreement whether the way of believing should be semantically or pragmatically encoded: Naive Russellians such as Soames (1987) typically argue that it should be merely pragmatically encoded, whilst contextualists like Crimmins and Perry (1989) embrace the idea that the way of believing ought to be semantically encoded. For example, according to Crimmins and Perry (1989) depending on the way a proposition is believed an agent stands in a different belief-relation to the proposition at stake.[16] Moreover, an agent can believe a proposition φ in one way but believe its negation in another way and, in this case, both Bφ and B¬φ should receive the semantic value true. According to the standard contextualist view the way of believing depends upon an unarticulated constituent of the attitude report and is provided by the context under consideration. In contrast, the semantic picture we are about to propose agrees with the contextualist that the way of believing impacts the semantic evaluation of the attitude report but we do not conceive of it as an unarticulated constituent of the attitude report.[17] Rather, the idea is that the way of believing BTt_φ is determined by the specific representation t_φ. Furthermore, in the absence of further information we are only guaranteed to believe Tt_φ—assuming we believe it at all—in this specific way, i.e. under the representation t_φ, and this idea will be hardwired into our semantics. In contrast, if no formula of the form Tt_ψ occurs in φ, then φ will be believed in an unspecific way, that is, a way of believing that does not depend on a particular representation of the belief. If we believe a proposition in such an unspecific way there is again no guarantee that we also believe it under a specific representation: Bφ and BTt_φ can only be assumed to be equivalent if we are guaranteed that φ and Tt_φ are believed in the same way. On our semantics this will be the case, if an agent is aware that by believing that t_φ is true, they commit themselves to believing that φ and vice versa, that is, if an agent is aware that t_φ, directly or indirectly, represents the belief that φ.

7.4 Semantics for Ways of Believing

In the previous section we argued that the way of believing impacts the semantic value of a belief ascription and that the way of believing depends on the representation t_φ, if Tt_φ occurs in the belief context. But this leaves open two alternatives on how t_φ can impact the way of believing: it can either have an impact qua expression of the language or in virtue of what it denotes. Which of the two alternatives one ought to pick will depend on the objects one takes the truth predicate to apply to. Again there are two options: if, as in the case of disquotational truth, a sentential truth predicate is assumed, the objects of truth are sentences (or utterances) and, as a consequence, the objects of truth will arguably be of a different category than the objects of belief, which typically are thought to be propositions.[18] However, if a propositional truth predicate is assumed, the objects of truth will be propositions, and the objects of truth and belief will coincide. Of course, depending on the view one opts for, a name t_φ will denote different types of objects, that is, either sentences or propositions. In this chapter we makes the simplifying assumption that whether the denotatum of a name t_φ is a sentence or a proposition is not reflected on the linguistic level, that is, we cannot distinguish between names of propositions and names of sentences on purely syntactic or linguistic grounds. Rather we consider it to be a conceptual decision which type of denotata one opts for. More generally, we conceive of a "name" t_φ to be any kind of nominalization that plausibly denotes a sentence or a proposition, that is, t_φ need not be a proper name but could, e.g., also be a definite description or a that-clause. Similarly, from the perspective of our formal language t_φ will simply be a singular term that denotes the sentence 'φ' or the proposition that φ.

Let us return to the question of whether t_φ impacts the way of believing qua name or in virtue of what it denotes and consider the case of the sentential truth predicate. In this case a term t_φ names a sentence, which, in turn, expresses a proposition. The term t_φ does not represent a belief directly but only via the sentence it denotes. We arrive at a framework of two-level belief representation as displayed in Figure 7.1. On this view, it seems plausible to assume that the reason why an agent may believe, say, that snow is white whilst at the same time not believe that 'Snow is white' is true, is that the agent is not aware that the sentence 'Snow is white' expresses the proposition that snow is white. In other words, from perspective of sentential truth the way we believe is determined by the sentence rather than its name.

We shall adopt the sentential, i.e. disquotational, perspective in formulating our semantics but the view that conceives of the objects of truth as sentences (or utterances) is highly contested. Rather it is often thought that the natural language truth predicate applies to propositions. On this view, t_φ denotes a proposition and believing that $\mathrm{T}t_\varphi$ is not mediated via a sentence that expresses a proposition but depends only on the name t_φ and the proposition itself. In this case we are working in a framework of one-level belief representation as displayed in Figure 7.2. Accordingly, if the objects of truth are conceived of as propositions and, at the same time, we wish to maintain the idea that t_φ impacts the way we believe $\mathrm{BT}t_\varphi$, we are forced to accept that t_φ impacts the way of believing qua expression of the language, i.e. qua name, rather than via its semantic contribution. Of course, this idea could also be implemented within the

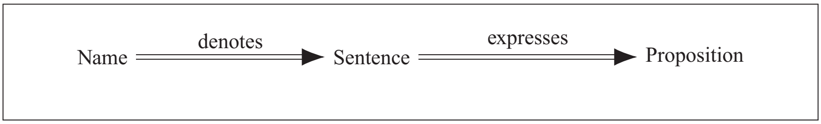


Figure 7.1 Two-level belief representation

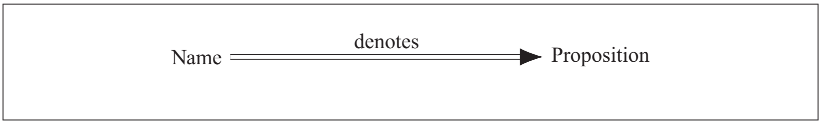


Figure 7.2 One-level belief representation

framework of two-level belief representation, but making the way of believing dependent on sentences rather than their names has the neat consequence that the semantics remains fully referential. In contrast, if the name t_φ is allowed to impact the semantics evaluation qua expression, the semantics will no longer be fully referential. So there seems to be at least a prima facie advantage of adopting a framework of two-level belief representation in which the way of believing depends on the denotatum of t_φ rather than t_φ itself.

Independently of whether we opt for a sentential or a propositional truth predicate the question arises how the contribution of the way of believing should be spelled out within the framework of PWS. To this end, it is helpful to coerce the believe relation into the framework of PWS: in PWS an agent *a* believes the proposition that φ, denoted by $\|\varphi\|$, at a world w iff

$$\forall v(wR_a v \Rightarrow v \in \|\varphi\|),$$

where $\|\varphi\|$—the proposition that φ—is the set of possible worlds in which φ is true and R_a a doxastic accessibility relation. From this definition of the belief-relation one can easily see that the only parameter in the definition is the doxastic accessibility relation. Consequently, if the way of believing is supposed to depend upon t_φ, the term needs to impact the accessibility relation. Indeed, the crucial point of departure of our semantics from standard PWS is that instead of assuming a primitive accessibility relation for every agent we now consider a function that outputs accessibility relations and takes either finite sets of sentences (sentential truth predicate) or finite sets of names (propositional truth predicate) of the language as inputs. We appeal to sets of sentences (terms) rather than to the sentences or names themselves, since we might have several formulas of the form $\mathrm{T}t$ in the belief-context. For example, the truth of $B(\mathrm{T}t_\varphi \wedge \mathrm{T}s_\psi)$, that is, the way we believe $\mathrm{T}t_\varphi \wedge \mathrm{T}s_\psi$, should depend on both t_φ and s_ψ, that is, on the set $\{t_\varphi, s_\psi\}$.

Before we spell out the formal semantics in some more detail, we need to reconsider the idea that (TDQ) and (OS) should hold in "ideal" circumstances. Following up on our remarks in Section 7.3 we take "ideal" or "disquotational" to be circumstances in which the agent is *aware* which proposition the sentence denoted by t_φ expresses (the name t_φ denotes). In this case their belief that $\mathrm{T}t_\varphi$ will be insensitive to the particular way of believing related to t_φ and coincide with the way the agent believes that φ, that is, relative to t_φ the truth predicate behaves transparently in the relevant belief-context. On our semantics, the information which sentences (names) an agent is aware of needs to be provided externally, i.e., it is not possible to compute the Awareness set on the basis of the information provided within the semantics. In this respect our semantics resembles classical Awareness semantics (cf. Fagin et al., 1995, Chapter 9.5), where an externally provided Awareness set controls the transition from idealized belief to non-idealized belief. It is perhaps best to view the Awareness set to be retrieved from the information provided by the common ground relevant to the particular belief report but we shall leave this issue open for the purpose of our formal semantics.

7.4.1 Formal Semantics

In this section we make our heuristic remarks precise and introduce a formal semantics for ways of believing for a language containing the belief operator B and the truth predicate T.[19] As we remarked in the previous section we shall assume a sentential truth predicate, that is, we shall develop a semantics for two-level belief representation. We wish to keep our semantics as general as possible. For this reason we appeal to an inner/outer domain semantics (cf., e.g., Garson, 2001), that is, we allow for a universe of discourse U, which comprises the domain of quantification, which is allowed to vary from world to world. We also allow for terms of the language to denote non-rigidly, as long as these terms are not expressions of the language of the syntax theory. The interpretation of the syntax theory, that is the syntactic vocabulary, will remain constant across worlds. In contrast to more customary formulations the syntax theory will not carry any explicit ontological commitments, as we shall not require the expressions of the language to be in the domain of quantification at each world. Rather, any potential commitment should be considered implicit and as a sine qua non condition of the theoretical framework; it is a different kind of commitment than the one we engage in when talking about, say, elephants. The question how this type of commitment is to be understood is left open but will obviously depend on one's interpretation of the universe U. To avoid explicit commitment to an ontology of expressions the language of syntax is conceived of as a quantifier-free language along the lines of certain formulations of PRA. However, our syntax theory and language need not be an arithmetical language where the syntactic operations operate on codes of expressions, i.e., natural numbers, but could—perhaps preferably—be a syntax theory that operates directly over an ontology of expressions (see, e.g., Halbach and Leigh, 2021).

Definition 7.4.1 (Universe, Language). Let $\mathrm{O} \neq \emptyset$ be the set of non-syntactic objects relevant to the discourse under consideration. Let $\mathcal{L}_{\mathrm{O}}$ be the language of a syntax theory such that $\mathcal{L}_{\mathrm{O}}$

- contains names $o_1, o_2, \ldots$ for all member of O;
- contains the logical constants $\neg$ and $\wedge$ and the identity predicate but is quantifier free;
- contains names for all expressions of $\mathcal{L}_{\mathrm{B}}$ (cf. below) and function symbols for all primitive recursive syntactic operations of $\mathcal{L}_{\mathrm{B}}$.

Let $\mathrm{U} := \mathrm{O} \cup \mathsf{Expr}_{\mathcal{L}_{\mathrm{B}}}$ be the universe of discourse for $\mathcal{L}_{\mathrm{B}}$ where $\mathsf{Expr}_{\mathcal{L}_{\mathrm{B}}}$ is the set of all expressions of $\mathcal{L}_{\mathrm{B}}$. $\mathcal{L}_{\mathrm{B}}$ extends $\mathcal{L}_{\mathrm{O}}$ by a countable number of individual constants c_1, c_2, … ; n-ary predicate symbols $P_1^n, P_2^n, \ldots$; the belief operator B; the truth predicate T; (possibly) the Awareness predicate A and the universal quantifier $\forall$. Other logical symbols are used merely as abbreviations. The syntax of $\mathcal{L}_{\mathrm{B}}$ is given by

$$\varphi ::= \psi \mid t_i = t_j \mid P^n t_1, \ldots, t_n \mid \mathrm{A}t \mid \mathrm{T}t \mid \neg\varphi \mid \varphi \wedge \varphi \mid \mathrm{B}\varphi \mid \forall x\varphi$$

with $\psi \in \mathsf{Frml}_{\mathcal{L}_{\mathrm{O}}}$ and $t_1, \ldots, t_n, t_i, t_j \in \mathsf{Term}_{\mathcal{L}_{\mathrm{B}}}$. $\mathsf{Frml}_{\mathcal{L}_{\mathrm{O}}}$ ($\mathsf{Term}_{\mathcal{L}_{\mathrm{B}}}$) is the set of formulas (terms) of $\mathcal{L}_{\mathrm{B}}$.

The definition implies that the cardinality of the language will depend on the set of contingent objects O the language is intended to talk about. In particular, if O is uncountable, then $\mathcal{L}_{\mathrm{O}}$ and $\mathcal{L}_{\mathrm{B}}$ will also be uncountable.

With the details of the languages $\mathcal{L}_{\mathrm{O}}$ and $\mathcal{L}_{\mathrm{B}}$ in place, we need to specify their semantic interpretation. To this end we introduce the notion of a belief frame. Models for $\mathcal{L}_{\mathrm{B}}$ will be defined relative to such a belief frame.

Definition 7.4.2 (Belief frame). A belief frame F is a tuple $\langle \mathrm{U}, W, H, D\rangle$ where U is the universe of discourse, $W \neq \emptyset$ is a set of worlds, and $D : W \to \mathcal{P}(\mathrm{U})$ for all $w \in W$ is a function that assigns the domain of quantification relative to a world w such that $\mathrm{O} \subseteq \bigcup_{w \in W} D(w)$. Finally, $H : \mathcal{P}(\mathsf{Sent}_{\mathcal{L}_{\mathrm{B}}}) \to W \times W$ is a function that generates a serial (right-unbounded) doxastic accessibility relation relative to a set of sentences.[20]

It's worth noting that every non-syntactic object needs to exist at some world. No condition of this kind is imposed on the syntactic objects, that is, the expressions of $\mathcal{L}_{\mathrm{O}}$. These may live in U without "coming into existence" at any world.

We now define an interpretation over a belief frame F. As mentioned at the beginning of this section the interpretation will act rigidly on $\mathcal{L}_{\mathrm{O}}$ but is allowed to vary from world to world for the remaining vocabulary of $\mathcal{L}_{\mathrm{B}}$. As we shall see later, not every interpretation over F gives rise to a belief model since only specific interpretations will ultimately be deemed adequate.

Definition 7.4.3 (Interpretation). Let F be a belief-frame. An interpretation I is a function that assigns an interpretation to the non-logical vocabulary of $\mathcal{L}_{\mathrm{B}}^{-}$ ($\mathcal{L}_{\mathrm{B}}$ without the truth predicate) relative to a possible world such that for all w, $v \in W$

(i) for all individual constants $k \in \mathcal{L}_{\mathrm{O}}$, $I(k, w) = I(k, v) \in U$; for all individual constants $k \in (\mathcal{L}_{\mathrm{B}} - \mathcal{L}_{\mathrm{O}}), I(k, w) \in \mathrm{O}$;

(ii) for all function symbols f^n of $\mathcal{L}_{\mathrm{O}}$, $I(f, w) : U^n \to U$ and for all $u_1, \ldots$, $u_n \in U$

$$I(f^n, w)(u_1, \ldots, u_n) = I(f^n, v)(u_1, \ldots, u_n);$$

(iii) $\mathrm{U} - \mathsf{Sent}_{\mathcal{L}_{\mathrm{B}}} \subseteq I(\mathrm{A}, w) \subseteq \mathrm{U}$;

(iv) If P^n is a predicate constant of $\mathcal{L}_{\mathrm{B}}$, then $I(P^n, w) \subseteq \mathrm{U}^n \times \mathrm{U}^n$; for $P^n \in (\mathcal{L}_{\mathrm{B}} - \mathcal{L}_{\mathrm{O}})$

- if $u_i \in \mathsf{Expr}_{\mathcal{L}_{\mathrm{B}}}$, for some $1 \leq i \leq n$, then for all $e \in \mathsf{Expr}_{\mathcal{L}_{\mathrm{B}}}$

$$\langle u_1, \ldots, u_i, \ldots, u_n \rangle \in I(P^n, w)^{(+/-)} \Leftrightarrow \langle u_1, \ldots, e, \ldots, u_n \rangle \\ \in I(P^n, w)^{(+/-)};$$

for $P^n \in \mathcal{L}_{\mathrm{O}}$,

- $I(P^n, w) = I(P^n, v) = \langle X, Y \rangle$ with $Y = U - X$.

The definition guarantees that the vocabulary of $\mathcal{L}_{\mathrm{O}}$ is interpreted rigidly and that the expressions of $\mathcal{L}_{\mathrm{B}} - \mathcal{L}_{\mathrm{O}}$ do not discriminate between syntactic objects but only objects of O. It is worth highlighting that Definition 7.4.3 does not guarantee that sentence denoting singular terms will always be rigid designators: we only know that if the term is an $\mathcal{L}_{\mathrm{O}}$-expression, then it will be a rigid designator.[21] We will examine the issue more closely in Section 7.6.1 but until then, to keep things as simple as possible, we conceive of all sentence denoting singular terms as rigid designators. Finally, condition (iii) of Definition 7.4.3 specifies that all objects in U that are not sentences will always be in the extension of the Awareness predicate; that is, the interpretation of the Awareness predicate at a world can only vary with respect to the sentences in its extension. This may seem awkward for one might wonder what it means to be aware of some non-syntactic object. However, the point is that we intend the Awareness predicate to discriminate only between different sentences rather than arbitrary objects.[22] As explained at the beginning of Section 7.4, in the semantics of two-level belief representation we adopted the doxastic accessibility relation under consideration will depend on which sentences an agent is aware of. As laid out in Remark 7.4.13 below, if we were to work in a system of one-level belief representation instead, the interpretation of the Awareness predicate should discriminate between names of sentences, or more correctly, names of propositions.

Definition 7.4.3 does not guarantee that the syntactic operations and expressions are interpreted in a desirable way, that is, that they are interpreted as the syntactic operations and the expressions they intend to denote. Interpretations that guarantee such an intended interpretation will be called adequate and give rise to a belief model.

Definition 7.4.4 (Adequate Interpretation, Belief Model, Assignment). Let (U, J) be the standard model (of the syntax theory of) of $\mathcal{L}_{\mathrm{O}}$. We call an interpretation function I adequate iff for all non-logical constants (individual, function, predicate constants) $\zeta \in \mathcal{L}_{\mathrm{O}}$ and all $w \in W$, $\mathsf{J}(\zeta) = I(\zeta, w)$. If

I is an adequate interpretation, then $M = (F,I)$ is called a belief model. An assignment $\beta : Var_{\mathcal{L}_{\mathrm{O}}} \to \mathrm{U}$ assigns to each variable an object in U.

We now turn to the interpretation of the truth predicate, which is provided by an evaluation function.

Definition 7.4.5 (Evaluation function). Let F be a belief frame. An evaluation function relative to F is a function $f : W \times W \to \mathcal{P}(\mathsf{Sent}_{\mathcal{L}_{\mathrm{B}}})$ that assigns to each pair of worlds a set of sentences—the interpretation of the truth predicate. The set of all evaluation functions relative to a frame F is denoted by Val_F.

Next we introduce the index set of a formula φ, which serves as the input to the H function that outputs an accessibility relation. Intuitively the index set of φ consists of all sentences ψ such that Tt_ψ is a subformula of φ and where the agent is not aware which proposition the sentence ψ expresses, i.e., the set of all those representations occurring (explicitly and implicitly) in φ which are not transparent to the agent. It is in the definition of the index set where the differences between systems of one-level and two-level belief representation become most apparent. While in the present case of two-level belief representation the index set will consist of a set of sentences, it will be a set of names (of propositions) in the case of one-level belief representation, namely, the set of those names t_ψ such that the agent is not aware of the proposition t_ψ denotes.

Definition 7.4.6 (Index set). Let φ be a formula of $\mathcal{L}_{\mathrm{B}}$, $t^{M,w}[\beta]$ be the interpretation of a term t in M at w relative to a variable assignment β. Then the index set φ^β_w of φ relative to a belief model $\mathcal{M}$, a world w and an assignment function β is defined by the following recursion:

$$\varphi^\beta_w := \begin{cases} \varnothing, & \text{if } \varphi \in \mathcal{L}^-_{\mathrm{B}} \text{ or } (\varphi \doteq \mathrm{T}t \;\&\; t^{M,w}[\beta] \notin \mathsf{Sent}_{\mathcal{L}_{\mathrm{B}}}); \\ \{\psi\}, & \text{if } \varphi \doteq \mathrm{T}t \;\&\; t^{M,w}[\beta] = \psi \in \mathcal{L}^-_{\mathrm{B}}; \\ \psi^\beta_w, & \text{if } \varphi \doteq \mathrm{T}t \;\&\; t^{M,w}[\beta] = \psi \in I(\mathrm{w},\mathrm{A}); \\ \{\psi\} \cup \psi^\beta_w, & \text{if } \varphi \doteq \mathrm{T}t \;\&\; t^{M,w}[\beta] = \psi \notin I(\mathrm{w},\mathrm{A}); \\ \psi^\beta_w, & \text{if } \varphi \doteq \neg\psi \text{ or } \varphi \doteq \mathrm{B}\psi; \\ \psi^\beta_w \cup \chi^\beta_w, & \text{if } \varphi \doteq \psi \wedge \chi; \\ \bigcup_{d \in D(w)} \psi(v)^{\beta(v:d)}_w, & \text{if } \varphi \doteq \forall v \psi. \end{cases}$$

We now define truth in a model at a world in a belief-model relative to an evaluation function f. Indeed, as already implicit in Definition 7.4.5 we define truth in a model relative to a pair of worlds (w, v) where w is the world in which we evaluate the given formula and v is the world relative to which the index set of the formula is defined. The reason for the two-dimensional interpretation is that we want to evaluate a formula relative to the interpretation of the Awareness predicate at the initial world of evaluation rather than one of its doxastic alternatives which may be relevant for evaluating one of its subformulae at a later stage of the semantic computation. For example, consider the formula $\mathrm{BBT}t_{\varphi}$. In evaluating the formula at w, that is (w, w), we first check whether $\mathrm{BT}t_{\varphi}$ is true in all doxastic alternatives v of w given the accessibility relation generated by the index set of $\mathrm{BT}t_{\varphi}$ at w. In the next step, we wish to consider whether $\mathrm{T}t_{\varphi}$ is true at all doxastic alternatives of v relative to the index set of $\mathrm{T}t_{\varphi}$ at w—rather than the index set of $\mathrm{T}t_{\varphi}$ at v—but without appealing to a two-dimensional interpretation there is no way we can retrieve the starting point of our semantic computation. Again, the point is that what matters in the semantic evaluation of the $\mathrm{BBT}t_{\varphi}$ is what the agent is aware of at world w rather than at a doxastic alternative v. Moreover, the index set of $\mathrm{T}t_{\varphi}$ relative to w may be different to the index set of $\mathrm{T}t_{\varphi}$ relative v to since the interpretation of the Awareness predicate may change and, as a consequence, $\mathrm{BT}t_{\varphi}$ may be true at v, say, relative to the index set of $\mathrm{T}t_{\varphi}$ at w but false relative to its index set at v.

Definition 7.4.7 (Strong Kleene truth in a belief model). Let F be a belief frame, $f \in \mathsf{Val}_F$ and β a variable assignment. We define the notion of truth in a belief model M relative to the evaluation function f and the assignment β at a world-pair (w, v) according to the strong Kleene scheme for formulas φ of $\mathcal{L}_{\mathsf{B}}$ $(M, (w, v) \vDash^{f}_{\mathsf{sk}} \varphi[\beta])$ by an induction on the positive complexity of φ:

(i) $M, (w, v) \vDash^{f}_{\mathsf{sk}} s = t[\beta] \Leftrightarrow s^{M,w}[\beta] = t^{M,w}[\beta]$

(ii) $M, (w, v) \vDash^{f}_{\mathsf{sk}} \neg s = t[\beta] \Leftrightarrow s^{M,w}[\beta] \neq t^{M,w}[\beta]$

(iii) $M, (w, v) \vDash^{f}_{\mathsf{sk}} P^n t_1, \ldots, t_n s = t[\beta] \Leftrightarrow \left\langle t_1^{\mathcal{M},ww}[\beta], \ldots, t_n^{\mathcal{M},w}[\beta] \right\rangle \in I(P^n, w)^{+};$

(iv) $M, (w, v) \vDash^{f}_{\mathsf{sk}} \neg P^n t_1, \ldots, t_n[\beta] \Leftrightarrow \left\langle t_1^{\mathcal{M},ww}[\beta], \ldots, t_n^{\mathcal{M},w}[\beta] \right\rangle \in I(P^n, w)^{-};$

(v) $M, (w, v) \vDash^{f}_{\mathsf{sk}} \mathrm{A}t[\beta] \Leftrightarrow t^{M,w}[\beta] \in I(\mathrm{A}, w)$

(vi) $M, (w, v) \vDash^{f}_{\mathsf{sk}} \neg \mathrm{A}t[\beta] \Leftrightarrow t^{M,w}[\beta] \notin I(\mathrm{A}, w)$

(vii) $M, (w, v) \vDash^{f}_{\mathsf{sk}} \mathrm{T}t[\beta] \Leftrightarrow t^{M,w}[\beta] \in f(w, v)$

(viii) $M, (w, v) \vDash^{f}_{\mathsf{sk}} \neg \mathrm{T}t[\beta] \Leftrightarrow (\underset{\cdot}{\neg}\, t)^{M,w}[\beta] \in f(w, v)$ or $t^{M,w}[\beta] \notin \mathsf{Sent}_{\mathcal{L}_{\mathrm{B}}},$

(ix) $M, (w, v) \vDash^{f}_{\mathsf{sk}} \neg\neg\psi[\beta] \Leftrightarrow M, (w, v) \vDash^{f}_{\mathsf{sk}} \psi[\beta]$

(x) $M, (w, v) \vDash^{f}_{\mathsf{sk}} \psi \wedge \chi[\beta] \Leftrightarrow (M, (w, v) \vDash^{f}_{\mathsf{sk}} \psi[\beta]$ and $M, (w, v) \vDash^{f}_{\mathsf{sk}} \chi[\beta])$

(xi) $M, (w, v) \vDash^{f}_{\mathsf{sk}} \neg(\psi \wedge \chi)[\beta] \Leftrightarrow (M, (w, v) \vDash^{f}_{\mathsf{sk}} \neg\psi[\beta]$ or $M, (w, v) \vDash^{f}_{\mathsf{sk}} \neg\chi[\beta])$

(xii) $M, (w, v) \vDash^{f}_{\mathsf{sk}} \forall x\psi[\beta] \Leftrightarrow$ for all $d \in D(w)(M, (w, v) \vDash^{f}_{\mathsf{sk}} \psi[\beta(x : d)])$

(xiii) $M, (w, v) \vDash^{f}_{\mathsf{sk}} \neg\forall x\psi[\beta] \Leftrightarrow$ there is an $d \in D(w)(M, (w, v) \vDash^{f}_{\mathsf{sk}} \neg\psi[\beta(x : d)])$

(xiv) $M, (w, v) \vDash^{f}_{\mathsf{sk}} \mathrm{B}\psi[\beta] \Leftrightarrow \forall z(H_{\psi_v^{\beta}} wz \Rightarrow M, (z, v) \vDash^{f}_{\mathsf{sk}} \psi[\beta])$

(xv) $M, (w, v) \vDash^{f}_{\mathsf{sk}} \neg\mathrm{B}\psi[\beta] \Leftrightarrow \exists z(H_{\psi_v^{\beta}} wz \;\&\; M, (z, v) \vDash^{f}_{\mathsf{sk}} \neg\, \psi[\beta]).$

If a formula φ is true in the belief model M at (w, w) relative to the evaluation function f and assignment β, we write $M, w \vDash^{f}_{\mathsf{sk}} \varphi[\beta]$ and say that φ is true in the belief model M at w relative to the evaluation function f an assignment β; if φ is true in M relative to the evaluation function f and the assignment β for all $w \in W(M, w \models^f_{\mathsf{sk}} \varphi[\beta])$, we write $M \models^f_{\mathsf{sk}} \varphi$ and say that φ is true in M relative to the evaluation function f and the assignment β. We say that φ is true in F under a given evaluation function f and assignment β $(F \models^f_{\mathsf{sk}} \varphi[\beta]))$ iff for all belief-models M on F, $M \models^f_{\mathsf{sk}} \varphi[\beta]$. In general, we drop reference to an assignment β if the formula is true relative to all assignments.

This concludes the specifications of the semantics for truth and belief. However, it remains to be shown that adequate interpretations of the truth predicate can be constructed over arbitrary belief models, that is, that there are evaluation functions f such that (TSW) holds at each world. Fortunately, this can be shown by running a standard Kripkean construction. To this end, we first define an ordering on Val_F.

DEFINITION 7.4.8 (Ordering). Let $f, g \in \mathsf{Val}_F$. We set $f \leq g$ iff $f(w, v) \subseteq g(w, v)$ for all $w, v \in W$.

It is not difficult to see that $\models_{\mathsf{sk}}$ is a monotone evaluation scheme relative to the ordering $\leq$.

FACT 7.4.9 (Monotonicity). For $f, g \in \mathsf{Val}_F$ and for all $\varphi \in \mathcal{L}_{\mathsf{B}}$

$$f \leq g \Rightarrow \forall w, v \in W(M, (w, v) \models^f_{\mathsf{sk}} \varphi \Rightarrow M, (w, v) \models^g_{\mathsf{sk}} \varphi).$$

Fact 7.4.9 guarantees the existence of fixed points for arbitrary belief frames F, that is, evaluation functions f such that for all $\varphi \in \mathsf{Sent}_{\mathcal{L}_{\mathrm{B}}}$ and all $w, v \in W$

$$F, (w, v) \models^f_{\mathsf{sk}} \varphi \Leftrightarrow \varphi \in f(w, v).$$

To construct such a fixed point we define a Kripke jump in the customary fashion:

DEFINITION 7.4.10 (Kripke jump). Let F be a frame and Val_F the set of evaluation functions relative to F and M a belief model. Then $\mathsf{SK}_{\mathrm{B}} : \mathsf{Val}_F \to \mathsf{Val}_F$ is an operation such that

$$[\mathsf{SK}_{\mathrm{B}}(f)](w, v) := \{\varphi \mid M, (w, v) \models^f_{\mathsf{sk}} \varphi\}.$$

The existence of fixed points of SK_{B} then follows by the usual arguments (see, e.g., Kripke, 1975; Visser, 1989).

PROPOSITION 7.4.11 (Fixed points). Let F be a frame and M a belief model on F. Then there exists an evaluation function $f \in \mathsf{Val}_F$ such that

$$\mathsf{SK}_{\mathrm{B}}(f) = f.$$

As an immediate corollary of Proposition 7.4.11, f provides an adequate interpretation of the truth predicate relative to every world $w \in W$:

COROLLARY 7.4.12 (Truth model). Let F be a frame and $f \in \mathsf{Val}_F$ a fixed point of SK_{B}. Then for all $w, v \in W$ and $\varphi \in \mathcal{L}_{\mathrm{B}}$

$$M,(w,v) \vDash^{f}_{\mathsf{sk}} \mathrm{T}t_{\varphi} \Leftrightarrow M,(w,v) \vDash^{f}_{\mathsf{sk}} \varphi.$$

This concludes out presentation of our semantics of truth and belief within the framework of two-level belief representation. Before we turn to some of the consequences of the semantics we sketch out a semantics for one-level belief representation,

REMARK 7.4.13 (One-level belief representation). The present semantics was developed with a two-level belief representation in mind, that is, the language contains names of sentences, which in turn express propositions (or some other attidunal object). A semantics of one-level belief representation can be obtained by implementing the following changes:

- the syntax theory needs to be conceived of as, or replaced by, a theory of structured propositions;[23]
- the arguments of H need to be sets of names of propositions rather than propositions;
- the interpretation of A will be a set of names rather than a set of their denotata;
- the set φ^{β}_{w} will also be a set of names; moreover when $\varphi \doteq \mathrm{T}t$, where t denotes a proposition ψ relative to β and $t(\overline{\beta(x)}/x) \notin I(w, \mathrm{A})$ one should probably set $\varphi^{\beta}_{w} := \{t(\overline{\beta(x)}/x)\}$ rather than $\psi^{\beta}_{w} \cup \{t(\overline{\beta(x)}/x)\}$ as suggested by Definition 7.4.6. The reason is that in this case if $t(\overline{\beta(x)}/x) \notin I(w, \mathrm{A})$ the agent will have no grasp of the proposition denoted by t and, in particular, they will not be aware of the belief-representation, i.e. names of propositions, the proposition appeals to. In contrast, in the case of two-level belief representations the agent is under no illusion what sentence a particular name refers to: they are only not aware which particular proposition the sentence expresses.

7.4.2 *Formal Consequences*

The semantics behaves as intended in the sense that (OS) is not generally true at a world w in a belief model, indeed both directions of (OS) fail:

$$(\not\Rightarrow) \quad M, w \vDash^{f}_{\mathsf{sk}} \mathrm{B}\varphi \not\Rightarrow M, w \vDash^{f}_{\mathsf{sk}} \mathrm{BT}t_{\varphi}.$$

$$(\not\Leftarrow) \quad M, w \vDash^{f}_{\mathsf{sk}} \mathrm{B}\varphi \not\Leftarrow M, w \vDash^{f}_{\mathsf{sk}} \mathrm{BT}t_{\varphi}.$$

However, as intended, (OS) holds under idealization conditions; that is, if $t^{M,w}_{\varphi}[\beta] \in I(\mathrm{A}, w)$, then

$$M, w \vDash^{f}_{\mathsf{sk}} \mathrm{B}\varphi[\beta] \Leftrightarrow M, w \vDash^{f}_{\mathsf{sk}} \mathrm{BT}t_{\varphi}[\beta].$$

This follows from the fact that if $t^{M,w}_{\varphi}[\beta] \in I(w, \mathrm{A})$, $\varphi^{\beta}_{w} = (\mathrm{T}t_{\varphi})^{\beta}_{w}$.

For the T-free fragment of the language the B-operator behaves like a standard modal operator in possible world semantics, that is, for the T-free fragment B is a modal operator of a normal modal logic. This is in stark contrast to the behavior of B if applied to sentences in which T occurs. In this case B is a non-normal and indeed a hyperintensional and non-algebraic operator, that is, let $\|\varphi\|^{f}_{M} := \{w \in W \mid M, w \models^{f}_{\mathsf{sk}} \varphi\}$, then

$$\|\varphi\|^{f}_{M} = \|\psi\|^{f}_{M} \Rightarrow \|\mathrm{B}\varphi\|^{f}_{M} = \|\mathrm{B}\psi\|^{f}_{M}$$

does not generally hold if the truth predicate occurs in either φ or ψ. In this respect our semantics is similar to semantics of non-idealized belief such as classical Awareness semantics or Impossible worlds semantics. But while in these semantics whether a formula $\mathrm{B}\varphi$ or $\mathrm{BT}t_{\varphi}$ is satisfied at some world would depend solely on the Awareness operator or the Impossible world, our semantics provides independent truth conditions for the two formulas, which happen to coincide if t_{φ} is in the interpretation of the Awareness predicate in the world under consideration.

One consequence of the non-normality of B is that the operator does not generally commute with conjunction at a world w if either of the conjuncts has a subformula of the form Tt such that $t^{M,w}[\beta] \notin I(\mathrm{A}, w)$:

$$M, w \models^{f}_{\mathsf{sk}} \mathrm{B}(\psi \wedge \chi) \leftrightarrow M, w \models^{f}_{\mathsf{sk}} \mathrm{B}\psi \wedge \mathrm{B}\chi.$$ [24]

The reason is that while we may believe ψ and φ in some way we might not believe them in the same way. It is precisely for this reason that it is possible for both $\mathrm{B}\varphi$ and $\mathrm{B}\neg\mathrm{T}t_{\varphi}$ to be true at a world w. However, it is impossible for $\mathrm{B}(\varphi\wedge\ \neg\mathrm{T}t_{\varphi})$ to be true at any world w since this would imply that we believe φ and $\neg\mathrm{T}t_{\varphi}$ in the same world, which contradicts Corollary 7.4.12. This kind of phenomena also arises in contextualist theories of belief, where it is possible for Clara to believe that Superman is strong and to believe that Clark Kent is not strong but impossible for her to believe that Superman is strong and that Clark Kent is not strong in the same way.

Moreover, due to the same phenomenon an agent will believe the T-scheme for grounded sentences despite the fact that (OS) will not generally hold at a world even for such sentences; that is, if $F \models^{f}_{\mathsf{sk}} \mathrm{T} \neg t_{\varphi} \leftrightarrow \neg\mathrm{T}t_{\varphi}$, then

$$F \models^{f}_{\mathsf{sk}} \mathrm{B}\,(\mathrm{T}t_{\varphi} \leftrightarrow \varphi).$$ [25]

At first glance the semantics seems to get a lot of things right, but how does it apply to the case of Max and Clara, respectively? Does it yield the correct semantic explanations and predictions?

7.5 Taking Stock: Anne, Clara, and Max

We started the chapter by observing that standard PWS for belief combined with the idea that semantic states-of-affairs supervene on non-semantic states of affairs leads to the undesirable consequence that believing and believing-true are semantically equivalent. Such a semantics, we argued, yields counterintuitive accounts and predictions in the case of Clara and Max, respectively. The moral we drew from this observation was that the semantic value of a belief report depends on the way an agent believes—akin to an idea prominent in contextualist theories of belief reports. In particular, we argued that the way we believe φ will depend on the syntactic representations of beliefs occurring in φ and that unless an agent is aware of t_φ, an agent will believe φ and $\mathrm{T}t_\varphi$ in different ways: they may believe a proposition in one way but fail to believe it in another way. On this semantic picture, believing and believing-true are no longer semantically equivalent and we obtain a neat semantic explanation of this fact. But what precisely happens in the cases of Max and Clara, respectively? Does the new semantics yield correct semantic predications?

Let us start with Clara and assume (DQ) is correct, that is, that Clara believes that Clark Kent is strong. In this case, it seems, we are compelled to accepting that Clara is not fully aware of the sentence 'Clark Kent is strong' for it is part of the story that Clara does not believe the proposition in a 'Clark Kent is strong'-way, indeed Clara does not seem fully aware of what proposition 'Clark Kent is strong' expresses. (OS) cannot be applied, that is, Clara does not believe that 'Clark Kent is strong' is true. Notice, however, that our semantics does not commit us to accepting (DQ), i.e., to accept that Clara believes that Clark Kent is strong because Clara accepts the sentence 'Superman is strong' and, according to our story, also believes that 'Superman is strong' is true: it is perfectly acceptable according to our semantics for Clara to believe that Clark Kent is strong in a 'Superman is strong'-way but not in a representation independent way, indeed, this seems to be a rather plausible view. But no matter one's position in this respect, the semantics provides the flexibility of reporting Clara's beliefs appropriately.

What about Max? We have already seen that Max's use of the truth predicate does not qualify as a disquotational use and (OS) should not be applicable in this context. Admittedly, even though we have assumed a two-level belief representation, in Max's case it would seem more appropriate to make the way of believing dependent on the name 'Goldbach's conjecture' rather than the sentence it denotes. After all, Max may be in no doubt about which proposition is expressed by a particular sentence expressing Goldbach's conjecture but, according to our story, he does not seem to be aware of the sentence the term 'Goldbach's conjecture' denotes.[26] Now, independently of the particulars of our semantics it seems clear that Max is not aware of the proposition represented, be it directly or indirectly, by the name 'Goldbach's conjecture'. Max does not believe that every even number > 2 is the sum of two primes in a representation-independent way. Rather the only way Max believes the proposition is in a 'Goldbach's conjecture'-way of believing—(OS) cannot be applied, that is, our semantics gives the correct semantic assessment: 'Max believes that Goldbachs conjecture is true' and 'Max believes that every even number > 2 is the sum of two primes' are not semantically equivalent.

It seems that our semantics yields the right outcome in Clara's and Max's cases where believing and believing-true are not semantically equivalent. But what about disquotational uses of the truth predicate: can our semantics accommodate such uses as we have claimed at the beginning of Section 7.4? We have seen that the disquotationalist will argue that (2) correctly reports Anne's beliefs but that this requires (OS). In our semantics, application of (OS) is only licensed if Anne is, for every axiom of Euclidean geometry, aware of at least one sentence expressing the axiom qua proposition. From the disquotational perspective this is arguably an acceptable assumption: the disquotationalist's claim is not that Anne would necessarily report her belief in this way. Rather, the claim is that Anne's belief is correctly reported by (2) if an external, observational perspective is assumed. In reporting Anne's belief by (2) the disquotationalist stipulates the transparency of the truth predicate, that is, they stipulate Anne's awareness of the relevant sentences, and it is precisely against the background of this stipulation that the disquotationalist's report of Anne's beliefs is acceptable.

Summing up, our semantics provides an intuitive explanation of why believing and believing-true ought to be semantically differentiated, which neatly applies to the cases of Max and Clara. Moreover, it is sufficiently flexible to accommodate disquotational uses of the truth predicate, that is, uses that treat the truth predicate in a transparent way and for which believing and believing-true turn out to be semantically equivalent. In this respect our semantics should be acceptable to the disquotationalist, although the disquotationalist will need to grant that there may also be non-disquotational uses of 'true'. However, as for every semantics our semantics also produces some—arguably—counterintuitive consequences and faces certain limitations. To conclude the chapter, we discuss some of these limitations and point toward some alternative semantic explanations for distinguishing between believing and believing-true.

7.6 Limitations and Alternative Semantic Explanations

The semantics we presented provides a greater amount of flexibility than standard PWS, which enables the semantics to differentiate between believing and believing-true. But the flexibility of the semantics has its limitations. In this section, we flag two such limitations before outlining some alternative explanations of the semantic difference between believing and believing-true. The first limitation stems from the fact that the way of believing is fully governed by the syntactic information provided by the formula in the scope of the B-operator and the Awareness set; all other contextual information is discarded as irrelevant. The second limitation is inherited from PWS semantics: propositions are still conceived of as sets of possible worlds (or states), which has some at least prima facie undesirable side effects.

To illustrate the first problem recall that in discussing some of the formal consequences of our semantics we noted that B was a non-normal modal operator, if applied to a formula which has a subformula of the form T*t*. As a consequence, virtually all logical reasoning will break down in such cases. But frequently agents are perfectly capable of reasoning logically and the semantics should be able to accommodate logical reasoning in such cases. For example, at least at the outset, disjunction introduction in the scope of B seems fine in many cases, i.e.,

$$\frac{\mathrm{B}\varphi}{\mathrm{B}(\varphi \vee \psi)}.^{27}$$

But the inference breaks down because the way of believing may change in course of our logical reasoning. While this may happen in some cases, surely, in most circumstances if an agent is engaged in reflective, logical thinking, we should expect the way of believing to remain constant throughout the reasoning. This suggests that the way of believing is not fully determined by the syntactic information available, but depends more directly on the context of the belief report. More generally, it seems reasonable to assume that the way of believing will frequently be determined by previous discourse and its common ground.[28] Perhaps then a semantics that explicitly appeals to the way of believing needs to embrace a contextualist approach to attitude reports more fully than we acknowledged in Section 7.3.

Let's turn to the second limitation. In our semantics a formula φ is true at a world w if and only if $\mathrm{T}t_\varphi$ is true at w. As a consequence, the (possibly) diverging semantic values of $\mathrm{B}\varphi$ and $\mathrm{BT}t_\varphi$ are due to the different range of quantification of the (interpretation of the) B-operator in the two cases rather than the semantic value of the formulas in the scope of B. But this also means that if a formula φ is true (false) in all worlds, then (OS) will be true for φ and every name t_φ of φ: independently of the range of quantification of B, there will simply be no worlds to falsify (verify) φ. $\mathrm{B}\varphi$ and $\mathrm{B}t_\varphi$ will be semantically equivalent. For example, let $\chi: = \psi \vee \neg\psi$ where ψ is a sentence of the language of syntax $\mathcal{L}_\mathrm{O}$, then (OS) holds, i.e., let M be a belief model on an arbitrary frame F, then for all $w \in W$ and names t_χ of χ

$$M, w \models^{f}_{\mathsf{sk}} \mathrm{B}\chi \Leftrightarrow M, w \models^{f}_{\mathsf{sk}} \mathrm{BT}\,t_\chi.$$

This feature clearly highlights the limited flexibility of our semantics and that the semantics inherits some of the conceptual limitations of standard PWS: it is not possible to distinguish between "different" necessarily true (false) propositions.

In devising semantics there is always a trade-off between the flexibility of the semantics and the availability of systematic, i.e. non ad hoc, semantic explanations. The question then arises whether we can increase the flexibility of the semantics, that is, to allow for the possibility to block all instances of (OS) whilst at the same time providing or retaining principled semantic explanations. Of course, there is hardly a clear-cut answer to the question, and where one theorists will push for a more flexible semantics another theorists will invoke pragmatic strategies to account for allegedly counterintuitive consequences. In this chapter we shall not enter this kind of debate but point to two alternative strategies for blocking (OS), which may allow for a greater amount of semantic flexibility.

7.6.1 Non-Rigid Terms and Scope Distinctions

Throughout this paper we have tacitly assumed that names of expressions of $\mathcal{L}_{\mathrm{B}}$ refer rigidly to these expressions, that is, a name t_φ refers to φ in all possible worlds. But this assumption may be questioned, even if one conceives of proper names as rigid designators because we frequently designate sentences or propositions via definite descriptions rather than proper names viz 'the sentence "..."' or 'the proposition that ...'. Even if definite description are treated as full-fledged singular terms of the language, rather than incomplete symbols à la Russell (1905), they are generally thought to be flaccid designators and to denote different objects at different worlds. Indeed, as we have mentioned in passing, under certain circumstances our semantics allows for non-rigid sentence denoting singular terms even though we have ignored this aspect of the semantics up to this point. But if non-rigid sentence denoting singular terms are envisaged (OS) would fail because a term t_φ can fail to denote the sentence (proposition) φ in all worlds but the actual world—for all we know it could also denote the sentence (proposition) ψ.[29] Of course, (TSW) would then fail likewise, that is, we would only be guaranteed that $\mathrm{T}t_\varphi$ and φ receive the same semantic value in the actual world but not in any other world. However, this would not imply that semantic states-of-affairs do not supervene on non-semantic states of affairs, since (TSW) no longer asserts that 'φ' holds at a world w, if and only if 'the sentence 'φ' (the proposition that φ) is true' holds at w, but possibly, using our previous example, that "φ' holds at w, if and only if, 'the sentence ψ' (the proposition that φ) is true' holds at w. So, at least prima facie, treating sentence/proposition-denoting singular terms as flaccid designators does not seem to clash with any fundamental semantic principle.

This opens up the possibility of semantically distinguishing between believing and believing-true without appealing to different ways of believing. Rather the distinction arises due to non-rigid designation. However, if sentence/proposition-denoting singular terms are treated as non-rigid designators, the question arises how we are to deal with cases like Anne's, i.e., belief reports like (2) where, at least from a disquotational perspective, appealing to (OS) seems to be legitimate and to yield the correct semantic predictions. The most immediate and plausible strategy to accommodate such cases is to appeal to scope distinctions triggered by the definite description and to distinguish between believing of the denotatum of t_φ that it is true and believing that t_φ is true, that is to distinguish between believing-true de re and de dicto. While (OS) fails on the de dicto-reading, it seems, or so the argument would go, acceptable on the de re-reading, that is, we would obtain the following version of (OS):

$$M, w \models^f \mathrm{B}\varphi \Leftrightarrow M, w \models^f \langle \lambda x.\mathrm{BT}x \rangle t_\varphi. \tag{OSR}$$

In Section 7.5 we argued that the disquotationalist assumes an external, observational position in reporting Anne's belief and that their report need not match the way Anne would report her beliefs. This position goes neatly with the strategy of accommodating belief reports like (2) by focusing on the de re reading of believing-true.

Still one may wonder whether understanding sentence/proposition-denoting singular terms to be flaccid designators is the correct analysis of the cases of Max and Clara. First, if one subscribes to the Millean/Kripkean-doctrine that names are rigid designators, then names of sentences/propositions such as Goldbach's conjecture should be taken to rigidly designate their referent.[30] As a consequence, theorists would need to provide an alternative explanation for why (OS) fails in Max's case, which suggests that alluding to non-rigid designation cannot fully replace the appeal to ways of believing in semantic explanations. Second, and more generally, it is not clear that in the cases of Clara and Max the failure of (OS) is due to a de dicto-reading of believing-true. Arguably, in Max's case the equivalence between believing and believing-true seems to break down on the de re-reading likewise. Similarly, at least focusing on the sentential truth predicate, Clara's case also remains puzzling under a de re-reading of believing-true. Treating sentence-denoting singular terms as flaccid designator then does not seem to resolve the original puzzle: the puzzle seems to be a puzzle about believing-true de re and the cases brought forward against (OS) appear to undermine (OSR) likewise. Of course, allowing for sentences and propositions to be denoted non-rigidly in the semantics—not all sentence/proposition-denoting expressions should be treated uniformly—may be interesting for other reasons, but, as mentioned, these non-rigid terms will not be helpful in semantically distinguishing believing from believing-true de re.

7.6.2 More Syntax Strategies

The aim of the chapter was to provide an adequate semantics for truth and belief. To this end, we appealed to PWS for belief, which, as we have seen, yields counterintuitive consequences unless special precautions are taken. However, many theorists working on the semantics of attitude reports will deem such an approach a non-starter, as it is well-known that PWS provides a very coarse grained-arguably too coarse-grained—analysis of attitude reports and semantic content.[31] According to these views, semantic content, i.e. propositions, should not be conceived as sets of worlds or "truth supporting circumstances" (cf. Soames, 1987) but as structured propositions: whilst in PWS the proposition that Mary is smart is conceived of as the set of all those worlds in which Mary is smart, on the structured proposition account it would be, roughly put, the tuple consisting of Mary and the property of being smart, i.e, $\langle$Mary, Smartness$\rangle$. In other words, on the structured proposition account a proposition conveys structural or syntactic information that has been extracted or retained from the sentences that express it. On this view, if belief is conceived as a relation between an agent and a proposition or, possibly, the constituents of the proposition, we should not expect (OS) to be true: 'the proposition that it is true that Mary is smart' will express a proposition along the lines of $\langle$True, $\langle$Mary, Smartness$\rangle\rangle$. While $\langle$Mary, Smartness$\rangle$ and $\langle$True, $\langle$Mary, Smartness$\rangle\rangle$ will be true at exactly the same worlds, there is no guarantee that an agent will be aware of this fact—they may believe the one without believing the other.

If one agrees with the idea that belief should not be a relation between an agent and a set of possible worlds but rather a relation between an agent and structured propositions, there is no puzzle: an agent may believe φ and not believe $\mathrm{T}t_\varphi$ and vice versa despite the fact that φ and $\mathrm{T}t_\varphi$ have the same truth value in every world. In light of this one may be tempted to abandon PWS. However, upon closer inspection explicit appeal to structured proposition seems inessential for producing correct semantic predictions, as the belief relation of the structured proposition theorist can be recovered in the PWS framework. On this view a structured proposition is just a set of worlds represented in a way that conforms to specific structural constraints. Accordingly, a formula $\mathrm{B}\varphi$ will be true at a world w iff φ is true at all doxastic alternatives of w and the agent believes the proposition that φ qua set of possible worlds under the representation φ. To make this idea precise on would need to say when a formula is true at a world under a specific representation. To this end, let

$$\mathrm{Rep} : \mathcal{P}(W) \times W \to \mathcal{P}(\mathsf{Frml}_{\mathcal{L}_\mathrm{B}})$$

be a function that selects the available representations of a proposition qua set of possible worlds relative to each world. This could, e.g., be a class of sentences that are intensionally isomorphic in the sense of Carnap (1947) or Lewis (1970). A formula of the form $\mathrm{B}\varphi$ will then be true in a belief model M and world w, iff

$$\forall v(wRv \Rightarrow M, v \vDash \varphi \;\&\; \varphi \in \mathrm{Rep}_v(\|\varphi\|)).$$[32]

On this view, $\mathrm{B}\varphi$ and $\mathrm{BT}t_\varphi$ will not always be semantically equivalent at w even though (TSW) will hold at every world, that is, $\|\varphi\| = \|\mathrm{T}t_\varphi\|$, because there is no guarantee that

$$\varphi \in \mathrm{Rep}_v(\|\varphi\|)) \Leftrightarrow \mathrm{T}t_\varphi \in \mathrm{Rep}_v(\|\varphi\|)).$$

In principle, such a semantics can individuate beliefs as finely as the sentences of the language but it can also allow for a much coarser individuation of beliefs. However, whereas the formal semantics thus settles the formal puzzle, that is, (OS) is no longer valid on such a semantics, there still remains the need for a principled philosophical explanation of why an agent may stand in the believing relation to ⟨Mary, Smartness⟩ without also standing in the believing relation to ⟨True, ⟨Mary, Smartness⟩⟩. Perhaps appealing to the way of believing may be useful to this effect and, in this case, the sketched semantics could potentially be combined with the semantics for one-level belief representation we outlined in Section 7.4.

7.7 Conclusion

The aim of the chapter was to provide a more adequate semantics for belief and truth, that is, a semantics in which one can semantically distinguish between believing and believing-true. The main novelty of the semantics proposed in this paper is to explicitly appeal to the way of believing in the semantic evaluation of a formula of the form $\mathrm{B}\phi$ where the way of believing is extracted from the syntactic information provided by ϕ. We argued that this provides a neat semantic explanation of why believing and believing-true are not semantically equivalent. We take it that our semantics provides, at the very least, an interesting first step towards an adequate semantics for belief and truth.[33]

Acknowledgments

My research was supported by the European Research Council (ERC) Starting Grant *Truth and Semantics* (TRUST, Grant no. 803684). I wish to thank Carlo Nicolai and Thomas Schindler for helpful comments on the chapter, and Catrin Campbell-Moore, Ollie Tatton-Brown, and the

audiences of talks virtually given at Glasgow, Birmingham, and Bristol for stimulating discussions of the material.

Notes

1. Caie (2012), Jerzak (2019), and, arguably, Halbach and Welch (2009), Campbell-Moore (2015), and Stern (2016), are notable exceptions to this claim. Yet, the semantics presented by all these authors produce the type of unintended consequence discussed in Section 7.2.
2. This is not supposed to be a controversial statement. Of course, the work by theorists working with structured propositions (see Section 7.6) or within certain forms of truth-maker semantics may not subscribe to such an analysis. But as far as developed formal semantics go PWS is, by far, still the dominating approach.
3. Arguably, to formulate (1) one would need to formalize belief as a predicate rather than a sentential operator as customary in PWS. We shall not discuss this issue but assume that a belief predicate can be retrieved in the language via some sort of "Kripke-reduction" (cf. Halbach and Welch, 2009; Stern, 2016, Chapter 4).
4. We do not assume that $\models$ is a classical satisfaction relation, that is, (TSW) is not necessarily equivalent to

   $$M, w \models^f \mathrm{T}t_\varphi \leftrightarrow \varphi. \tag{TS}$$

   Indeed in the semantics for belief and type-free truth we shall construct $\models$ will be a non-classical satisfaction relation according to which (TS) and (TSW) are not equivalent.
5. In particular (OS) holds in the semantics for belief and truth proposed by Caie (2012) and Jerzak (2019).
6. In contrast to the Quinean analysis of attitude reports (Quine, 1956), in studying a semantics for truth in doxastic contexts there is no presupposition that all attitudinal relations implicitly appeal to the truth predicate, e.g. we do not assume the 'believes that "φ"' ought to be always reconstructed as 'believes-true "φ"'.
7. Here, we employ the slightly more explicit formulation given in, e.g., Nelson (2019). Kripke (1979) originally formulated the disquotational principles using 'assents to' instead of 'accepts'. Kripke also lists a number of qualifications that are intended to rule out unusual or atypical circumstances that would interfere with the agent assenting or expressing dissent with a sentence s. We implicitly adopt these qualifications.
8. Cf. Field (1994, pp. 251–252). Field makes a number of qualifications to which we will come back to in due course. See also Künne (2003) and Heck (2020) for a discussion of cognitive equivalence.
9. $\bar{s}$ is a name of the sentence s.
10. Admittedly, Max does not accept 'Every even number > 2 is the sum of two prime numbers.' and, as we shall discuss below, this yields an argument against (TDQ).
11. See, e.g., Picollo and Schindler (2020) for an endorsement of this view.
12. A similar point is made by Heck (2020).
13. The case of Clara is somewhat different. Arguably, the use of the truth predicate is a disquotationally legitimate use. But the disquotationalist would presumable say that 'Superman is strong' does not express that Clark Kent is

strong. Rather it expresses whatever Clara qua competent speaker understands ‘Superman is strong’ to say, or something along these lines.

14. Indeed, we are not aware of a single disquotationalist who defends (OS). Field (2006) seems to explicitly reject the conclusion. Heck (2020) agrees with our assessment that disquotationalist like Field are committed to (OS).
15. We take disquotationalists in contrast to other deflationists to conceive of ‘true’ as a predicate of sentences or utterances rather than of propositions (cf. Künne, 2003).
16. Of course, Crimmins and Perry (1989), like other Russellians, conceive of propositions as structured entities, which is at odds with conceiving propositions as sets of possible worlds, states, or situations as customary in PWS. At this point, our comparison only pertains to the idea that the way of believing impacts the attitudinal relation. See Section 7.6.2 for a discussion of the structured propositions approach.
17. Or rather we remain neutral whether the way of believing has an impact on the semantic evaluation of the attitude report, if it is not explicitly conveyed in reporting the attitude.
18. It is not important for our purpose whether the objects of belief are propositions or some other sort of attitudinal object. The relevant issue is whether the objects of truth and belief coincide or not.
19. To keep the presentation as concise as possible we only allow for one agent—this allows us to omit an index for B and simplifies some of our definitions, yet nothing hinges on this simplifying assumption.
20. Of course, further properties of the doxastic accessibility relation could be imposed. Here, we take seriality to be a minimal condition of a doxastic accessibility relation. However, nothing we say in this chapter will depend on the properties of the doxastic accessibility relation.
21. Suppose the function symbol q represents a function on U that, similar to the num-function, if applied to some element of U outputs the “standard” name of the object. Then $\mathsf{q}(c_1)^\frown\dot{=}^\frown\mathsf{q}(c_2)$ is a name of a sentence. But since the interpretation of c_1 and c_2 may change from world to world, $\mathsf{q}(c_1)^\frown\dot{=}^\frown\mathsf{q}(c_2)$ may denote, say, the sentence $o_1 = o_2$ at world w but the sentence $o_3 = o_5$ at world v.
22. Indeed we could have also defined the interpretation of the Awareness predicate to be a set of sentences only and modify Definition 7.4.6 below accordingly.
23. Admittedly, this sits ill with PWS where propositions are conceived of as sets of possible worlds. We shall ignore this issue for the purpose of the formal semantics but see Section 7.6.2.
24. Similarly, disjunction introduction in the scope of B fails, i.e.,

$$M, w \models^{f}_{\mathsf{sk}} \mathrm{B}\psi \;\not\Rightarrow\; M, w \models^{f}_{\mathsf{sk}} \mathrm{B}(\psi \vee \chi).$$

25. We take this to be a neat feature of our semantics but, to be sure, it is not unproblematic. Because of basically the same phenomenon $\mathrm{B}(\mathrm{T}t_\varphi \vee \varphi)$ and $\mathrm{BT}t_\varphi$ will be equivalent on our semantics, which seems problematic if you think of a case like Max’s. Thanks to Ollie Tatton-Brown for raising this issue.
26. This suggests that the uniform treatment of sentence-denoting singular terms in our semantics may be too coarse-grained and that we should treat different types of singular terms differently.
27. It seems particularly unfortunate that disjunction introduction is valid iff ψ does not contain a subformula of the form $\mathrm{T}t$ but invalid otherwise. The

reason for this asymmetry is that in the former case ψ does not affect the way of believing while in the latter case it does.

28. To some extent this can be accommodated by the fact that the Awareness set is contextually controlled but arguably this will only suffice in cases where the syntactic information remains constant throughout the discourse.
29. Without special restrictions in place a term t_φ could denote, e.g., the symbol '(' or some arbitrary other object in the domain. In our semantics sentence denoting singular terms will always denote sentences but the philosophical question remains of whether this is a plausible assumption once we have allowed for non-rigid designators: what guarantees that non-rigid designators always designate objects of right (syntactic) category?
30. Arguably, Goldbach's conjecture is a descriptive name but even descriptive names are typically thought to rigidly designate their denotatum. Moreover, the failure of (OS) in Max case does not seem due to the descriptive property conveyed by 'Goldbach's conjecture'.
31. See, Soames (1987), King (2013) or Nelson (2019) for discussion.
32. Again, $\|\varphi\|$ is is the set of worlds in which φ is true, i.e.,

$$\|\varphi\| := \{w \in W \mid M, w \models \varphi\}$$

We can assume $\|\varphi\|$ to be defined at this stage of the semantic evaluation since φ is of lower complexity than Bφ. Ultimately, the semantics would amount to a variant of Awareness semantics in the sense of Fagin et al. (1995) in which the Awareness set is constrained by a number of specific rules.

33. It is perhaps worth noting that our proposal is not committed to the notion of full belief, as opposed to the notion partial belief or credences. The principal strategy underlying our semantics, that is, the idea of making the accessibility relation dependent on the way of believing can also be used to provide a semantics for truth and partial belief.

References

Asher, N. and Kamp, H. (1989). Self-reference, attitudes and paradox. In Chierchia, G., Partee, B. H., and Turner, R., editors, *Properties, Types, and Meaning. Vol. I: Foundational Issues*, pages 85–158. Kluwer.

Caie, M. (2012). Belief and indeterminacy. *Philosophical Review*, 121(1): 1–54.

Campbell-Moore, C. (2015). How to express self-referential probability. A Kripkean proposal. *The Review of Symbolic Logic*, 8(4): 680–704.

Carnap, R. (1947). *Meaning and Necessity*. University of Chicago Press.

Crimmins, M. and Perry, J. (1989). The prince and the phone booth: Reporting puzzling beliefs. *The Journal of Philosophy*, 86(12): 685–711.

Fagin, R., Halpern, J. Y., Moses, Y., and Vardi, M. Y. (1995). *Reasoning about Knowledge*. MIT Press.

Field, H. (1994). Deflationist views of meaning and content. *Mind*, 103: 249–285.

Field, H. (2006). Compositional principles vs. schematic reasoning. *The Monist*, 89(1): 9–27.

Frege, G. (1892). Über Sinn und Bedeutung. *Zeitschrift für Philosophie und philosophische Kritik*, 100: 25–50.

Garson, J. W. (2001). Quantification in Modal Logic. In Gabbay, D. and Guenther, F., editors, *Handbook of Philosophical Logic*, volume 3, pages 267–323. Kluwer Academic Publishers, 2nd edition. First published in 1984.

Gupta, A. and Belnap, N. (1993). *The Revision Theory of Truth*. The MIT Press.

Halbach, V. and Leigh, G. (2021). *The Road to Paradox: A Guide to Syntax, Truth, and Modality*. Cambridge University Press, forthcoming.

Halbach, V. and Welch, P. (2009). Necessities and necessary truths: A prolegomenon to the use of modal logic in the analysis of intensional notions. *Mind*, 118: 71–100.

Heck, R. K. (2020). Disquotationalism and the Compositional Principles. In Nicolai, C. and Stern, J., editors, *Modes of Truth: The Unified Approach to Modality, Truth, and Paradox*. Routledge, forthcoming.

Hintikka, J. (1962). *Knowledge and Belief*. Cornell University Press.

Jerzak, E. (2019). Non-classical knowledge. *Philosophy and Phenomenological Research*, 98(1): 190–220.

King, J. C. (2013). On fineness of grain. *Philosophical Studies*, 163(3): 763–781.

Kripke, S. A. (1972). Naming and necessity. In Davidson, D. and Harman, G., editors, *Semantics of Natural Language*, pages 253–355. Springer Netherlands.

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

Kripke, S. A. (1979). A puzzle about belief. In Margalit, A., editor, *Meaning and Use*, pages 239–283. Springer.

Künne, W. (2003). *Conceptions of Truth*. Oxford University Press.

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

Lewis, D. (1970). General semantics. *Synthese*, 22(1/2): 18–67.

Nelson, M. (2019). Propositional attitude reports. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Spring 2019 edition.

Picollo, L. and Schindler, T. (2020). Is deflationism compatible with compositional and Tarskian truth theories. In Nicolai, C. and Stern, J., editors, *Modes of Truth: The Unified Approach to Modality, Truth, and Paradox*. Routledge, Forthcoming.

Quine, W. V. O. (1956). Quantifiers and propositional attitudes. *The Journal of Philosophy*, 53.

Russell, B. (1905). On denoting. *Mind*, 14(56): 479–493.

Soames, S. (1987). Direct reference, propositional attitudes, and semantic content. *Philosophical Topics*, 15(1): 47–87.

Stern, J. (2014a). Modality and axiomatic theories of truth I: Friedman-Sheard. *The Review of Symbolic Logic*, 7(2): 273–298.

Stern, J. (2014b). Modality and axiomatic theories of truth II: Kripke-Feferman. *The Review of Symbolic Logic*, 7(2): 299–318.

Stern, J. (2016). *Toward Predicate Approaches to Modality*, volume 44 of Trends in Logic. Springer, Switzerland.

Tarski, A. (1944). The semantic conception of truth. *Philosophy and Phenomenological Research*, 4: 341–376.

Visser, A. (1989). Semantics and the liar paradox. In Gabbay, D., editor, *Handbook of Philosophical Logic*, pages 617–706. Dordrecht.

Yablo, S. (1982). Grounding, dependence, and paradox. *Journal of Philosophical Logic*, 11: 117–137.

8 Indeterminate Truth and Credences

Catrin Campbell-Moore

8.1 Introduction

Suppose you find yourself in the following unfortunate situation:

> PASSPORT
>
> If you have credence ≥ 0.5 that you'll remember your passport, then when the time comes you'll end up forgetting it (you'll get on with other things).
>
> And if you do not have credence ≥ 0.5, then you *will* end up remembering it (you'll spend your time worrying about it).
>
> And you know this about yourself. What should your credence be?
>
> (Closely related to the Archer case from Joyce, 2018, or Basketball from Caie, 2013)

Suppose you adopt credence 0.2 that you'll remember your passport. Then you'll get on with other things and will forget your passport. And you know this. So then you should be certain that you'll forget your passport, i.e., adopt credence 1. But were you to adopt credence 1 that you'll forget your passport, you would spend all your time worrying about it, and thus would remember it. And since you know this about yourself, this would recommend adopting credence value 0. More generally, any credence value you assign will undermine itself in this sort of way.

Such scenarios have recently been discussed as a challenge for rationality.[1] A rational opinion state should not undermine its own adoption. If it does, it cannot be relied on. And in a case like PASSPORT, every credence value is undermining so there seems to be no rational options.

We propose to parallel the kind of underminingness found in PASSPORT to that due to the liar sentence:

> Liar: Liar is not true.

We might naturally reason about the liar sentence as follows: Is it true or not? Suppose it were true. Then since it says "Liar is not true", and Liar is true, we can conclude that it is false. Suppose it were not true. Then since it says "Liar is not true", we can conclude it is true. Reflecting on its truth value always results in a contradictory truth value; we might describe this as truth value assignments undermining themselves.

McGee (1989, 1990) argues that one should consider definite truth. Some sentences are definitely true, others are definitely false, but some, such as Liar, are indefinite. Inspired by the influential construction of Kripke (1975), in its supervaluational form, one can obtain an account of revision of definite truth value verdicts which allows for "fixed points". We might describe such fixed points as accounts of indefinite truth which are not undermining.

In this chapter, we propose to apply similar considerations to rational credence. The chapter starts with a summary of the account of truth in Section 8.2, and a presentation the classical account of credences and cases like Passport in Section 8.3.

To account for these kinds of cases we propose allowing the notion of credence to be indeterminate, in particular, allowing that no definite credence value is assigned to Passport. Our first question is how to model this (Section 8.4). One could just focus on questions such as whether one's credence in φ is definitely equal to r or definitely not equal to r, or neither. But we will suggest that it's more natural to directly consider the indeterminate credal state as a set of precise credence functions: its set of precisifications. So instead of focusing on definite judgements and considering the collection of precisifications as derivative, we directly work with the set of precisifications. This is a model of belief which is of independent interest in formal epistemology. Moreover, we can understand other supervaluational models by considerations of the precisifications, so by directly working with the set of precisifications, we can then also apply the results to other proposed models.

We then need to consider how to revise one's indeterminate credences, i.e., describe a supervaluational Kripkean jump. Section 8.5 proposes a very natural supervaluational jump, which revises a set of precisifications simply by revising each of the individual precisifications, we call this $\mathcal{R}$. But we will see that this can sometimes result in triviality.

To see how we might avoid this triviality, Section 8.6 returns to considering the supervaluational Kripkean jump for truth. For truth, one typically focuses simply on whether sentences are definitely true, definitely not true, or indefinite instead of focusing on the sets of precisifications themselves, even though it's this resultant set of precisifications that's important for defining the supervaluational Kripkean jump. We might also ask what happens when we focus on the set of precisifications themselves in the case of truth, and revise it by simply revising each of its members ($\mathcal{R}$). It too leads to triviality due to the McGee sentence. But the usual supervaluational Kripkean jump for truth (thought of as applying to an assignment of definite truth value verdicts) does not correspond to $\mathcal{R}$, but instead additional precisifications are added: those that agree on any definite truth value verdicts. This is what allows the usual account to avoid triviality.

We propose to apply this idea when we focus on sets of precisifcations themselves, as we did in the case of credences. Section 8.7 considers an alternative jump, $\mathsf{closure} \circ \mathcal{R}$, which explicitly adds additional precisifications that are 'limits' of the precisifications obtained by revising each member of the set ($\mathcal{R}$). Formally, we take a topological closure. This will allow triviality to be avoided. It is moreover important to note that any focus on particular definite judgements for credences will not allow the triviality to be avoided in the way that it did for truth.

Section 8.8 proposes an account of when an imprecise credal state is non-undermining: when $\mathcal{R}(\mathbb{C}) \subseteq \mathbb{C} \subseteq \mathsf{closure}(\mathcal{R}(\mathbb{C}))$. Since there is always some credal state which is a fixed-point of $\mathsf{closure} \circ \mathcal{R}$, there is always some credal state which is non-undermining in this sense. We thus propose that such credal states are candidates for being rational attitudes to adopt.

Section 8.9 demonstrates that the account we have provided is in fact very general and could apply to a whole range of target notions, one just needs to spell out a range of objects that play the role of the precisifications, and to describe how to revise each of these.

8.2 Truth

We first briefly present the usual account for truth.

8.2.1 Classical, Precise Truth

Setup 8.2.1. Let $\mathcal{L}$ be a base language in which we have the ability to code sentences, for ease we might take this to be the language of Peano Arithmetic. Let $\mathcal{L}_T$ extend this with the addition of a unary predicate, T. Sent_T denotes the sentences of this language. We will assume we have a fixed model of our base language, which we assume is the standard model of arithmetic, denoted $\mathbb{N}$.

Definition 8.2.2. A *precise interpretation of truth*, Q, is given by a collection of sentences, i.e., $Q \in \mathcal{P}(\mathsf{Sent}_T)$. The collection of all precise interpretations of truth is $\mathsf{AllPrecsT} = \mathcal{P}(\mathsf{Sent}_T)$.

The collection Q gives the collection of true sentences. We could equivalently think of it as assigning a truth value, true or not-true, to each sentence.

Setup 8.2.3. $(\mathbb{N}, Q)$ refers to the classical model of $\mathcal{L}_T$ resulting from expanding the standard model of arithmetic for the base language, $\mathbb{N}$, with Q providing the sentences whose codes are in the extension of the truth predicate. So we have $(\mathbb{N}, Q) \models T\ulcorner\varphi\urcorner$ iff $\varphi \in Q$.

Definition 8.2.4. The *Tarskian revision jump*, τ, is a function τ: $\mathsf{AllPrecsT} \to \mathsf{AllPrecsT}$ with $\tau(Q)$ given by

$$\varphi \in \tau(Q) \text{ iff } (\mathbb{N}, Q) \models \varphi.$$

Note that we have $\varphi \in Q$ iff $T\ulcorner\varphi\urcorner \in \tau(Q)$.

The function τ can be understood as a stage of reflecting on the supposed truth values. To be materially adequate, an interpretation of truth should be a fixed point of τ, i.e., Q such that $Q = \tau(Q)$. But the liar paradox shows us that this is not possible because for the liar sentence, Liar, which is equivalent to $\neg T\ulcorner\text{Liar}\urcorner$, we have that Liar $\in Q$ iff Liar $\notin \tau(Q)$. For an analogy with credences, we might say that every precise interpretation of truth is thus undermining.

8.2.2 Definite Truth

We now consider (in)definite truth (McGee, 1989, 1990). Some sentences are definitely true, for example '0 = 0', some definitely not true, for example '0 ≠ 0', and some neither, for example Liar. A specification of which sentences are which is given by a definite verdict assignment:

Definition 8.2.5. A *definite verdict assignment*, S, is given by two sets of sentences, S^+ and S^-.

S^+ contains the sentences that are definitely true, and S^- the sentences that are definitely not true.[2] Some sentences may be neither definitely true nor definitely not true.[3]

Associated with any definite verdict assignment is a collection of precise interpretations of truth, its "precisifications":

Definition 8.2.6. $Q \in \mathsf{AllPrecsT}$ is a *precisification* of $S = (S^+, S^-)$ iff

- If $\varphi \in S^+$ then $\varphi \in Q$.
- If $\varphi \in S^-$ then $\varphi \notin Q$.

We call the collection of precisifications of S, $\mathsf{Precs}(S)$.

That is, Q is a precisification if it agrees with the definite verdicts given by S: any sentence that S assigns as definitely true should be true in Q, and any sentence assigned as definitely not true should be not true in Q.

One might also consider adding "admissibility conditions" which restrict the precisifications to, for example, those that are maximally consistent. A more natural implementation of this in our framework is just to restrict $\mathsf{AllPrecsT}$ to such interpretations. For the purposes of this chapter, we will not consider such restrictions but it is easy to apply all our considerations with such restrictions, as we will mention in Section 8.9.

We can also consider a set of precisifications giving rise to a definite verdict assignment:

Definition 8.2.7. Given a set of precisifications, $\mathbb{Q} \subseteq \mathsf{AllPrecsT}$, $\mathsf{Def}(\mathbb{Q})$ is a definite verdict assignment given by:

- $\varphi \in \mathsf{Def}(\mathbb{Q})^+$ iff $\varphi \in Q$ for all $Q \in \mathbb{Q}$.
- $\varphi \in \mathsf{Def}(\mathbb{Q})^-$ iff $\varphi \notin Q$ for all $Q \in \mathbb{Q}$.

That is, if φ is determinately true in $\mathbb{Q}$, then it is assigned as definitely true by $\mathsf{Def}(\mathbb{Q})$, and if it is determinately not true in $\mathbb{Q}$ it is assigned as definitely not true by $\mathsf{Def}(\mathbb{Q})$.

The supervaluational Kripke jump revises a definite verdict assignment as follows:

Definition 8.2.8. $\Delta(S)$ is the definite verdict assignment given by:

- $\varphi \in \Delta(S)^+$ iff $(\mathbb{N}, Q) \models \varphi$ for all $Q \in \mathsf{Precs}(S)$.
- $\varphi \in \Delta(S)^-$ iff $(\mathbb{N}, Q) \models \varphi$ for all $Q \in \mathsf{Precs}(S)$.

Since $(\mathbb{N}, Q) \models \varphi$ iff $\varphi \in \tau(Q)$, we can give an alternative description of this: $\Delta(S)$ is the definite verdict assignment given by looking at the determinate judgements of the collection of precisifications $\tau(Q)$ for $Q \in \mathsf{Precs}(S)$. That's exactly what Def allowed us to state, so:

$$\Delta(S) = \mathsf{Def}(\{\tau(Q) | Q \in \mathsf{Precs}(S)\}).$$

We can further simplify this by making another definition.

Definition 8.2.9. For a set of precisifications $\mathbb{Q} \subseteq \mathsf{AllPrecsT}$,

$$\mathcal{R}_\tau(\mathbb{Q}) := \{\tau(Q) | Q \in \mathbb{Q}\}.$$

The operation $\mathcal{R}_\tau$ just revises each member of the set in accordance with τ, so $\{\tau(Q) \mid Q \in \mathsf{Precs}(S)\} = \mathcal{R}(\mathsf{Precs}(S))$. We then immediately have:

$$\Delta(S) = \mathsf{Def}(\mathcal{R}_\tau(\mathsf{Precs}(S))).$$

We can use $\circ$ to denote concatenation, so write this as

$$\Delta(S) = \mathsf{Def} \circ \mathcal{R}_\tau \circ \mathsf{Precs}(S).$$

We can describe this by the following procedure for obtaining $\Delta(S)$:

(i) Starting with a definite verdict assignment, use Precs to move to the corresponding set of precisifications.
(ii) Revise each of the precise interpretations in the set according to τ (i.e., apply $\mathcal{R}_\tau$).

(iii) Use Def to move from the resultant collection of revised precise interpretations to the definite verdicts.

There are fixed points of Δ: accounts of definite truth that are non-undermining. Moreover, there are non-trivial such fixed points.

8.3 Precise Credences

We now move to developing the analogous tools for the case of credences. We start with the classical, precise setting.

8.3.1 Credence Functions

We first start just by specifying the notion of credence function that we're working with in the classical setting.

Setup 8.3.1. We start with a non-empty set of sentences, $\mathcal{A}$, which we call our *agenda*.[4]

This could be all sentences of a given language, but it can also be more restrictive, for example we might consider cases where we are only looking at your credence in a single sentence, so where $\mathcal{A}$ is a singleton. For example we might just be interested in the credence that you'll forget your passport, so $\mathcal{A}$ might just contain the sentence saying that you'll forget your passport.[5]

Definition 8.3.2. A *credence function* on an agenda $\mathcal{A}$ is a function, c, from $\mathcal{A}$ to $[0, 1]$; i.e., it associates with each sentence in $\mathcal{A}$ a degree of belief, which is a real number between 0 and 1 inclusive. $\mathsf{Creds}_{\mathcal{A}}$ is the set of all credence functions, i.e., all functions from $\mathcal{A}$ to $[0, 1]$. If $\mathcal{A}$ just contains a single sentence, a credence function can be thought of simply as a value in $[0, 1]$ and we will call this a *credence value*.

It wouldn't affect our account if we restrict $\mathsf{Creds}_{\mathcal{A}}$ to just those functions that are finitely-additive probabilities (or, more carefully, which are extendable to functions on a Boolean algebra which satisfy the axioms of finitely-additive probability theory[6]), as this is still a compact space. However, we could not restrict attention to *countably*-additive probabilities.[7]

8.3.2 Revision of Precise Credences

The Tarskian revision jump, τ, was a way of revising precise interpretations of truth. For credences, we will suppose we have a revision function as given, and our supervaluational account will be a general one that can apply to any given revision function.[8] Formally:

Definition 8.3.3. A *revision function* is a function, ρ, from $\mathsf{Creds}_{\mathcal{A}}$ to $\mathsf{Creds}_{\mathcal{A}}$.

A precise credence function is undermining if $\rho(c) \neq c$.

Passport is a story which directly describes how one should revise one's credence in the target-proposition, that you'll remember your passport, under a step of reflection on the proposed credence value. If we let $\mathcal{A}$ simply contain this one proposition, then credence functions are just values $x \in [0, 1]$, and the revision function that Passport gives rise to is:

$$\rho_{\text{Passport}}(x) = \begin{cases} 1 & x < 0.5 \\ 0 & x \geq 0.5 \end{cases}$$

One can see that there are no fixed points of ρ_{Passport}. That is, every precise credence function undermines its own adoption. In this sense Passport can be related to the liar sentence: in both cases, all precise options are undermining.

There are other cases that also give rise to the same revision function as Passport, for example:

Bad Navigator

> You've come to a crossroads and are wondering whether you need to turn left or right to get to your hotel. You know you're a really bad navigator. In particular, you believe that if you have credence ≥ 0.5 that left is the way to your hotel, then it's actually right; and if not, then it's actually to the left. What should your credence be that it's actually right?
>
> (Extremal version of an example in Egan and Elga, 2005)

In this case, the change in one's credence isn't due to the causal structure, but instead simply that the credence that one adopts affects the evidence that one has about the situation. But the same revision function describes this case. The same revision function would also arise when considering the following self-referential sentence:

> CredLiar: Your credence in CredLiar is not ≥ 0.5.

In this case, the revision is due to semantic features of the sentence.

You might also be uncertain about whether you are in a Passport-like case:

Golf

> You think there's a small chance, 1%, that whether you'll be able to successfully get this hole-in-one is dependent on the credence you adopt in it in a Passport-style way, i.e., where if you have credence ≥ 0.5 then you'll fail, and if not then you'll succeed. But you're 99% sure that it's just a normal case and you have a 50% chance of success.

This leads to the revision function

$$\rho_{\textsc{Golf}}(x) = \begin{cases} 0.01 \times 1 + 0.99 \times 0.5 = 0.505 & x \geq 0.5 \\ 0.01 \times 0 + 0.99 \times 0.5 = 0.495 & x < 0.5 \end{cases}$$

It also might be that one's credence doesn't directly provide evidence about the truth of the sentence, but instead it affects the chances. Consider, for example, the following scenario discussed by Greaves (2013):

Promotion

> "Alice is up for promotion. Her boss, however, is a deeply insecure type: he is more likely to promote Alice if she comes across as lacking in confidence. Furthermore, Alice is useless at play-acting, so she will come across that way iff she really does have a low degree of belief that she's going to get the promotion. Specifically, the chance of her getting the promotion will be $1 - x$, where x is whatever degree of belief she chooses to have in the proposition P that she will be promoted. What credence in P is it epistemically rational for Alice to have?"
>
> (Greaves, 2013, pp. 1–2)

(Moreover Greaves assumes "that the agent is aware of the specification of ... her case".) If Alice considers adopting credence 0.2 in P; then the chance of P would be 0.8, and she knows that, so that would recommend adopting credence 0.8. More generally, the description of this case directly provides us with the revision function

$$\rho_{\textsc{Promotion}}(x) = 1 - x.$$

Unlike for the Passport revision function, this function does have a fixed point, 0.5.

All the cases we've seen so far are unusual cases. In normal cases, the credence one adopts provides no additional evidence about the situation at hand.

Rain

> The credence that you adopt that it is going to rain tomorrow provides no additional evidence about the likelihood of rain.

In this case, $\rho(x) = x$. More generally, in normal cases $\rho(c) = c$ for all c, or at least all c which are probabilistic.[9] And most theorising about rationality has focused on these "safe" cases.

This same revision function might also arise in a case where the credence one adopts does provide additional information:

Leap

> The chance you'll successfully leap across this chasm is identical to your credence.
>
> (Greaves, 2013, see also James, 1897)

We do not assume any further modelling of this revision function, we simply assume that any scenario gives rise to such a revision function. Any further modelling of ρ would have allow for the range of cases mentioned so far. It has to allow for logical, causal and evidential impact (as in CredLiar, Passport, and Bad Navigator), and this might go via chance (like Promotion) or be directly about the proposition (as in Passport), or be associated with further uncertainty. Our account does not depend on any further specifics of the revision function, we can simply take it as an input to our account. The revision function should encode the idea of reflecting on one's credences, and we are adopting the idea that to be rational, a precise credence function should be a fixed point of this revision function.[10] Otherwise a credence function is undermining.

There are two suggestions for how one could include further modelling or explanation of the revision function.

Firstly, one might simply take $\rho(c)$ to be c conditionalised on "c is my credence function". If we assume that one's initial credence function satisfies certain other constraints of rationality such as deferring to chances (by satisfying the so-called Principal Principle) this should lead to the revision functions proposed. The idea of this is to result in notions like those of Joyce (2018) or Konek and Levinstein (2019), as opposed to the "consequentialist" recommendation notions of, for example, Pettigrew (2018) and Caie (2013).[11]

Alternatively, one might want to explicitly include a possible worlds structure in the modelling and then define ρ using this. This is particularly natural for accounting for a language with sentences that can talk about the credence one has in that very sentence, i.e., cases like PrLiar. This kind of picture has commonly been used when developing accounts for languages with modal predicates (Halbach et al., 2003; Stern, 2015; Campbell-Moore, 2015; Halbach and Welch, 2009; Nicolai, 2018). Given a fixed possible world structure, with various worlds and probabilistic accessibility relations between them, we consider a precise interpretation to be given by an assignment of a credence function at each world: a credal-evaluation-function. We can then directly define the revision of a credal-evaluation-function by taking the weighted proportion of the accessible worlds where the sentence is evaluated as true when the initial credal-evaluation-function provides the interpretation of the credence function symbol at the various worlds.[12]

There are two reasons to focus simply on credence functions rather than using a possible world structure. Firstly, it is simpler and all our considerations will immediately apply to the more general setting (Section 8.9).

Secondly, it is not directly obvious how to provide a possible worlds model for the cases like Promotion where the impact goes via chances. And insofar as it deals with Passport or Bad Navigator it just treats them like CredLiar, for example, we wouldn't represent 'the hotel is to the left' as an atomic sentence, as would be most natural, but instead as a sentence that refers to itself. Whilst this leads to the right revision function, it does not seem to be the right analysis of the sentence itself. We thus find it valuable to not encode further modelling such as this, but to simply provide the account for any specified revision function; though in section 8.9 we note how our account immediately applies to the possible worlds setting.

8.4 How to Model Indeterminate Credences

In the case of truth we focused on definite truth value verdicts. When considering indeterminate credences, what should we think about? We suggest that we directly work with a set of precise credence functions, that is, we consider one's indeterminate credal state to be given by a non-empty set $\mathbb{C} \subseteq \mathsf{Creds}_{\mathcal{A}}$. The precise credence functions in the set will be called the 'precisifications' of the indeterminate credal state. So, for example, if our agenda contains a single proposition, a precise credence function is some real number between 0 and 1, whereas an indeterminate credal state is given by a set of numbers, e.g., $\{0.2, 0.3\}$, or $[0.2, 0.3]$. It remains indefinite which of the credence values in the set it is, but it is, for example, definitely not 0.9.

This model of belief is closely related to one that is familiar in formal epistemology under the term 'imprecise probabilities', 'indeterminate probabilities', or 'mushy credences'.[13] It has been proposed for a range of reasons, including being able to represent incomparability as distinct from indifference, distinguishing between lack of evidence and symmetric evidence, allowing for suspension of judgement, and rationalising intuitively rational responses to certain decision problems (Joyce, 2010; Bradley, 2015; Levi, 1978; Jeffrey, 1984). Its interpretation is debated.

To more closely match the application to truth, we might instead identify some particular judgements and ask whether they definitely hold, definitely don't hold, or neither. For example, we might only care about whether one's credence in φ is definitely equal to r, definitely not equal to r, or neither. But the account will be then be very weak. For example, a case like Golf would not get assigned a credence value at all. This thus doesn't respect the fact that your credence should definitely be ≥ 0.3, which we would get out of the more expressive framework when we look directly at the set of precisifications. Furthermore, this is a difference that might be used in decision making. One might instead then try to be more expansive about the kinds of definite verdicts that are being considered. We might consider whether your credence is definitely $\geq r$ or definitely not $\geq r$. Again, this can be criticised for leaving out potentially definite judgements such as that φ is more likely than ψ, or that φ is evidence for ψ, or that I'm certain in at least one of $\varphi_1, \varphi_2, \dots$.

By focusing on sets of precisifications themselves, however, every definite judgement is encoded. Any $\mathbb{B} \subseteq \mathsf{Creds}$ can be thought of as a property of one's credences, for example, $\mathbb{B} = \{c \mid c(\varphi) > c(\psi)\}$ is the property that you think φ is more likely than ψ; $\mathbb{B} = \{c \mid c(\varphi \mid \psi) > c(\varphi)\}$ is the property that you take ψ to be evidence for φ; and $\mathbb{B} = \{c \mid c(\varphi_k) = 1 \text{ for some } k)\}$ is the property that you are certain of at least one of $\varphi_1, \varphi_2, \dots$. For any set of precisifications $\mathbb{C}$, we can say whether it definitely satisfies that property, definitely doesn't, or neither, by considering whether $\mathbb{C} \supseteq \mathbb{B}$, $\mathbb{C} \bigcap \mathbb{B} = \emptyset$, or neither. Focusing on the set of precisifications, $\mathbb{C}$, itself is equivalent to focusing on definite judgements on *all* properties, at least when we ignore any differences in definite judgement assignments that don't correspond to differences in resultant precisifications.[14] And since differences that don't constitute differences in precisifications will not affect the supervaluational jump, considering sets of precisifications themselves is the most general model available for our purposes. Focusing on any particular definite judgements can then be considered as special cases of our general account. Unlike for truth, though, the kinds of triviality issues we face when working with sets of precise credences will often arise for these other models, at least whenever one is interested both in whether a property is definitely satisfied and whether it is definitely not satisfied. (See Section 8.9 for further discussion.)

8.5 The Jump $\mathcal{R}$ Revising Indeterminate Credence

8.5.1 $\mathcal{R}$ Applied to Credences

We now turn to revision of one's indeterminate credences. Recall our presentation of the supervaluational Kripkean jump for truth as:

$$\Delta(S) = \mathsf{Def} \circ \mathcal{R}_\tau \circ \mathsf{Precs}\,(S).$$

We described this with the following procedure: (i) use Precs to move from a definite verdict assignment to the corresponding set of precisifications; (ii) use the Tarskian revision function, τ, to revise each of these ($\mathcal{R}_\tau$); and (iii) use Def to move from the resultant set of revised precise interpretations back to the define verdicts assignment.

I suggest that what motivates this definition of Δ is just the revision of each of the members of the set, $\mathcal{R}_\tau$, but since Δ is defined on the definite verdicts model rather than sets of precisifications, we have to also introduce stages (i) and (iii) to find $\mathcal{R}_\tau$'s treatment of definite verdicts.

In the case of credences, however, we have suggested working directly with a set of precisifications, and want to know how to revise that. So stages (i) and (iii) aren't needed and we might suggest that the analogous way to revise a set of precisifications is just to revise each of the precisifications, i.e., just apply stage (ii). We defined $\mathcal{R}_\tau$ for the case of truth as $\mathcal{R}_\tau(\mathbb{Q}) := \{\tau(Q) \mid Q \in \mathbb{Q}\}$. We now simply present this as a more general definition that can apply to any revision function:

Definition 8.5.1. For a given (fixed) revision function ρ,

$$\mathcal{R}_\rho(\mathbb{C}) := \{\rho(c) \mid c \in \mathbb{C}\}.$$

We will generally drop the subscript as it's typically clear which revision function is used.

This simply takes the collection of revised individuals. It is a very intuitive notion of revision applied to a indeterminate credence, understood as a set of precisifications. It also seems to follow naturally from the supervaluationist idea that what happens on the supervaluational-side supervenes on what happens on the precise side.

Consider Passport, Bad Navigator or CredLiar. We simply focus on an agenda consisting of the single sentence at stake in each of these scenarios, so credence functions are given by real numbers between 0 and 1. Supervaluational credences are given by sets of precise credences, so this will be a set of real numbers between 0 and 1. Consider adopting the set consisting just of the two extremal credences, $\{0, 1\}$. Revision of precise credences is spelled out by ρ_{Passport}. In particular, credence 0 recommends adopting credence 1; and 1 recommends 0. When we apply $\mathcal{R}$ to the set $\{0, 1\}$ we just revise each member, so we have $\mathcal{R}(\{0, 1\}) = \{\rho(0), \rho(1)\} = \{1, 0\} = \{0, 1\}$. The indeterminate credal state {0, 1} is a fixed point of $\mathcal{R}$. Whilst each precisification is undermining, the set, as a whole, is a non-undermining attitude to adopt in these cases.

We might consider in general saying that an imprecise credal state is non-undermining iff it's a fixed point of $\mathcal{R}$. However, there is a formal issue facing this proposal: $\mathcal{R}$ may not have any (non-trivial) fixed points. Thus, if our notion of underminingness is spelled out just with $\mathcal{R}$, we still might end up with a situation where every credal state, precise or imprecise, is undermining, and thus not a candidate for the rational response to the situation. Ultimately, we will suggest an alternative notion of underminingness which will always allow for a non-undermining response.

8.5.2 $\mathcal{R}$ Doesn't Always Have a Fixed Point

Consider the following kind of scenario:

> Spring
>
> You know that you're always overconfident in this type of situation. Except you also know that a credence value of 0 would be wrong.

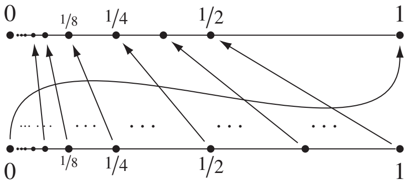


Figure 8.1 Illustration of ρ_{SPRING}

What revision function does this lead to? It will have that $\rho(0) > 0$, and for all $x > 0$, $\rho(x) < x$. To say more about it, though, we need further details about this case: ‘how overconfident?’ ‘how wrong?’. In fact, I think natural ways of adding to this story will not guarantee a notion of a particular credence value being recommended, instead it might allow for ties. But for simplicity, this chapter focuses on the case where we have a fully specified revision function.[15] In fact it doesn’t matter how we spell it out, any revision function with these properties leads to a $\mathcal{R}$ which has no fixed points. To work with a concrete example, we suppose that additional details are added to the case so that we obtain the following revision function:[16]

$$\rho_{\text{SPRING}}(x) = \begin{cases} 1 & x = 0 \\ \dfrac{x}{2} & x > 0 \end{cases}$$

See Figure 8.1 for an illustration.

As in the cases like PASSPORT, every credence value is undermining. But, unlike in PASSPORT, there is also no indeterminate credal state which is a fixed point of $\mathcal{R}$. Even though $\mathcal{R}$ is monotone, we result in the empty set of precisifications, which is not a legitimate imprecise opinion state.

PROPOSITION 8.5.2. There is no (non-empty) fixed point of $\mathcal{R}_{\text{SPRING}}$.

Proof. We will first observe that for any $\mathbb{C}$ and $n \geq 1$, any $x \in \mathcal{R}^n(\mathbb{C})$ has $0 < x \leq \frac{1}{2^{n-1}}$. (See Figure 8.2.) Recall that $x \in \mathcal{R}^n(\mathbb{C})$ iff there is $y \in \mathcal{R}^{n-1}(\mathbb{C})$ with $x = \rho(y)$. So we equivalently need to show that any $x \in \mathcal{R}^n(\mathbb{C})$ has $0 < \rho(x) \leq {}^1\!/_{2^n}$.

Base case: For any $x > 0$, $\rho(x) = x/_2 > 0$ (and $\rho(x) \leq 1$). Also $\rho(0) = 1 > 0$. Thus, any $x \in \mathbb{C} \subseteq [0, 1]$ has $0 < \rho(x) \leq 1$, as required.

Inductive step: For any $x \in \mathcal{R}^n(\mathbb{C})$, $0 < x \leq {}^1\!/_{2^{n-1}}$. So $\rho(x)$, which $= x/_2$ has $0 < \rho(x) \leq \frac{{}^1\!/_{2^{n-1}}}{2} = {}^1\!/_{2^n}$, as required.

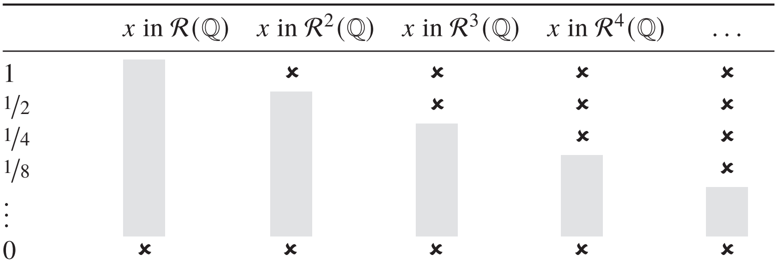

x in $\mathcal{R}(\mathbb{Q})$
x in $\mathcal{R}^2(\mathbb{Q})$
x in $\mathcal{R}^3(\mathbb{Q})$
x in $\mathcal{R}^4(\mathbb{Q})$
...
1
1/2
1/4
1/8
⋮
0

Figure 8.2 Illustration of $\mathcal{R}$ with Spring. x can lie only in the gap between the crosses

Now, suppose $\mathbb{C} = \mathcal{R}(\mathbb{C})$. Then $\mathbb{C} = \mathcal{R}^n(\mathbb{C})$ for all n. So any $x \in \mathbb{C}$ has $x \leq {}^1\!/_{2^{n-1}}$ for all n. But the only such x is 0, and we also require that $x > 0$. So $\mathbb{C} = \varnothing$. □

So, this supervaluational jump does not guarantee that undermining credal states can be avoided. So we shouldn't use this alone to characterise underminingness.

8.6 Supervaluational Kripkean Jump for Truth as It Applies to Sets of Precisifications

In order to develop a notion of underminingness which will always allow for non-undermining credal states, we first consider how the supervaluational Kripkean jump for truth avoids triviality. This will lead us to consider an alternative jump, $\mathsf{closure} \circ \mathcal{R}$. Both jumps, $\mathcal{R}$ and $\mathsf{closure} \circ \mathcal{R}$, will be used in the characterisation of underminingness.

8.6.1 $\mathcal{R}_\tau$ Has No Fixed Points

Some revision functions, such as that for Passport, do lead to fixed points of $\mathcal{R}$, whereas the revision function for Spring rules them out. What about the Tarskian revision function for truth, τ? Does just revising a set of precise interpretations of truth by revising each according to τ ($\mathcal{R}_\tau$) have fixed point? It does not.[17] To show this, we note that the McGee sentence leads to Spring-style phenomena. The McGee sentence, McGee, is given by:

> McGee: Some truth iteration of McGee is not true.

Or, more formally, where

$$\mathsf{McGee} \text{ is equivalent to } \neg\forall k > 0\,\overbrace{T\ulcorner T\ldots\ulcorner T}^{k}\,\ulcorner\mathsf{McGee}\urcorner\urcorner\ldots\urcorner.$$

	Q in $\mathcal{R}(\mathbb{Q})$	Q in $\mathcal{R}^2(\mathbb{Q})$	Q in $\mathcal{R}^3(\mathbb{Q})$	Q in $\mathcal{R}^4(\mathbb{Q})$	$\cdots$
McGee		true	true	true	
$T\ulcorner$McGee$\urcorner$	some not-true		true	true	
$T^2\ulcorner$McGee$\urcorner$		some not-true		true	
$T^3\ulcorner$McGee$\urcorner$			some not-true		
$\vdots$				some not	

Figure 8.3 Illustration of $\mathcal{R}$ with the McGee sentence.

One can then use this to show:

Proposition 8.6.1 (See also Halbach, 2014, Theorem 14.11). There is no (non-empty) fixed point of $\mathcal{R}_\tau$.

Proof. We will first observe that for any $\mathbb{Q}$ and $n \geq 1$, any $Q \in \mathcal{R}^n(\mathbb{Q})$ has $T^i\ulcorner\mathsf{McGee}\urcorner \in Q$ for all $0 \leq i < n-1$, but also has some k with $T^k\ulcorner\mathsf{McGee}\urcorner \notin Q$. (See Figure 8.3.) That is, any $Q \in \mathcal{R}^n(\mathbb{Q})$ has $T^i\ulcorner\mathsf{McGee}\urcorner \in \tau(Q)$ for all $0 \leq i < n$, but also some $T^k\ulcorner\mathsf{McGee}\urcorner \notin \tau(Q)$.

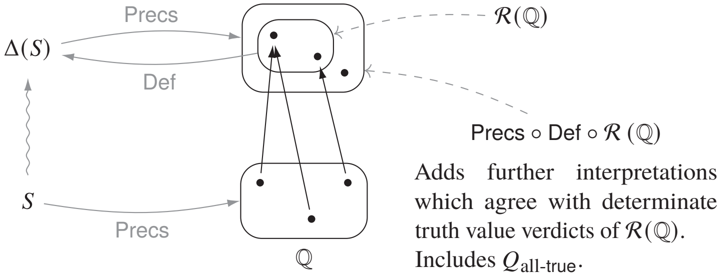
Base case: We just need to show that for any $Q \in \mathbb{Q}$, some $T^k\ulcorner\mathsf{McGee}\urcorner \notin \tau(Q)$. Since we always have $\varphi \in Q$ iff $T\ulcorner\varphi\urcorner \in \tau(Q)$, if $T^k\ulcorner\mathsf{McGee}\urcorner \notin Q$, then $T^{k+1}\ulcorner\mathsf{McGee}\urcorner \notin \tau(Q)$. So we just need to consider Q where there is no such k. For such Q, $(\mathbb{N}, Q) \not\models \mathsf{McGee}$, so $\mathsf{McGee} \notin \tau(Q)$ giving us our $k = 0$.

Inductive step: by our inductive hypothesis we have that any $Q \in \mathcal{R}^n(\mathbb{Q})$ has $T^i\ulcorner\mathsf{McGee}\urcorner \in Q$ for all $0 \leq i < n-1$, but also has some $T^k\ulcorner\mathsf{McGee}\urcorner \notin Q$. So, since $T\ulcorner\varphi\urcorner \in \tau(Q)$ iff $\varphi \in Q$, we have $T^{i+1}\ulcorner\mathsf{McGee}\urcorner \in Q$ for all $0 \leq i < n-1$ and $T^{k+1}\ulcorner\mathsf{McGee}\urcorner \notin Q$. Thus $T^i\ulcorner\mathsf{McGee}\urcorner \in \tau(Q)$ for all $1 \leq i < n$ and some $T^k\ulcorner\mathsf{McGee}\urcorner \notin \tau(Q)$. Also $(\mathbb{N}, Q) \not\models \mathsf{McGee}$, so $\mathsf{McGee} \in \tau(Q)$ as required for $i = 0$.

Now, suppose $\mathbb{Q} = \mathcal{R}(\mathbb{Q})$. Then also $\mathbb{Q} = \mathcal{R}^n(\mathbb{Q})$ for all n. So any $Q \in \mathbb{Q}$ has $T^n\ulcorner\mathsf{McGee}\urcorner \in Q$ for all n. But also $T^k\ulcorner\mathsf{McGee}\urcorner \notin Q$ for some k. So $\mathbb{Q} = \varnothing$. □

8.6.2 How Δ Acts on Sets of Precisifications

We have thus seen that like for our revision function for the Spring case, the Tarskian revision function for truth means that revising a set of precisifications by just revising each of the members leads to triviality. However, the usual supervaluational Kripke jump for truth, Δ, does not act on sets of precisifications simply by revising each member of the set (i.e., $\mathcal{R}$) instead it acts in accordance with $\mathsf{Precs} \circ \mathsf{Def} \circ \mathcal{R}$: If $\mathsf{Precs}(S) = \mathbb{Q}$, then $\Delta(S) = \mathsf{Def} \circ \mathcal{R}\,(\mathbb{Q})$, so $\mathsf{Precs}(\Delta(S)) = \mathsf{Precs} \circ \mathsf{Def} \circ \mathcal{R}\,(\mathbb{Q})$. See Figure 8.4.


Precs
$\Delta(S)$
Def
$\mathcal{R}(\mathbb{Q})$
Precs $\circ$ Def $\circ$ $\mathcal{R}$ $(\mathbb{Q})$
S
Precs
$\mathbb{Q}$

Adds further interpretations which agree with determinate truth value verdicts of $\mathcal{R}(\mathbb{Q})$. Includes $Q_{\text{all-true}}$.

Figure 8.4 Revising a definite verdict assignment and revising the set of precisifications

By applying $\mathsf{Precs} \circ \mathsf{Def}$ to $\mathcal{R}(\mathbb{Q})$, additional precisifications are added. In particular any additional precisifications that agree with any of the determinate truth value verdicts, i.e., those truth value verdicts which are unanimously agreed on. More carefully: $Q^* \in \mathsf{Precs} \circ \mathsf{Def}\,(\mathbb{Q})$ iff for any φ,

- if $\varphi \in Q$ for all $Q \in \mathbb{Q}$ then $\varphi \in Q^*$, and
- if $\varphi \notin Q$ for all $Q \in \mathbb{Q}$ then $\varphi \notin Q^*$.

Any further relationships between sentences do not need to be respected by all Q^* in $\mathsf{Precs} \circ \mathsf{Def}\,(\mathbb{Q})$. For example, even if every $Q \in \mathbb{Q}$ has at least one of φ or ψ true, if both are indeterminate we can have some $Q^* \in \mathsf{Precs} \circ \mathsf{Def}\,(\mathbb{Q})$ with both φ and ψ not true. This means, for example, that even though $\mathcal{R}(\mathbb{Q})$ only contains maximally consistent precise interpretations, $\mathsf{Precs} \circ \mathsf{Def} \circ \mathcal{R}\,(\mathbb{Q})$ can contain inconsistent precise interpretations.

Similarly, each $Q \in \mathcal{R}(\mathbb{Q})$ puts at least one of $T^k\ulcorner\mathsf{McGee}\urcorner$ as not-true. However, since each $T^k\ulcorner\mathsf{McGee}\urcorner$ is indeterminate according to $\mathcal{R}(\mathbb{Q})$, so we can find some $Q_{\text{all-true}}$ in $\mathsf{Precs} \circ \mathsf{Def} \circ \mathcal{R}\,(\mathbb{Q})$ which agrees with any determinate truth value verdicts of $\mathcal{R}(\mathbb{Q})$ but which puts all $T^k\ulcorner\mathsf{McGee}\urcorner$ as true. This is what allows Δ to have fixed points, where $\mathcal{R}$ does not.

Proposition 8.6.2. There is some $Q_{\text{all-true}}$ in $\mathsf{Precs} \circ \mathsf{Def} \circ \mathcal{R}\,(\mathsf{AllPrecsT})$ such that we have $T^k\ulcorner\mathsf{McGee}\urcorner \in Q_{\text{all-true}}$ for all k. There is no such $Q_{\text{all-true}}$ in $\mathcal{R}(\mathsf{AllPrecsT})$.

Proof. We argued in 8.6.1 that there is no such $Q_{\text{all-true}}$ in $\mathcal{R}(\mathsf{AllPrecsT})$. We need to show there is some such in $\mathsf{Precs} \circ \mathsf{Def} \circ \mathcal{R}\,(\mathsf{AllPrecsT})$.

Consider any $Q_0 \in \mathsf{AllPrecsT}$. Set $Q_n = \tau^n(Q_0)$. Note that $Q_n \in \mathcal{R}^n(\mathsf{AllPrecsT})$, and also that $Q_n \in \mathcal{R}(\mathsf{AllPrecsT})$.

Define $Q_{\text{all-true}}$ by $\varphi \in Q_{\text{all-true}}$ iff φ is stably true in $\langle Q_n \rangle$, that is there is some k with $\varphi \in Q_n$ for all $n > k$. Since every $T^n \ulcorner \mathsf{McGee} \urcorner$ is stable true, it is in $Q_{\text{all-true}}$.[18]

We can then show that $Q_{\text{all-true}} \in \mathsf{Precs} \circ \mathsf{Def} \circ \mathcal{R}\,(\mathsf{AllPrecsT})$. Note that we have $\{Q_1, Q_2, \ldots\} \subseteq \mathcal{R}(\mathsf{AllPrecsT})$. If $\varphi \in Q$ for all $Q \in \mathcal{R}(\mathsf{AllPrecsT})$, then $\varphi \in Q_n$ for all $n > 0$, so $\varphi \in Q_{\text{all-true}}$. If $\varphi \notin Q$ for all $Q \in \mathcal{R}(\mathsf{AllPrecsT})$, then $\varphi \notin Q_n$ for any $n > 0$, so $\varphi \notin Q_{\text{all-true}}$. Thus, since $Q_{\text{all-true}}$ agrees with any determinate truth value verdicts of $\mathcal{R}(\mathsf{AllPrecsT})$, it is in $\mathsf{Precs} \circ \mathsf{Def} \circ \mathcal{R}\,(\mathsf{AllPrecsT})$ (this is an immediate consequence of the definitions of Precs and Def). □

8.7 An Alternative Jump for Indeterminate Credence

How can we take these insights and apply them to credence, where we are working directly with the set of precisifications model? When viewed through the lenses of sets of precisifications, the usual supervaluational jump for truth, Δ, adds additional precisifications: it corresponds to $\mathsf{Precs} \circ \mathsf{Def} \circ \mathcal{R}$ rather than $\mathcal{R}$. We similarly propose an alternative jump for imprecise credences, $\mathsf{closure} \circ \mathcal{R}$, which also adds additional precisifications. Which ones? For truth, $\mathsf{Precs} \circ \mathsf{Def} \circ \mathcal{R}$ adds to $\mathcal{R}$ any precise interpretations which agree on any truth value verdicts that are unanimously agreed on by all $Q \in \mathcal{R}(\mathbb{Q})$. For example, in the case of McGee it adds some $Q_{\text{all-true}}$. For credences, we will instead directly use underlying structure of the real numbers, and take a topological closure.

That is, we will directly consider the jump given by first revising each member of the set, and then taking the closure of the resultant set, where the notion of closure is given as follows:

Definition 8.7.1. $c^* \in \mathsf{closure}(\mathbb{C})$ iff there is a sequence, $\langle c_\alpha \rangle$, (not necessarily following the revision function) with each $c_\alpha \in \mathbb{C}$ and where $\langle c_\alpha \rangle$ converges to c^*, i.e., where for all $\varphi \in \mathcal{A}$ and for all $\epsilon > 0$, there is some β such that for all $\alpha > \beta$, $|c_\alpha(\varphi) - c^*(\varphi)| < \epsilon$.[19]

To see how these definitions work, consider how they apply in the case of Spring (see Figure 8.5).

$\mathcal{R}$ just revises each member of the set, so recalling the revision function for this case as spelled out in Section 8.5.2, we have

$$\mathcal{R}(\{0, 1, {}^1/_2, {}^1/_4, \ldots\}) = \{1, {}^1/_2, {}^1/_4, {}^1/_8, \ldots\}.$$

The sequence $\langle 1, {}^1/_2, {}^1/_4, \ldots \rangle$ converges to 0, and each member of the sequence is a member of $\mathcal{R}(\{0, 1, {}^1/_2, \ldots\})$. So when we take the closure of this set we will add 0, resulting in:

$$\begin{aligned} \mathsf{closure} \circ \mathcal{R}\,(\{0, 1, {}^1/_2, {}^1/_4, \ldots\}) &= \mathsf{closure}(\{1, {}^1/_2, {}^1/_4, {}^1/_8, \ldots\}) \\ &= \{0, 1, {}^1/_2, {}^1/_4, {}^1/_8, \ldots\}, \end{aligned}$$

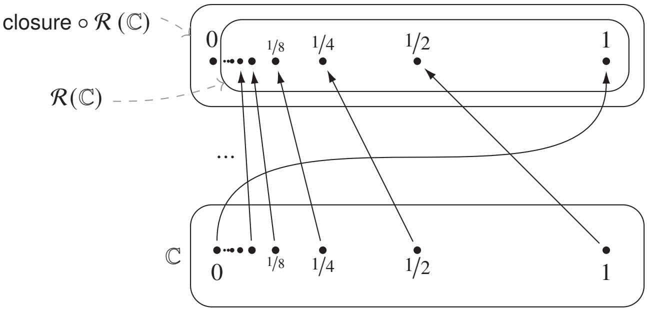

closure ∘ $\mathcal{R}$ ($\mathbb{C}$)
$\mathcal{R}$($\mathbb{C}$)
0
1/8
1/4
1/2
1
...
$\mathbb{C}$
0
1/8
1/4
1/2
1

Figure 8.5 By including the additional credence, 0, we find a fixed point in the case of Spring

This set is a fixed point of $\mathsf{closure} \circ \mathcal{R}$ in the Spring case.

In order to show our general result that $\mathsf{closure} \circ \mathcal{R}$ always has fixed points, we need to say a bit more about this notion of closure.

A space, in this case $\mathsf{Creds}_{\mathcal{A}}$, with a notion of closure gives us a topology.[20] In fact, the notion of closure defined in Definition 8.7.1 gives us the so-called topology of pointwise convergence, with the underlying topology on the real numbers being the standard one. There is an important property of this topology: it is *compact*. This is the property that allows us to show that $\mathsf{closure} \circ \mathcal{R}$ has fixed points.

We can now move to defining compactness (which we state in the form that we need for our main result[21]).

Definition 8.7.2. $\mathbb{C}$ is *closed* if $\mathbb{C} = \mathsf{closure}(\mathbb{C})$.

Definition 8.7.3. A space with a notion of closure is *compact* iff whenever $\mathcal{C}$ is a collection of *closed* sets which has the finite intersection property,

> i.e., for any finite sub-collection $C_1, \ldots, C_k \in \mathcal{C}$, $C_1 \cap \ldots \cap C_k \neq \emptyset$ then $\bigcap \mathcal{C} \neq \emptyset$.

Proposition 8.7.4. $\mathsf{Creds}_{\mathcal{A}}$ (with the specified notion of closure) is compact.

Proof Idea. Note that [0, 1] is compact as it is a closed and bounded subset of $\mathbb{R}$. Tychonoff's theorem (see, e.g., Willard, 1970, Theorem 17.8) says that the product of compact spaces is compact. We gave the notion of closure which corresponds to the topology of pointwise convergence, which is just the product topology $[0,1]^{\mathcal{A}}$, and is therefore compact. □

There are some other properties that are equivalent to compactness: that every convergent sequence has a limit point, or that every sequence whatsoever has a cluster point.[22] If we call a collection of sets consistent if it has some common member, i.e., has non-empty intersection, then we can describe this as: any collection of *closed* sets which is finitely consistent is consistent. Since we are looking for fixed points of $\mathsf{closure} \circ \mathcal{R}$ rather than $\mathcal{R}$ we can focus just on closed properties, as is done in the topological definition of compactness.

This allows us to show our main result:

THEOREM 8.7.5. For any ρ, there is a non-empty fixed point of $\mathsf{closure} \circ \mathcal{R}$.

Proof. Define a sequence

- $\mathbb{C}_0 := \mathsf{Creds}_{\mathcal{A}}$,
- $\mathbb{C}_{\alpha+1} := \mathsf{closure} \circ \mathcal{R}\,(\mathbb{C}_\alpha)$,
- $\mathbb{C}_\mu := \bigcap_{\alpha<\mu} \mathbb{C}_\alpha$.

$\mathsf{closure} \circ \mathcal{R}$ is monotone, that is:

SUBLEMMA 8.7.5.1. If $\mathbb{C} \supseteq \mathbb{C}'$ then $\mathsf{closure} \circ \mathcal{R}\,(\mathbb{C}) \supseteq \mathsf{closure} \circ \mathcal{R}\,(\mathbb{C}')$.

Proof. It is easy to observe that $\mathcal{R}$ is monotone, that is, if $\mathbb{C} \supseteq \mathbb{C}'$ then $\mathcal{R}(\mathbb{C}) \supseteq \mathcal{R}(\mathbb{C}')$.

Also, $\mathsf{closure}$ is monotone: Suppose $\mathbb{C} \supseteq \mathbb{C}'$. For any, $c^* \in \mathsf{closure}(\mathbb{C}')$, there is a sequence $\langle c_\alpha \rangle$ in $\mathbb{C}'$ which convergs to c^*. This sequence is also a sequence in any $\mathbb{C} \supseteq \mathbb{C}'$. So $c^* \in \mathsf{closure}(\mathbb{C})$.

And thus, $\mathsf{closure} \circ \mathcal{R}$, which is the result of composing these, is also monotone. □

So, by starting with $\mathbb{C}_0 = \mathsf{Creds}_{\mathcal{A}}$, where we have $\mathbb{C}_0 \supseteq \mathsf{closure} \circ \mathcal{R}\,(\mathbb{C}_0)$, we have that for $\alpha < \beta$, $\mathbb{C}_\alpha \supseteq \mathbb{C}_\beta$; and there must be a (possibly empty) fixed point of $\mathsf{closure} \circ \mathcal{R}$.

We need to check that this fixed point is non-empty, which we do by induction.[23]

- Base case: $\mathbb{C}_0 = \mathsf{Creds}_{\mathcal{A}} \neq \varnothing$.
- Successor case: For any $c \in \mathbb{C}_\alpha$, $\rho(c) \in \mathcal{R}(\mathbb{C}_\alpha)$, and since $\mathsf{closure}(\mathbb{C}') \supseteq \mathbb{C}'$ for any $\mathbb{C}'$ (as the constant sequence $\langle c', c', \ldots \rangle$ converges to c'), also $\rho(c) \in \mathsf{closure}(\mathcal{R}(\mathbb{C}_\alpha)) = \mathbb{C}_{\alpha+1}$.
- Limit case: Suppose each $\mathbb{C}_\alpha \neq \varnothing$ for $\alpha < \mu$. $\{\mathbb{C}_\alpha \mid \alpha < \mu\}$ is a collection of closed subsets of Creds. For $\alpha < \beta$, $\mathbb{C}_\alpha \supseteq \mathbb{C}_\beta$, so any finite subcollection has a non-empty intersection. Thus, by definition 8.7.3, $\mathbb{C}_\mu \neq \varnothing$. □

We will now spell out a notion of underminingness applied to imprecise credences, and use this result to show that there are always some non-undermining credal states.

8.8 Characterising Underminingness

We had originally considered suggesting that an imprecise credal state is non-undermining iff it is a fixed point of $\mathcal{R}$. But with this definition, in the case of SPRING, not only are all precise credence functions undermining, also all the imprecise credences are too. However, we have now considered also taking the closure of $\mathcal{R}$, and seen that $\mathsf{closure} \circ \mathcal{R}$ will always have fixed point imprecise credences.

We propose to characterise undermining imprecise credal states by:[24]

DEFINITION 8.8.1. $\mathbb{C}$ is *non-undermining* iff $\mathcal{R}(\mathbb{C}) \subseteq \mathbb{C} \subseteq \mathsf{closure}\ (\mathcal{R}(\mathbb{C}))$.

This says that every $c \in \mathbb{C}$ should have its recommended credence in the set, i.e., $\mathcal{R}(\mathbb{C}) \subseteq \mathbb{C}$, and that every credence function in the set should either be recommended by some member of the set or be the limit of a sequence of such recommended functions, i.e., $\mathbb{C} \subseteq \mathsf{closure}(\mathcal{R}(\mathbb{C}))$.

This definition allows that any $\mathbb{C}$ which is a fixed point of $\mathcal{R}$ is non-undermining.[25] So are any fixed points of $\mathsf{closure} \circ \mathcal{R}$, and thus by Theorem 8.7.5 there is always some non-undermining credal state. A state can also be non-undermining without being a fixed point of either of these if it contain some but not all members of the closure.[26]

What credal states are non-undermining in the cases mentioned? (Most of these were introduced in Section 8.3.)

- For the PASSPORT case, {0, 1} is the only credal state which is non-undermining. The same revision function is used for BAD NAVIGATOR and CredLiar, so the same holds for these cases too.
- GOLF is similar, and the only non-undermining credal state is {0.495, 0.505}.
- For normal cases, like RAIN, where $\rho(c) = c$ for all c, every imprecise credal state, $\mathbb{C}$, is a fixed point of $\mathcal{R}$, and thus is non-undermining. Note that if we *required* states to be fixed points of $\mathsf{closure} \circ \mathcal{R}$, then a set which is not closed, such as $(0.2, 0.8) = \{x \mid 0.2 < x < 0.8\}$ would be undermining as it doesn't contain its limit points of 0.2 and 0.8. But we would like to say that it is non-undermining, and our definition allows this.
- Since the revision function of LEAP is identical to that of RAIN, it too says that any $\mathbb{C}$ is non-undermining.
- For an extremal version of LEAP where $\rho(x) = 1$ if $x \geq 0.5$ and $\rho(x) = 0$ if $x < 0.5$, the non-undermining options are 0, 1 and {0, 1}.[27]

- For Promotion, the precise credence 0.5 is non-undermining. But so is any imprecise credence with $x \in \mathbb{C}$ iff $1 - x \in \mathbb{C}$, e.g., {0.2, 0.8}, or (0.2, 0.8).[28]
- For Spring, $\{0, 1, {}^1/_2, {}^1/_4, \ldots\}$ is a fixed point of $\mathsf{closure} \circ \mathcal{R}$; it is the only non-undermining state.

8.9 Other Applications and General Considerations

We have here considered the notions of truth and rational credence. But the considerations and construction we give here is very general. It could fruitfully apply to a whole range of target domains, for example: reference or satisfaction; membership or exemplification; necessity or knowledge; or decision theoretic or game theoretic rationality.[29] All one needs in order to apply it to a target domain is to specify the collection of all potential precisifications, $\mathsf{AllPrecs}$, and how to revise each of them, i.e., specify a revision function, $\rho : \mathsf{AllPrecs} \to \mathsf{AllPrecs}$. The revision function should be such that one would like to find fixed points of it, though this may not be possible. This can be done for all the domains mentioned.

For truth, $\mathsf{AllPrecs}$ was given by $\mathcal{P}(\mathsf{Sent}_T)$ and the revision function was spelled out with Tarskian truth revision jump, τ. But we could also consider variations of this setup, for example, we might consider restricting to just certain kinds of precise interpretations, for example those that are maximally consistent.

For credences, $\mathsf{AllPrecs}$ was given by $\mathsf{Creds}_{\mathcal{A}} = [0, 1]^{\mathcal{A}}$ and we just took the revision function to be (externally) given. We could also modify this setup. We could restrict it to just the functions that are (finitely additive) probability functions. We could also consider credence, or probability, as spelled out over possible world structures. (See Section 8.3.2 for a brief description of this and the associated revision function.) We can thus obtain a supervaluational variant of a Kripkean account of probability paralleling the strong Kleene version developed in Campbell-Moore (2015), or a probabilistic variant of the supervaluational Kripkean account of necessity as in Nicolai (2018).[30] There are some advantages of the supervaluational approach over the strong Kleene one, especially as understood as giving sets of credal-evaluation-functions. For example, we can immediately read off a whole range of definite facts, for example about conditional probability, whereas it's not immediately clear how to consider conditional probability in a strong Kleene framework. We might also consider joint theories of credence and truth.

Once one has a collection of potential precisifications, we can then consider the indeterminate variant of one's target notion to be given by sets of precisifications. And we can define an operator $\mathcal{R}$ which applies to a set of precisifications just by revising each precise interpretation in accordance with the specified revision function. Whilst very natural, this will also typically not guarantee fixed points.

Having further investigated the usual supervaluational account for truth as it applies to sets of precisifications, we proposed allowing one to add additional precisifications which are not individually recommended, but are in the closure of the set of recommended precisifications. To apply this in general, one also needs to define an appropriate notion of closure, which should be monotone and increasing. Many domains come along with natural notions of closure.[31] Often this will be compact, as for example our topology on Creds was. There are some limitations, for example, if we restrict the precise interpretations of truth to those that are ω-consistent, or the credence functions to those that are countably-additive, compactness is lost. Whenever the space it is compact, fixed points of $\mathsf{closure} \circ \mathcal{R}$ can be found.

We then proposed a notion of when indeterminate credal states are non-undermining: when $\mathcal{R}(\mathbb{C}) \subseteq \mathbb{C} \subseteq \mathsf{closure}(\mathcal{R}(\mathbb{C}))$. And we observed that since $\mathsf{closure} \circ \mathcal{R}$ has non-trivial fixed points, there are always some non-undermining credal states. The analogous definition could be applied to other domains and further investigated. Or one might explicitly focus on $\mathsf{closure} \circ \mathcal{R}$.[32]

Our main focus has been on sets of precisifications, but one might be interested some independently specified supervaluational models. For example, just focusing on certain definite judgements as is usually done for truth. Again, the general account we have developed here can immediately apply. To apply it, one should spell out operators analogous to Precs and Def, which we used for truth, that allow us to associate sets of precisifications with one's independently specified models. One can then consider two jumps on these supervaluational models, one which tracks the action of $\mathcal{R}$ on one's supervaluational models ($\mathsf{Def} \circ \mathcal{R} \circ \mathsf{Precs}$), and the other that tracks the action of $\mathsf{closure} \circ \mathcal{R}$ ($\mathsf{Def} \circ \mathsf{closure} \circ \mathcal{R} \circ \mathsf{Precs}$). Usually, the jump of one's supervaluational models corresponding to $\mathsf{closure} \circ \mathcal{R}$ will obtain fixed points whereas that corresponding to $\mathcal{R}$ often does not.

However sometimes the action of $\mathcal{R}$ on one's supervaluational models does lead to fixed points even though $\mathcal{R}$ itself does not. This is what we observed in the case of truth where the action of $\mathcal{R}$ on definite verdict assignments was given by Δ and had non-trivial fixed points even though $\mathcal{R}$ did not. This is because in this special case, the jumps $\mathcal{R}$ and $\mathsf{closure} \circ \mathcal{R}$ are identical insofar as they act on definite verdict assignments. The formal reason for this is that any $Q^* \in \mathsf{closure}(\mathbb{Q})$ is in $\mathsf{Precs} \circ \mathsf{Def}\,(\mathbb{Q})$.

But this typically won't be the case, especially for a notion like credence. For example, if we work with definite judgements regarding one's credences, $\mathcal{R}$ and $\mathsf{closure} \circ \mathcal{R}$ will act differently. Consider focusing on definite judgements as to whether one's credence in φ is $> r$ or not. In the case of

Spring, for example, every $c \in \mathcal{R}(\mathsf{Creds})$ has $c(\textsc{Spring}) > 0$, so this is a definite judgement, but when we consider $\mathsf{closure} \circ \mathcal{R}\,(\mathsf{Creds})$ we find that $c(\varphi) > 0$ is no longer definitely satisfied as some member of the closure has $c(\varphi) = 0$. This means that even once we consider definite judgements in the case of credences, we have to work with the jump analogous to $\mathsf{closure} \circ \mathcal{R}$ rather than $\mathcal{R}$ to ensure there are fixed points.

It would not help to focus on definite judgements regarding different properties. For example, regarding whether one's credence is $\geq r$, or equal to r, at least if one is interested both in whether it definitely holds and whether it definitely doesn't hold. This is because for either the positive or negative component, taking a closure can make a difference, and thus the jumps corresponding to $\mathcal{R}$ and $\mathsf{closure} \circ \mathcal{R}$ will differ, and it's only $\mathsf{closure} \circ \mathcal{R}$ that will guarantee fixed points. Formally, this is because, for the case of credences, a set and its complement won't both be closed (unless one of them is empty). It was a special feature of truth-values (they're discrete) that meant that focussing on definite truth value verdicts allowed triviality to be avoided.[33]

There are still ways to avoid such worries. Consider modelling one's supervaluational credences with a partial function, where to to some sentences you do not form any attitude whatsoever, and to the others, you assign normal precise values. This is essentially caring about whether one's credence is definitely equal to r without caring about whether it is definitely not equal to r. Since taking a closure of a set cannot reject any definite credence value assignments, $\mathcal{R}$ and $\mathsf{closure} \circ \mathcal{R}$ act identically on these definite credence value assignments, and thus there will be fixed points of the jump associated with $\mathcal{R}$ (as there are fixed points of the jump associated with $\mathsf{closure} \circ \mathcal{R}$). This will be the case whenever we only consider positive definite judgements on closed properties. But one very legitimately might be interested in other properties, such as whether one's credence is definitely $> r$, or if it's definitely not equal to r, in which case $\mathsf{closure} \circ \mathcal{R}$ needs to be considered.

In summary: We have proposed an account of underminingness applying to indeterminate credal states where there are always non-undermining indeterminate credal states even in the face of certain scenarios which have recently been considered as a challenge to rationality, and which bear a close relationship to the liar paradox. Along the way, we have obtained a deeper understanding of the supervaluational Kripkean account of truth, especially as it applies to sets of precisifications, and offered a very general account that could be applied to a whole range of target domains.

Acknowledgments

Thank you very much to Johannes Stern and Carlo Nicolai for insightful conversations and detailed feedback on this chapter. Variants of this talk have been in numerous conferences and seminars, I'm grateful for the organisers and the many helpful questions and conversations. My research was supported by a Leverhulme Trust Research Fellowship.

Notes

1. Recent discussion of such cases was initiated by Caie (2013) and Greaves (2013). However they both consider rationality considerations to apply to these cases in a consequentialist manner, which is not the way I am working with them. Instead, I am following Konek and Levinstein (2019) and Joyce (2018). See also Carr (2017) and Pettigrew (2018) for further discussion.
2. We have chosen to follow McGee (1989, 1990) in thinking about this as definite truth rather than a partial interpretation of truth as Kripke (1975) did. All the formal work would equally well apply to the partial interpretations picture.
3. Whilst we officially allow that sentences may be both in S^+ and S^-, such definite verdict assignments will have no precisifications, and thus will be trivial.
4. It would not affect our account if we took them to be propositions understood in a different way, e.g., they could be sets of possible worlds.
5. We may need to ensure that $\mathcal{A}$ contains all relevant sentences to obtain the revision notion; that is, the rationally recommended credence value of a sentence in $\mathcal{A}$ is settled by a hypothesis about the credence value of sentences in $\mathcal{A}$. See the notion of self-ref agenda from Campbell-Moore (2016, section 6.2).
6. See, e.g., Pettigrew (2016, Definition 1.0.1).
7. To see that this is not compact, observe that the limit of a convergent sequence of countably additive probabilities might be merely finitely additive. This is not the case for finite additivity.
8. In fact, many situations will not give rise to a recommendation *function*. Instead, maybe there are ties. Our whole account can be expanded to deal with ties, see note 23. However, this would complicate the presentation and the parallel to the truth case, so I assume that the notion of revision is functional.
9. In fact, one might want to specify ρ so that we only have $\rho(c) = c$ if c satisfies further principles of rationality, for example the Principal Principle.
10. For arguments for this in our "unusual" cases see especially (Joyce, 2018). For this argument in "safe" cases, see discussions of immodesty, e.g., Joyce (2009) and Lewis (1971). Someone like Pettigrew (2018) who disagrees with Joyce on these "unusual" cases might be thought of as agreeing with the idea that one's credence should be a fixed point of ρ, but instead works with an implementation of ρ which is consequentialist. It turns out then that ρ does not depend on the input value at all, and thus it always has a fixed point.
11. Joyce in fact proposes that $\rho(c)$ is the function that minimises expected inaccuracy, given that c is chosen (Joyce, 2018, p. 257), but this will typically simply be the c thus conditionalised.
12. See Campbell-Moore (2016, Section 5.3). In Section 4.1, I also extended this to the imprecise setting and considered $\mathcal{R}$. However, it is there claimed that $\mathcal{R}$ will guarantee fixed points, but these might be empty due to considerations in this chapter.
13. There is in fact a whole range of models of belief that are discussed. Many of these are weaker than arbitrary sets of probabilities, e.g., upper-lower probability models. However, especially under the term "imprecise probabilities", some models are stronger as they can encode opinions that would only be captured by non-Archimedean probability functions. Campbell-Moore and Konek (2019) present a model of belief that is more general than all

those considered in the imprecise probability literature, and where considerations such as those in this chapter might also be able to be applied. Moreover, some of the issues we find of ensuring fixed points may be more easily avoided and will be presented in a future article.

14. For example if two definite judgement assignments both have no precisifications, they are treated as identical.
15. See also note 23.
16. Our choice of $\rho(0) = 1$ is an extreme way to spell out the details of the story: if you assign credence 0, you think it's definitely true. We have made this choice as it is then parallel the McGee sentence which we will discuss in Section 8.6.1.
17. When we conceive of it as apply to the whole language, which includes, for example, the McGee sentence. It does have fixed points when applied, e.g., just to Liar.
18. This style of argument can directly be used to show that $\mathsf{closure} \circ \mathcal{R}$ has non-trivial fixed points, see note 22.
19. If $\mathcal{A}$ is countable, we just need to look at ω-length sequences.
20. See Willard (1970, Theorem 3.7). To apply these considerations in general one does not need the notion of closure to satisfy all the usual properties, it suffices to assume that $\mathsf{closure}$ is monotone and increasing.
21. This is equivalent to the usual definition of compactness, see Willard (1970, Theorem 17.4).
22. See, e.g., Willard (1970, Theorem 17.4). See also notes 22 and 25 for use of this alternative picture and a comment that it then offers a close relationship to the revision theory.
23. Alternative arguments are possible: Firstly, we could observe that what compactness shows us is that the non-empty closed subsets of $\mathbb{C}$ forms a ccpo in the sense of Visser (1984). And since we can restrict attention to the closed subsets for the purposes of $\mathsf{closure} \circ \mathcal{R}$, there must be a fixed point. Secondly, we can define a revision sequence: $c_0 \in \mathsf{Creds}_{\mathcal{A}}$, $c_{\alpha+1} = \rho(c_\alpha)$, and let c_μ be a cluster point of the preceding sequence (in Campbell-Moore, 2019 we proposed using this as the limit criterion in the revision theory), and observe that $c_\alpha \in \mathbb{C}_\alpha$. This relies on the ability to always find a cluster point, which is equivalent to compactness.
24. To extend this to the case where ρ doesn't pick out a unique credence function but can allow for ties, we will say that $\mathbb{C}$ is non-undermining iff each $c \in \mathbb{C}$ has at least one of their maximally recommended credences in $\mathbb{C}$, and everything in $\mathbb{C}$ is in the closure of recommended credences.
25. This is a key reason to give our definition rather than saying that it has to be a fixed point of $\mathsf{closure} \circ \mathcal{R}$. I would like to say the further thing that they are preferable: if they exist then they are required, but the criterion does not do this. To account for this intuition, we might also define a notion of recommendation for imprecise credences as: $\mathbb{C}$ recommends $\mathcal{R}(\mathbb{C})$ and say that ideally one's credal state should be self-recommending, but if it cannot be, it should at least be non-undermining.
26. A further advantage of this definition is it leads to a nice relationship with the revision theory of Gupta and Belnap (1993). If one has a revision sequence, following ρ as the revision step and using the limit criterion that the limit be a cluster point of the preceding sequence, then the collection of members of the looping part of the sequence is non-undermining. So are any unions of such revision loops. (However, there are non-undermining states which are not unions of such revision loops.)
27. This could be described as a "truth-teller" variant of $\mathsf{CredLiar}$.

28. In fact we can find further non-undermining states such as (0.2,0.8).
29. The first three of these collections of examples are taken from Gupta and Belnap (1993, Section 7.2), game theory is another place where revision theory has been applied (Bruni and Sillari, 2018), and decision theoretic cases such as Death in Damascus (Gibbard and Harper, 1978) where every action is undermining would be another natural application.
30. Nicolai, along with all the other work on predicate approaches to modality, does not consider sets of precisifications but rather, in our terms, whether a sentence is definitely necessary (at world w) or definitely not necessary (at w).
31. Often one has an underlying domain of "values", such as $\{\mathsf{true}, \mathsf{not\text{-}true}\}$ or $[0, 1]$, and the space of precisifications is the collection of functions from some other objects such as sentences, or pairs of worlds and sentences, to these values. One then just needs to impose a topology (or notion of closure) on the underlying values and then can use the topology of pointwise convergence (the product topology) to obtain a notion of closure on the collection of all possible precisifications. By Tychonoff's theorem this will be compact so long as the topology on the underlying values is compact (as $\{\mathsf{true}, \mathsf{not\text{-}true}\}$ or $[0, 1]$ is). One might also consider only special kinds of functions (for example those that are maximally consistent), in which case one should also check that the set of all such functions is closed in the full function space. See a text on general topology such as Willard (1970). See also Campbell-Moore (2019) where we used this topology for truth and probability to take limits in the revision theory.

    One could also work with alternatives, such as simply defining $\mathsf{closure}(\mathbb{Q})$ as $\mathsf{Precs} \circ \mathsf{Def}(\mathbb{Q})$ itself. This would suffice, but this will generally be more permissive than $\mathsf{closure}$ defined by the topology of pointwise convergence. An interesting exception is if we restrict $\mathsf{AllPrecsT}$ to those that are maximally consistent where then these are identical.
32. For credences, this seems unmotivated, but perhaps it is appropriate for other domains.
33. As both the property of a sentence being true and being not-true are closed properties.

References

Bradley, S. (2015). Imprecise probabilities. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*, Summer 2015 edition.

Bruni, R. and Sillari, G. (2018). A rational way of playing: Revision theory for strategic interaction. *Journal of Philosophical Logic*, 47(3): 419–448.

Caie, M. (2013). Rational probabilistic incoherence. *Philosophical Review*, 122 (4): 527–575.

Campbell-Moore, C. (2015). How to express self-referential probability. A Kripkean proposal. *The Review of Symbolic Logic*, 8: 680–704.

Campbell-Moore, C. (2016). *Self-Referential Probability*. PhD thesis, Ludwig-Maximilians-Universität, München.

Campbell-Moore, C. (2019). Limits in the revision theory. *Journal of Philosophical Logic*, 48(1): 11–35.

Campbell-Moore, C. and Konek, J. (2019). Believing probabilistic contents: On the expressive power and coherence of sets of sets of probabilities. *Analysis*, 80 (2): 316–331.

Carr, J. R. (2017). Epistemic utility theory and the aim of belief. *Philosophy and Phenomenological Research*, 95(3): 511–534.

Egan, A. and Elga, A. (2005). I can't believe I'm stupid. *Philosophical Perspectives*, 19(1): 77–93.

Gibbard, A. and Harper, W. L. (1978). Counterfactuals and two kinds of expected utility. In *Ifs*, pages 153–190. Springer.

Greaves, H. (2013). Epistemic decision theory. *Mind*, 122: 915–952.

Gupta, A. and Belnap, N. D. (1993). *The Revision Theory of Truth*. MIT Press.

Halbach, V. (2014). *Axiomatic Theories of Truth*. Cambridge University Press, Revised edition.

Halbach, V., Leitgeb, H., and Welch, P. (2003). Possible-worlds semantics for modal notions conceived as predicates. *Journal of Philosophical Logic*, 32: 179–222.

Halbach, V. and Welch, P. (2009). Necessities and necessary truths: A prolegomenon to the use of modal logic in the analysis of intensional notions. *Mind*, 118 (469): 71–100.

James, W. (1897). The will to believe. In *The Will to Believe and Other Essays in Popular Philosophy*, pages 1–15. New York: Longmans, Green, and Co.

Jeffrey, R. (1984). Bayesianism with a human face. In Earman, John, editor, *Testing Scientific Theories*. University of Minnesota Press, pages 133–156.

Joyce, J. M. (2009). Accuracy and coherence: Prospects for an alethic epistemology of partial belief. In *Degrees of Belief*, pages 263–297. Springer.

Joyce, J. M. (2010). A defense of imprecise credences in inference and decision making. *Philosophical Perspectives*, 24: 281–323.

Joyce, J. M. (2018). Accuracy, ratification, and the scope of epistemic. In Ahlstrom-Vij, K. and Dunn, J., editors, *Epistemic Consequentialism*, chapter 10. Oxford University Press.

Konek, J. and Levinstein, B. A. (2019). The foundations of epistemic decision theory. *Mind*, 128(509): 69–107.

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

Levi, I. (1978). On indeterminate probabilities. In *Foundations and Applications of Decision Theory*, pages 233–261. Springer.

Lewis, D. (1971). Immodest inductive methods. *Philosophy of Science*, 38: 54–63.

McGee, V. (1989). Applying Kripke's theory of truth. *The Journal of Philosophy*, 86(10): 530–539.

McGee, V. (1990). *Truth, Vagueness, and Paradox: An Essay on the Logic of Truth*. Hackett Publishing.

Nicolai, C. (2018). Necessary truths and supervaluations. *From Arithmetic to Metaphysics*, 73: 309.

Pettigrew, R. (2016). *Accuracy and the Laws of Credence*. Oxford University Press.

Pettigrew, R. (2018). Making things right: The true consequences of decision theory in epistemology. In Ahlstrom-Vij, K. and Dunn, J., editors, *Epistemic Consequentialism*, chapter 9. Oxford University Press.

Stern, J. (2015). *Toward Predicate Approaches to Modality*, volume 44 of Trends in Logic. Springer.

Visser, A. (1984). Four valued semantics and the liar. *Journal of Philosophical Logic*, 13(2): 181–212.

Willard, S. (1970). *General Topology*. Courier Corporation.

9 The Fourth Grade of Modal Involvement

Volker Halbach

9.1 Four Grades of Modal Involvement

Modalities are at the centre of many philosophical debates. They include apriority, analyticity, metaphysical, as well as logical and physical necessity, future and past truth, and being knowable or known. In philosophical logic the main tool for analyzing modal discourse is modal or intensional logic in the wide sense that includes temporal, epistemic, and deontic logic. The expressive weaknesses of modal logic have been known for a long time, and it is somewhat of an embarrassment of philosophical logic that the standard formalization of modality in its straightforward forms does not suffice as a framework for many well-known philosophical debates.

In this chapter I give some hints for developing formal languages with possible worlds semantics that can analyze modal discourse in its full force. Much of what I am going to say is tentative and experimental. This is because the project is huge: There is at least as much scope for work on this approach than on quantified modal logic. Therefore, I can only sketch some basic considerations. There are many aspects that will not be addressed at all. In particular, I will not say anything about extensions such as actuality operator, two-dimensional semantics, and so on.

I focus on metaphysical necessity as my main example of a modality; but the machinery sketched below is also adaptable to other modalities. Future and past truth should be straightforward, but epistemic modalities may be more difficult.

As a starting point, I revisit Quine's "Three grades of modal involvement" (1976, first published 1953). The three grades are three ways to view necessity, namely as a *semantical predicate*, a *statement operator*, or a *sentence operator*.

In modal logic modalities are conceived as *operators*: An operator $\Box$ is combined with a formula φ to obtain a new formula $\Box\varphi$. The difference between statement and sentence operators is that the latter allow quantification into the scope of the modal operator. Thus a sentence operator

$\Box$ can be combined with a formula φ containing free variables into a formula $\Box\varphi$ that contains the same free variables as φ itself. Quine's terminology is somewhat misleading, because sentences are understood here as *open* sentences with free variables. A statement operator does not permit quantification into its scope. *De re* modalities can be analyzed with sentence, but not with statement operators.

According to Quine, the lowest grade of modal involvement permits only 'semantic predicates': A semantic predicate Nec for necessity is combined with a singular term into a formula.[1] Quine thinks of semantic predicates as applying to sentences, not propositions or other objects that Quine called 'creatures of darkness' elsewhere; but this is not the point here. It is rather that a semantic predicate can be applied to a variable: $\mathsf{Nec}\ x$ is a formula with x free.

Quine thought of semantic predicates as a light grade of modal involvement, but in another sense they are expressively richer than operators. Using modal predicates one can express quantifications. Quantified statements such as 'There are synthetic judgements a priori', 'All theorems of arithmetic are analytic', or 'There are necessary a posteriori truths' cannot be expressed using only an operator $\Box$, at least not in a straightforward way. However, if modalities are conceived as predicates, quantified statements can be easily expressed: Kant's rejection 'There are synthetic judgements a priori' of empiricism, becomes the sentence $\exists x\ (\mathsf{Syn}(x) \wedge \mathsf{Apriori}(x))$ (omitting the restriction to judgements).[2]

For these reasons we require modal predicates. They are required to analyze the full force of modal discourse, especially in philosophy. With a semantic predicate for a modality comes the expressive power of what Quine called a 'statement operator'. That is, under fairly general assumptions everything that can be expressed using a statement operator $\Box$ can be expressed also with a semantic predicate. As long as φ does not contain $\Box$, a sentence $\Box\varphi$ can be replaced with $\mathsf{Nec}\ulcorner\varphi\urcorner$ where $\ulcorner\varphi\urcorner$ is a quotation or structurally descriptive name for the *sentence* φ. This induces a translation from the entire language with a statement operator to a language with a modal predicate only. However, this reduction or translation does not work for what Quine called 'sentence operators'. When a formula $\Box\varphi$ with free variables is replaced with $\mathsf{Nec}\ulcorner\varphi\urcorner$, the free variables disappear; they are only mentioned, not used in $\mathsf{Nec}\ulcorner\varphi\urcorner$. In other words, Quine's semantic predicates do not permit *de re* modality.

Famously, Quine argued against *de re* modalities and sentence operators that apply to formulae with free variables—and failed completely at convincing the philosophical community to reject them. *De re* modalities are nowadays used without qualms almost everywhere; and it is hard to imagine many discussions in contemporary philosophy without the additional expressive strength gained from using *de re* modalities. Among the *de re* modalities analyzed in modal or intensional logics are temporal modalities (future and past truth), different epistemic modalities such as knowledge, and several varieties of necessity, especially metaphysical necessity. However, semantic predicates in Quine's sense do not suffice for analyzing *de re* modality; in particular, sentence operators are not reducible to semantic predicates.

I would not like to miss the expressive power of quantifying over the bearers of necessity or of quantifying into modal contexts. The full strength of modal discourse requires both: a predicate conception of modality and *de re* modality. None of Quine's three grades of modal involvement offers the strength of modal predicates *and* that of sentence operators. There is a conspicuous omission from Quine's list of grades of modal involvement: There are no predicates expressing *de re* modalities.[3]

One could hope that nevertheless semantic predicates together with sentential operators in Quine's sense would suffice: For quantified statements of the kind mentioned above a predicate, and for *de re* modality an operator that allows quantifying-in could be used. This is in itself unsatisfactory, because we need to switch between quantified modal logic for dealing with *de re* modalities and modal predicates for dealing with general quantified claims. But there is a class of statements for which both, an operator or *de dicto* predicate are insufficient. For this kind of statement the fourth grade of modal involvement is required. An example is the following claim:

(E) There is an object *o* and there is a *P* such that *P* is necessary for *o*.

This claim may be taken as a statement of some kind of essentialism. Of course metaphysicians will disagree what kind of object *P* can be: The traditional answer will be that *P* must be a universal or property, while Quine and others might view *P* as a formula.

The point is that neither a semantic predicate nor a sentential operator suffice to express (E) without further assumptions or additional devices. This can be seen as follows. If we use an operator as in modal logic, $\exists x \, \Box Px$ fails as a formalization of (E), because it only claims that the specific property *P* is necessary for some object. In order to quantify over *P* we need a modal predicate (or higher-order quantification or the like). A unary predicate Nec for *de dicto* necessity may also be insufficient, but for a different reason: With a semantic predicate we cannot quantify any longer easily 'into the modal context', that is, over the objects that may have a property necessarily, at least if the semantic predicates apply to sentences. We may be able to express that some sentence of the form *Pt* is necessary, where *t* is some closed term; but, of course, the claim (E) just says that *P* is necessary for some *o*, whether *o* has a name or not.[4] Quine's semantic predicates cannot capture this higher kind of modal involvement. Discussions on various versions of essentialism, *ante rem* and *in rebus* conceptions of universals, and so on require *de re* modalities as predicates. Nothing should keep us from quantifying at the same time into modal contexts and over properties or predicates. Only the fourth grade of modal involvement affords this in a straightforward way.

In the following diagram the four grades of modal involvement are ordered by their strength. An arrow from one grade to another means that the grade at the origin of the arrow is stronger than that at the target, that is, the grade at the end of an arrow is reducible to that at its origin.

$$\begin{tikzcd}
 & \textit{de re}\text{ semantic predicate} \arrow[dl] \arrow[dr] & \\
\text{semantic predicate} \arrow[dr] & & \text{sentence operator} \arrow[dl] \\
 & \text{statement operator} &
\end{tikzcd}$$

At the bottom, as the lowest form of modal involvement, is the conception of operator modal logic without any quantifying-in.[5] As explained above, treating a modality as a semantic predicate or permitting quantifying-in increases the expressive strength. Neither is a semantic predicate for a modality reducible to the corresponding sentence operator nor *vice versa*: Semantic predicates and sentence operators add different kinds of expressive strength. The latter add *de re* modality, the former the possibility of quantifying over objects to which the modality is ascribed. At the top sits the highest, the fourth grade of modal involvement under which all others can be subsumed, a device that gives both kinds of expressive strength.

A formal framework for this fourth grade of modal involvement will need to deal not only with the traditional problems of modal predicates and *de re* modality as in quantified modal logic; there are also a few additional puzzles. There is not much literature on modal *de re* predicates. George Bealer's contributions (1982, 1993, 1998) are probably the most sophisticated.

9.2 *De Re* Semantic Predicates

Different strategies have been tried to reach the fourth grade of modal involvement. One option is to use a sentence operator for necessity (permitting quantifying-in) and to expand the language with second-order order, substitutional, or propositional quantification. To emulate propositional quantification, quantifiers ranging over sets of possible worlds can be used, as suggested by Kripke (1959). Bull (1969), Fine (1970), Kaplan (1970), and subsequently many others elaborated on Kripke's suggestion. The simple approaches formulated over propositional logic have to be generalized to quantified modal logic to recover the strength of the fourth grade of modal involvement. This will yield some kind of second-order quantified modal logic as in Williamson (2013). I cannot provide a discussion of these approaches here. I refer the reader to Halbach and Leigh (2021) for a sketch of reasons why this approach is less promising than the strategy pursued here. One worry is that it will be difficult to express quantification over predicates of arbitrary arity: In higher-order logic one can quantify over propositions (0-place predicates), unary predicates, binary predicates, and so on; but simultaneously quantifying over all predicates of arbitrary arity is not easily possible without further resources.

Another option, with a similar problem, is the use of modal predicates of different arities: There would be a binary predicate that expresses that a single object has a necessary property or that a formula with one free variable is necessary for that object. A ternary predicate would be required for binary relations or formulae with two free variables, and so on. This becomes very clumsy, because there will be a necessity predicate for each arity (or one with variable arity). Moreover, we cannot easily express quantification over predicates with arbitrary arity. Examples of such quantifications will be given below.

The obvious solution consists in adapting Tarski's (1935) trick for truth to modalities. Tarski used a satisfaction predicate applying to formulae and variable assignments to define the unary truth predicate. Corresponding to the truth predicate we have the unary *de dicto* predicate for necessity; and corresponding to the binary satisfaction predicate, applying to formulae and variable assignments, we have the binary *de re* necessity predicate applying to formulae (or universals) and variable assignments.

A language with such a binary modal predicate is expressively richer than one with modal predicates for each arity. By quantifying over formulae (or corresponding universals) and variable assignments we are not restricted to a fixed arity. For instance, we can express essentialist claims in the style of (E) above in a more cautious way by saying that some objects necessarily stand in some relation. With this formulation we do not commit ourselves to a relation of a specific arity. By using a binary predicate applying to predicates of arbitrary arity and sequences of objects, this can be expressed with the formula $\exists x\ \exists y\ \mathsf{Nec}(x,y)$ or $\exists x\ \exists y\ (\mathsf{For}(x) \wedge \mathsf{As}(y) \wedge \mathsf{Nec}(x,y))$, where $\mathsf{For}(x)$ expresses that x is a formula (relation) and $\mathsf{As}(y)$ that y is a variable assignment.

Using a predicate like Nec we can also express *de re* factivity. This is the claim that if a formula (or relation) is necessary of some objects, then it is true of those objects. This can be parsed as follows: If a formula x is necessary of a string y of objects, then x is satisfied by y, or formally

$\forall x\, \forall y\, (\mathsf{Nec}(x, y) \rightarrow \mathsf{Sat}(x, y))$. Again the quantifiers may be restricted to formulae and assignments as above. The factivity of *de re* knowledge can be expressed by replacing the binary predicate for necessity with a corresponding binary predicate for *de re* knowledge. *De re* factivity implies *de dicto* factivity, that is, the claim that whatever is necessary is true, but the converse implication does not hold.

At this point it is worth emphasizing the following: I am far from claiming that binary modal predicates of the kind described above are close to the methods of expressing the fourth degree of modal involvement in natural languages. A linguistic analysis is not the target of the present paper, although there are clearly some very interesting differences between modalities in the natural languages. For instance, temporal notions are usually expressed in the languages with which I am familiar by modifying the verb, either directly or with an auxiliary verb. The tense of a verb is probably most faithfully represented in a formal language by a tense operator, that is, a sentence operator in Quine's sense, not by a primitive predicate for future or past truth. Moreover, in many cases modalities can be expressed in many different ways in natural languages. We have modal operators such as *necessarily* as well as predicate phrases such as *is necessary* in English. I aim to provide a formal framework that allows us to analyze philosophical modal discourse in its full strength; this requires quantification over universals or formulae and variable assignments without fixed arity. If possible, I prefer to be parsimonious and not introduce devices that are reducible to binary predicate for *de re* modalities, even if there are analogues of these devices in natural languages.

All my claims so far about the strength of various devices for expressing modalities rely on appeals to some informal semantics. Of course, a plethora of various semantic systems exists for modal logics; but they are largely missing for corresponding binary predicates. Therefore, to substantiate my claims and to assess the prospects of analyzing *de re* modalities using binary predicates of the kind described, a formal semantics for languages with such predicates are needed.

9.3 Possible Worlds Semantics

In this section I extend and adapt the possible worlds semantics for a unary modal predicate developed by Asher and Kamp (1989), Halbach et al. (2003), Halbach and Welch (2009), and Halbach and Leigh (2021) to a semantics for the binary predicate $\mathsf{Nec}(x,y)$. I think of the semantics given below as a starting point. Several assumptions may be tweaked. Moreover, the account below can be generalized in many ways to different modalities and multimodal settings.

As pointed out at the end of the previous section, a formal semantics is needed in order to substantiate various claims about the strength of modal devices such as the binary modal predicate Nec, unary semantic predicates, the operator $\Box$ of modal logic as sentence and statement operators, and perhaps further expressions. This requires a certain degree of adequacy of the semantics for which I will not argue.

There are further reasons for developing possible worlds semantics for Nec. In particular, I would like to transfer insights that have been obtained using possible worlds semantics for the modal operator $\Box$. Giving up possible worlds semantics together with the operator conception of modality would be a huge loss. I do not intend to reject sentence and statement operators. My complaint is that they cannot capture the full force of modal talk; but as long as this force is not required, modal operators and their accompanying possible worlds semantics can shed light on many questions and puzzles.

The aim of this section is to define possible worlds models such that the formula $\mathsf{Nec}(\ulcorner\varphi(x_1,\ldots,x_n)\urcorner, a)$ holds at a world w iff $\varphi(x_1,\ldots,x_n)$ holds at all worlds v with wRv, if the free variables $x_1,\ldots, x_n$ are assigned values $a(x_1),\ldots, a(x_n)$. Here $\ulcorner\varphi(x_1,\ldots, x_n)\urcorner$ is a name for the formula $\varphi(x_1,\ldots, x_n)$, and $a(x_k)$ the value assigned to the variable x_k by a. I will be sloppy with notation and, for instance, use a in the object language, even though we might lack a name for it.

When setting up possible worlds semantics for Nec, the same choices as for quantified operator modal logic have to be faced, and, as is well-known, there are many (see, for instance, Garson, 2001). In particular, we have to decide whether to opt for a possibilist treatment of quantifiers with a fixed domain for all worlds or for actualism with domains that can vary between world.

There are also decisions that do not arise in quantified operator modal logic. On the predicate approach, we can quantify over the objects that can be necessary and they can inhabit possible worlds along with other objects (unlike propositions conceived as sets of possible worlds, which cannot themselves be elements of a world). Which of these objects exist would differ from world to world, if we decided to ascribe necessity to sentence tokens. I think of them as types and assume that they exist in all worlds. Metaphysical questions and decisions of this kind are suppressed in the usual variants of quantified modal logic; on the predicate approach they can be addressed in the object language.

Finite sequences of objects are needed as variable assignments for *de re* modalities. I assume that all worlds are closed under the formation of sequences. This is again a strong ontological assumption, which I endorse for metaphysical necessity, but not for other modalities. Only objects existing in the world can form part of a sequence in that world. I discuss potential problems arising from this assumption after having introduced formal possible worlds semantics for languages with modal predicates.

For my purposes here there is no need to be very specific about the language. The language should contain vocabulary that allows one to talk about either expression in the language or corresponding 'universals'. I have a preference for expansions of languages described in Halbach and Leigh (2021), but expansions of arithmetic or set theory are equally suitable. The language should provide vocabulary for talking about syntax or universals, finite sequences of objects, and possibly other 'ordinary' objects. I use the syntactic approach and terminology. The reader preferring universals will have to replace 'sentence' with 'proposition', 'formula with one free variable' with 'property', and so on. Given that we need arbitrarily long finite sequences of arbitrary objects as variable assignments anyway, we can understand expressions as a specific kind of such sequences, namely those with symbols as members of the sequence.

Each expression in the non-logical vocabulary falls into at least one of three groups:

1. *The syntactic vocabulary* contains at least a unary predicate that applies exactly to all expressions of the language. Quantifiers relativized to this predicate range exactly over all syntactic objects. The other syntactic vocabulary may express operations such as concatenation and substitution and allow one to define grammatical categories such as variables, predicate expressions, connectives, etc.
2. *The sequence vocabulary* includes again a relativizing predicate that applies exactly to finite sequences. Further vocabulary may express concatenation of finite sequences, projections, etc.
3. *The 'contingent' vocabulary* contains all remaining non-logical expressions. A relativizing predicate is not required, because 'contingent' objects are exactly those that are not syntactic or sequences.

As mentioned above, it is not assumed that finite sequences and expressions do not overlap. However, the contingent objects are exactly those that are neither expressions nor sequences. The use of the term 'contingent' may be somewhat misleading. It only means that no assumptions about the interpretation of the contingent vocabulary are made in possible worlds semantics. This does not rule out that the contingent vocabulary is interpreted at all worlds in the same way in a given model. For instance, some vocabulary of pure mathematics may form part of the contingent vocabulary, and we may confine our attention to interpretations that assign the same extensions to these expressions at all possible worlds. But this restriction is not imposed by possible worlds semantics, but rather by other considerations.

The following definition of a pre-model is close to the definition of a possible worlds model in modal logic, the main difference being that it is assumed that all expressions exist in every world and that worlds are closed under the formation of finite sequences.

Definition 9.3.1. A triple $\langle W, R, I\rangle$ is a *pre-model* iff

1. W is a non-empty set.
2. R is a binary relation on W.
3. I is a function that assigns each world $w \in W$ a domain $\mathcal{D}_w$. $\mathcal{D}_w$ contains (codes of) all strings of $\mathcal{L}_N$-symbols, possibly further 'contingent' objects and all finite sequences of all objects in $\mathcal{D}_w$. Moreover I provides interpretations of the contingent vocabulary over the domain $\mathcal{D}_w$.

Of course $\langle W, R\rangle$ is a frame in the usual sense in modal logic. Each world contains contains all $\mathcal{L}_N$-expressions and therefore infinitely many objects. Moreover, finite sequences can be formed from all objects in the world. Sequences themselves can be elements of sequences. Therefore each domain $\mathcal{D}_w$ is closed recursively under the formation of sequences. I do not make any assumption on whether expressions conceived as sequences of symbols are identical with these finite sequences.

Consequently, the domain $\mathcal{D}_w$ of each world $w \in W$ consists of three types of objects:

1. $\mathcal{D}_w$ contains all strings of symbols.
2. $\mathcal{D}_w$ contains arbitrarily many further 'contingent' objects.
3. $\mathcal{D}_w$ contains all finite sequences of objects from 1, 2, and 3. This includes mixed sequences that have syntactic, contingent, and finite sequences as entries. Thus the domain of each world is defined recursively.

A pre-model provides interpretations for the entire language at each world $w \in W$, except for the binary necessity predicate Nec. At any world the syntactic and sequence vocabulary is interpreted in the standard way. This means that the syntactic vocabulary is interpreted in the same way at all worlds. However, worlds differ with respect to the objects that exist in them and consequently also with respect to the sequences that exist in them. Thus the interpretation of the vocabulary about sequences at a world depends only on the domain $\mathcal{D}_w$ of that world, because only sequences whose ultimate components exist at that world also exist at that world.

If φ is a formula not containing the modal predicate Nec and a a variable assignment over $\mathcal{D}_w$, I write $\langle W, R, I\rangle \models_w \varphi[a]$ to indicate that the formula φ is satisfied by the variable assignment a at world w. As mentioned above, variable assignments are conceived as finite sequences and thus the question arises about situations where φ contains variables outside the domain of a. I could rule this out without real loss of generality, but prefer to stipulate that $a(x)$ is always a certain expression, say $\neg$

and thus effectively make every variable assignment a total function from the set of variables. By our stipulations, $\neg$ is in the domain of every world.

What is missing from this account is an interpretation of the *de re* modal predicate Nec. The interpretation B of the binary predicate symbol Nec is provided separately from I: B is a function that assigns to each world a binary relation (set of ordered pairs) on the set of objects that exist at w. I write $\langle W, R, I, B\rangle \vDash_w \varphi[a]$ iff φ holds at world w under the variable assignment a if Nec is interpreted according to $B(w)$. I now define what a possible worlds model ('pw-model' for short) is.

Definition 9.3.2. $\langle W, R, I, B\rangle$ is a pw-model iff $\langle W, R, I\rangle$ is a pre-model and B a function satisfying the following properties:

(i) B is a function assigning to each world $w \in W$ a set of pairs $\langle \varphi, a\rangle$ such that φ is a formula and a is a finite sequence in $\mathcal{D}_w$.

(ii) $\langle \varphi, a\rangle \in B(w)$ iff for all v (if wRv then $\langle W, R, I, B\rangle \vDash_v \varphi[a]$).

In (ii) the variable assignment a in $[a]$ belongs to the metatheory, while pair $\langle \varphi, a\rangle$ is a possible element of the extension $B(w)$ of Nec at w and thus a an element of $\mathcal{D}_w$. That is, variable assignments belong to both, the object and metatheory.

Condition (ii) of Definition 9.3.2 forces the interpretation of Nec as truth in all accessible worlds. Suppressing the variable assignment, it means that a sentence φ is in the extension of Nec at a world w iff φ is true in all worlds accessible from w. By condition (i), if $\langle \varphi, a\rangle \in B(w)$, then $a \in \mathcal{D}_w$ and therefore $a(x) \in \mathcal{D}_w$ for all variables x. That is, a variable assignment in a world w assigns variables only values that exist in w. But of course $a(x)$ need not exist in another world, that is, we may have $a(x) \notin \mathcal{D}_v$ for some other world $v \neq w$. In particular, when we pass from a world w to all worlds v that can be seen by w, a variable assignment with $\langle \varphi, a\rangle \in B(w)$ might assign a value to the variable x that is not in $\mathcal{D}_v$. Thus we need to make a stipulation about how to understand the right-hand-side of (ii), that is, $\langle W, R, I, B\rangle \vDash_v \varphi[a]$ if φ contains a free variable x such that $a(x) \notin \mathcal{D}_v$. The situation is similar to free logic and individual constants that do not denote. There are various options. Here I adopt the policy of negative free logic. If φ is an atomic formula with a free variable x such that $a(x) \notin \mathcal{D}_v$, then $\langle W, R, I, B\rangle \nvDash_v \varphi[a]$. For instance we have $\langle W, R, I, B\rangle \nvDash_v x = x\ [a]$. Of course, we still have $\langle W, R, I, B\rangle \vDash_v \forall x\ x = x\ [a]$ as quantifiers range over objects in $\mathcal{D}_v$.

The situation is different from ordinary quantified modal logic, because we have variable assignments not only in the metatheory (as in quantified modal logic), but we are also talking about variable assignments in the object language. In $\langle W, R, I, B\rangle \vDash_v \varphi[a]$ the variable assignment a is in the metatheory of course, while in $\exists y\ \mathsf{Nec}(x, y)$ we are quantifying over variable assignments in the object language. The metatheory is extensional. Every variable assignment that exists at some world is also available in the metatheory. But of course not every variable assignment in our metatheory exists in every world: only if all values exist in w, the variable assignment exists in w.

In the usual possible worlds semantics for modal operators, one defines recursively the semantics for $\Box$ from the interpretation of the other vocabulary at each world. Definition 9.3.2, in contrast, does *not* imply that for every pre-model $\langle W, R, I\rangle$ there is exactly one B such that $\langle W, R, I, B\rangle$ is a pw-model. In fact, both, existence uniqueness fail. Halbach et al. (2003) and Halbach and Leigh (2021) provide more information on existence and uniqueness conditions. I mention only some observations that apply to the present framework.

A relation R is converse wellfounded iff there is no infinitely ascending sequence $w_1 R w_2 R w_3 \ldots$ of objects.

LEMMA 9.3.3. If the accessibility relation R of a pre-model $\langle W, R, I\rangle$ is converse wellfounded on W, then there is a unique B such that $\langle W, R, I, B\rangle$ is a pw-model.

The lemma is proved by defining B on all $w \in W$ inductively on the converse of R, starting from the dead ends of R, that is, worlds $w \in W$ such that there is no $v \in W$ with wRv. At a dead end $w \in W$, we have $\langle \varphi, a\rangle \in B(w)$ for all formulae φ and variable assignments $a \in \mathcal{D}_w$, because trivially $\langle W, R, I, B\rangle \models_v \varphi[a]$ for all worlds v with wRv, as there are no such worlds. Once $B(v)$ has been defined for all worlds v with wRv, we can set

$$B(w) := \bigcap_{wRv} \{\langle \varphi, a\rangle : \ \langle W, R, I, B\rangle \models_v \varphi[a]\} \cap \mathcal{D}_w$$

The intersection with $\mathcal{D}_w$ ensures that we only include variable assignments that exist at world w. If R is converse wellfounded, this definition fixes $B(w)$ for all $w \in W$.

Lemma 9.3.3 gives us a sufficient condition for the existence of a pw-model $\langle W, R, I, B\rangle$ based on a given pre-model $\langle W, R, I\rangle$. I turn now to necessary conditions.

Under fairly general circumstances, there cannot be a pw-model $\langle W, R, I, B\rangle$ with a reflexive or symmetric accessibility relation R. This is a direct consequence of the paradoxes. Since φ is a sentence, the variable assignment a is irrelevant and $\forall a\, \mathsf{Nec}(\ulcorner \varphi \urcorner, a)$ expresses *de dicto* necessity of φ in the same way the unary truth predicate can be defined as the satisfaction of a sentence by all variable assignments. If R is reflexive, we have the following for all sentences φ and worlds $w \in W$, as in operator modal logic:

$$\langle W, R, I, B\rangle \models_w \forall a\, \mathsf{Nec}(\ulcorner \varphi \urcorner, a) \to \varphi \tag{9.1}$$

Using the diagonal lemma, a sentence λ can be obtained such that $\lambda \leftrightarrow \neg \forall a\ \mathsf{Nec}(\ulcorner \lambda \urcorner,\ a)$ holds at all worlds. Of course, this is the liar sentence for the 'truth predicate' $\forall a\ \mathsf{Nec}(x, a)$. Combining this with (9.1) for λ yields the following for all $w \in W$:

$$\langle W, R, I, B\rangle \models_w \lambda \tag{9.2}$$

Since λ holds at all worlds, $\forall a\ \mathsf{Nec}(\ulcorner\lambda\urcorner, a)$ holds at all worlds. This clashes with following consequence of the diagonal property of λ:

$$\langle W, R, I, B\rangle \models_w \neg\forall a\ \mathsf{Nec}(\ulcorner\lambda\urcorner, a)$$

This contradiction is only a variant of Montague's (1963) theorem, as explained in Halbach and Leigh (2021). This can be generalized as follows:

Theorem 9.3.4. If the relation R is converse illfounded on W, the syntactic vocabulary sufficiently expressive, and the contingent vocabulary contains at least one sentential parameter, then there is a I such that $\langle W, R, I\rangle$ is a pre-model, but there is no B such that $\langle W, R, I, B\rangle$ is a pw-model.

For a detailed proof the reader is referred to Halbach and Leigh (2021). First it is shown that one can define the modal predicate Nec^* associated with the transitive closure of R. Using the diagonal lemma one proves that Löb's theorem holds for this predicate Nec^*. Löb's theorem is of course a principle of transfinite induction that may fail if R^* is not converse wellfounded. To show this, one can choose an interpretation I that makes the sentential parameter true at all converse wellfounded worlds, but false at the converse illfounded worlds. If there is no contingent vocabulary (the sentential parameter in the theorem), there can be a converse illfounded R such that $\langle W, R, I, B\rangle$ is a pw-model and the situation becomes complicated. See Halbach et al. (2003).

9.4 The Challenge of the Paradoxes

Theorem 9.3.4 means that the accessibility relation cannot be total on all worlds; and we cannot expect modal predicates to satisfy the schemata of the modal systems T, B, $\mathsf{S4}$, or $\mathsf{S5}$ for all sentences (especially those sentences without equivalent in ordinary modal operator logic such as sentences obtained by Gödel's diagonal lemma). If we are confined to converse wellfounded frames for possible worlds semantics of predicates, then one might suspect that we cannot have a reasonable semantics for Nec as metaphysical necessity. Perhaps the accessibility relation need not be total and, perhaps, not even transitive for metaphysical necessity; but at last it should be reflexive. This is ruled out by Theorem 9.3.4.

The reader may wonder whether this shows that the predicate approach is to be rejected. In the light of the paradoxes Montague (1963) rejected 'syntactic' treatments of modality, that is, treatments of modality as predicates of sentences. The restrictions in Theorem

9.3.4 only spell out the consequences of the paradoxes as restrictions on the accessibility relations and they thereby systematize them. Should one agree with Montague, abandon predicate approaches, and resort to operator modal logic? The price for this move will be a significant reduction in expressive power of our language. This matters most to philosophers, who are interested in the quantified claims mentioned in the first section.

The situation can be compared with that for truth: We can reject the predicate conception of truth and opt for an operator treatment. This will result the trivial modal logic with $\Box\varphi \leftrightarrow \varphi$ as characteristic axiom. Although this move immediately blocks the paradoxes, it has not been very popular. What saved truth from the fate of metaphysical necessity and other modalities is that the truth predicate becomes trivial and boring, once it is stripped of its ability to express generalizations. When metaphysical necessity is treated in the same way, there are still many interesting things to say. But that is not a good justification for accepting the weakening of expressive power caused by the transition from $\Box$ to Nec.

Rejecting modal predicates because of the paradoxes is as sensible as rejecting the theory of distances because of Zeno's paradoxes and rejecting set theory because of Russell's and Burali–Forti's paradox. We may have to revise or refine some of our naive expectation, but we will gain expressive strength in return. This strength is needed for general claims in philosophy. If we give it up in favour of modal operators, the way philosophy is done will have to be changed profoundly. Therefore, we should seek solutions to the paradoxes of modal predicates in the same spirit as in the case of truth.

Unless we are prepared to restrict ourselves to converse wellfounded accessibility relations, we need to adapt and tweak possible worlds semantics. Stern (2016) discussed various strategies. First, one can apply the usual techniques known from the semantic paradoxes. In particular, classical logic can be abandoned. Halbach and Welch (2009) employed Kripke's (1975) fixed-point semantics for their possible worlds semantics without really defending it as a real solution.

Alternatively, condition (ii) of Definition 9.3.2 can be weakened: We do no longer stipulate for *all* sentences that $\mathsf{Nec}\ulcorner\varphi\urcorner$ holds at w iff φ holds in all accessible worlds; we merely stipulate this for certain sentences. Condition (ii) should be satisfied at least for all those sentences φ that are expressible with an operator $\Box$. This guarantees that the possible worlds semantics for predicates is not 'worse' than possible worlds semantics for the operator $\Box$. The class of sentences expressible with an operator $\Box$ can be defined as follows using a recursively defined mapping I from the language with the operator $\Box$ to the language with the corresponding predicate Nec. Since we deal with *de re* modality, we need to take care of the free variables in the scope of $\Box$. Pick some fixed variable assignment a_0. The operation that changes the assignment of a given variable v_n, so that the object x is assigned to v_n can be expressed in the object language and I write $a_0(x/\ulcorner v_n \urcorner)$ for this operation.

(i) If φ does not contain $\Box$, then $I(\varphi) = \varphi$.

(ii) $I(\varphi \wedge \psi) = I(\varphi) \wedge I(\psi)$, $I(\neg\varphi) = \neg I(\varphi)$, $I(\forall x\ \varphi) = \forall x\ I(\varphi)$, and so on.

(iii) $I(\Box\varphi(x_1, \ldots, x_n)) = \mathsf{Nec}(\ulcorner I(\varphi) \urcorner, a_0(x_1/\ulcorner x_1 \urcorner, \ldots, x_n/\ulcorner x_n \urcorner))$, where the assignment $a_0(x_1/\ulcorner x_1 \urcorner, \ldots, x_n/\ulcorner x_n \urcorner)$ is obtained by iterated application of the operation mentioned above. Only finitely many such applications are required, because φ contains only finitely many variables.

This embedding of the operator languages into the language with a predicate is well-known, and its origin can be traced back at least to Carnap (1934, IV.B.e).[6] Here I have only added the free variables. The class of formulae expressible with an operator $\Box$ is now simply the set of all $I(\varphi)$ such that φ is a formula of the language with the operator $\Box$ only. If condition (ii) of Definition 9.3.2 is restricted to such sentences, there are pw-models for arbitrary frames.

Sentences and formulae generated with Gödel's diagonal construction cannot be obtained with I from sentences of operator modal logic, but also not general formulæ of the form $\forall x\ (\chi(x) \rightarrow \mathsf{Nec}\ x)$. Some but not all formulae of the latter kind can be unproblematic and be included as permissible instances of condition (ii) of Definition 9.3.2.

A more comprehensive class of permissible instances of condition (ii) of Definition 9.3.2 can be obtained by singling out the 'grounded' formulae and sentences, defined along the lines of Kripke's (1975). If we apply this account to formulæ in general, the notion of groundedness relative to a variable assignment will have to be defined. In the brief following comments I restrict myself to sentences. The groundedness approach has the disadvantage that the set of such sentences is no longer definable in the object language (under fairly general assumptions). Moreover, whether a sentence is grounded may depend on contingent factors. If all pw-models with this restricted condition (ii) are considered, many general principles such as the distribution of necessity over → in pw-models without dead ends have to be abandoned; they will only hold for grounded sentences. This flies in the face of the predicate approach: After all, I adopted the predicate approach in order to express quantified principles. Not being able to get them as quantified generalizations seems odd. At any rate, the desirable models may be sought among those that are pw-models in the weaker sense of satisfying (ii) only for grounded sentences. A different option may be to introduce a new primitive predicate for determinacy or groundedness into the language as in Fujimoto and Halbach (2019) for truth.

An elegant and philosophically useful way to obtain generalization without compromising too much on the expected properties of possible worlds semantics is given by Stern (2014a,b, 2016). The idea is to state at least quantified principles with a truth predicate, as we routinely do in informal philosophical discourse. For this strategy, however, a sophisticated theory of truth will be required that avoids the paradoxes.

The challenge of the paradoxes arises for *de dicto* and *de re* modalities, although some paradoxes require quantification. In particular, McGee's ω-inconsistency (1985) and the Yablo–Visser paradox are of this kind (Visser, 1989; Yablo, 1985, 1993). A binary predicate may lend itself to the study of these paradoxes. Finally, the move to a binary predicate can be used to recover diagonalization and other syntactic operations without any explicit syntax-theoretic axioms (Halbach and Zhang, 2016; Halbach and Leigh, 2021). I will not pursue this topic here, but rather turn to the application of the fourth grade of modal involvement to an issue in metaphysics.

9.5 A Problem for Actualism

In this section I illustrate my claim that using the fourth grade of modal involvement can shed new light on topics and discussions in metaphysics, using an argument by Bealer (1993). In many ways the predicate conception of modalities is a more versatile framework for modal metaphysics than first-order quantified modal logic. The ontology of the objects to which the modalities are ascribed is not hidden away in the metatheory. The expressive power generated by the predicate conception permits an analysis of many questions in the object theory instead of the metatheory, which is usually set-theoretic and extensional.

For instance, it has become common to assume that propositions, as the objects that can be necessary, do not exist in a possible world; as sets of possible worlds, they live in the 'modal æther' (Forster, 2005) and are thereby incomparable to other normal objects. On the predicate conception, we talk about the objects that can be necessary in the object language, and we have to decide whether they exist in a world, as sets of worlds, or wherever. In the outline above they exist in all worlds. But this is only one option; moreover, they are assumed to have the structure of formulae. A metaphysician more serious about proposition may structure them in a different way. Of course, the ontology of propositions has received much attention, and I will not pursue this topic here. Instead I focus on the second argument place of the binary modal predicate Nec, the variable assignments. When they are discussed in the usual set-theoretic metatheory, they receive little attention. If they can be talked about in the object language, they become interesting and puzzling. In particular, they can shed new light on the discussion between actualism and possibilism (or possibilism and necessitism in Williamson's 2013 terminology). The semantics outlined above is actualist in the sense that the domains $\mathcal{D}_w$ can vary between worlds w; at each w the quantifiers range only over $\mathcal{D}_w$. Possibilist semantics is a special case: Nothing rules out that $\mathcal{D}_v = \mathcal{D}_w$ for all $v, w \in W$, that is, constant domain semantics is a special case of the actualist semantics.

The argument in this section is not intended as a definitive argument in favour of possibilism or some other position; the section merely demonstrates the effects of injecting objects, in particular, variable assignments from the metalanguage into the object language and how this move can shed light on metaphysical questions.

Bealer (1982, 1993) and others have used arguments of the kind discussed in this section to argue for an *ante rem* conception of universals. I could follow Bealer here, but in the present setting—which is different from Bealer's—it would mean that variable assignments exist prior to their elements. There would be variable assignments that assign non-existing objects to some variables, which is hardly acceptable.

What is going to follow is a partial recantation of Halbach and Sturm (2004). The discussions there and Bealer's (1993) paper suffer from the absence of formal semantics. With a framework such as the one outlined above this family of puzzles can now be discussed in a more rigorous way.

As I mentioned already, in quantified modal logic variable assignments are used only in the set-theoretic, extensional metatheory, where questions about possibilism and actualism cannot arise. In our semantics the variable assignments are pushed into the object language that contains modalities. In the presence of modality, the theory of sequences (and sets) becomes more complicated in actualist semantics: Sequences exist only at a world, if all its members exist at that world. This assumption is built into our semantics. In the following example our semantics may yield an unexpected result because of this assumption.

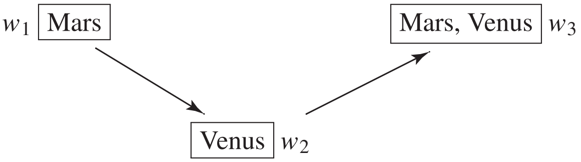
> There is a planet, and instead of this planet another planet could have existed which could have coexisted with the first. (PLA)

There may be more than one reading of this sentence. The reading I have in mind, can be made explicit using possible worlds: In our world w_1 there is a planet A and there is a possible world w_2 accessible from our world where A does not exist, but planet B does; moreover, there is another world w_3 accessible from w_2 where both planets A and B exist. The reading is captured by the following formalization in quantified modal logic:

$$\exists x \Big(\mathrm{Pla}\, x \wedge \Diamond\big(\neg\exists z\, z{=}x \wedge \exists y\, (\mathrm{Pla}\, y \wedge \Diamond(\exists z\, z{=}x \wedge \exists z\, z{=}y))\big)\Big) \qquad (\text{PLA}\Box)$$

The formula $\exists z\ z = x$ expresses that x exists. That is, $\exists z\ z = x$ becomes false at a world w under a variable assignment a if $a(x) \notin \mathcal{D}_w$. This is the case for our semantics for Nec, and we assume that the same holds for the possible worlds semantics for the operator $\Box$.

A minimal possible worlds model $\mathcal{M}$ of (PLA$\Box$) looks as follows:



The three worlds of the model are related by the accessibility relation as in the diagram. The domain of w_1 is {Mars}, and so on for the other worlds. The unary predicate Pla x applies at a world to all objects that exist in that world. It is easily seen that then (PLA$\Box$) is true at w_1 in this model.

The crucial feature of the example is that the first quantifier $\exists x$ binds an occurrence of x in the scope of *two* possibility operators $\Diamond$. The witness of the existentially quantified sentence (PLA$\Box$) at w_1 is Mars, which exists at w_1, but not in the next world w_2; it exists again in the third world w_3. In operator modal logic this does not pose a problem; it is not required that x exists in the intermediate world w_2 for the existential quantifier $\exists x$ to bind an occurrence of x in the scope of two modal operators. As I will show, however, it does cause a problem if a modal predicate is used. The formalization of (PLA) with a binary predicate Nec instead of the modal operator $\Box$ is false at w_1 in the pw-model corresponding to the model $\mathcal{M}$ above, while (PLA$\Box$) is true at w_1, as pointed out above.

To substantiate my claim that the predicate formalization gives a different truth value, I specify the predicate formalization of PLA and the (predicate) pw-model corresponding to the operator model $\mathcal{M}$ above.

First, I sketch how to transform $\mathcal{M}$ into a pw-model $\langle W, R, I, B\rangle$ for the language with Nec instead of $\Box$. W and R stay the same. The domain $\mathcal{D}_w$ for one of the three worlds $w \in W$ is obtained by adding all syntactic objects to the domain of w in $\mathcal{M}$ and then closing under the formation of sequences. Thus $\mathcal{D}_{w_1}$ contains Mars, but not Venus, $\mathcal{D}_{w_2}$ only Venus, while $\mathcal{D}_{w_3}$ contains both. Consequently, $\mathcal{D}_{w_2}$ does not contain any sequences involving Mars. Pla applies at w exactly to all planets in w. The existence of such a model is guaranteed by lemma 9.3.3 above, because the accessibility relation is converse wellfounded.

The formalization of (PLA) with the model predicate Nec looks as follows, if $\mathsf{Pos}(x, y)$ is an abbreviation for $\neg\mathsf{Nec}(\dot{\neg}x, y)$ where $\dot{\neg}$ represents the function of negating a sentence. Thus, Pos stands for possibility.

$$\exists x\Big(\mathrm{Pla}\,x \wedge \mathsf{Pos}\,(\underbrace{\ulcorner\neg\exists z\, z{=}x \wedge \exists y(\mathrm{Pla}\,y \wedge \mathsf{Pos}(\ulcorner\exists z\, z{=}x \wedge \exists z\, z{=}y\urcorner,\langle\tfrac{x}{\ulcorner x\urcorner},\tfrac{y}{\ulcorner y\urcorner}\rangle))\urcorner}_{},\langle\tfrac{x}{\ulcorner x\urcorner}\rangle)\Big)\,\text{(N)}$$

The expression $\langle\frac{x}{\ulcorner x\urcorner},\frac{y}{\ulcorner y\urcorner}\rangle$ in the underbraced quotation name and the less complex $\langle\frac{x}{\ulcorner x\urcorner}\rangle$ require an explanation. I assume that the vocabulary permits to describe the variable assignment assigning an object to a specific variable. Here I have treated $\langle\frac{x}{x_1},\frac{y}{y_1}\rangle$ like a function expression with four free variables x, x_1, y, and y_1 expressing the function that gives applied to objects x and y and variables x_1 and y_1 the relevant variable assignment; but this may have to circumscribed if no suitable function symbol is available. In addition, we need the quotation function sending an expression (here a variable) to a name for that expression. In arithmetical contexts the numeral function serves this purpose.

Since there are no further free variables, a variable assignment of length 2 in the underbraced name suffices. As mentioned above, $\neg$ is stipulated to be the value of all other variables. The 'outer' variable assignment $\langle\frac{x}{\ulcorner x\urcorner}\rangle$ has only length 1, because the preceding formula in corners has only x free. In $\frac{x}{\ulcorner x\urcorner}$ the upper occurrence of x is used, while $\ulcorner x\urcorner$ is only a mentioning of the same variable; of course, different variables could be used. If the bound variable x in the formula above is renamed as v and the relativization to planets is omitted for readability, the following formula is obtained:

$$\exists v\,\mathsf{Pos}\Big(\ulcorner\neg\exists z\,z=x\wedge\exists y\,\mathsf{Pos}(\ulcorner\exists z\,z=x\wedge\exists z\,z=y\urcorner,\langle\frac{x}{\ulcorner x\urcorner},\frac{y}{\ulcorner y\urcorner}\rangle)\urcorner,\langle\frac{v}{\ulcorner x\urcorner}\rangle\Big)$$

Of course, any occurrences in the quotation name are not affected by the renaming and x is retained.

In a nutshell, the reason why (N) fails at w_1 is the following: If (N) were true at w_1, there would have to be a variable assignment in $\mathcal{D}_{w_2}$ that assigns Mars to x and Venus to y. But there is no such variable assignment in $\mathcal{D}_{w_2}$, because Mars does not exist in w_2, that is Mars is not an element of $\mathcal{D}_{w_2}$. In fact, such a variable assignment exists only at w_3 where both, Mars and Venus exist.

A more explicit version of this argument can be given as follows. Let $\langle W,R,I,B\rangle$ be the pw-model described above. To show that (N) fails at w_1, assume to the contrary that it is true at w_1. Since Mars is the only potential witness of the existential quantifier, also the following must hold:

$$\begin{aligned}\langle W,R,I,B\rangle\models_{w_1}\ &\mathsf{Pos}\Big(\ulcorner\neg\exists z\,z=x\wedge\exists y\,(\mathrm{Pla}\,y\wedge\\&\mathsf{Pos}(\ulcorner\exists z\,z=x\wedge\exists z\,z=y\urcorner,\langle\frac{x}{\ulcorner x\urcorner},\frac{y}{\ulcorner y\urcorner}\rangle))\urcorner,\langle\frac{x}{\ulcorner x\urcorner}\rangle\Big)\quad[\langle\frac{\mathrm{Mars}}{\ulcorner x\urcorner}\rangle]\end{aligned}$$

That is, we assume that the formula $\mathsf{Pos}(\ldots)$ holds at world w_1 under the variable assignment that assigns Mars to x. The expression in square brackets is the variable assignment in the metalanguage, for which we conveniently use the same notation as in the object language. It follows from the assumption that the formula denoted by the quotation name must be satisfied by Mars for x at the world accessible from w_1:

$$\langle W, R, I, B\rangle \models_{w_2} \neg\exists z\, z{=}x \land \exists y \Big(\mathrm{Pla}\, y \land \mathsf{Pos}(\ulcorner \exists z\, z{=}x \land \exists z\, z{=}y \urcorner, \langle \tfrac{x}{\ulcorner x \urcorner}, \tfrac{y}{\ulcorner y \urcorner}\rangle)\Big) \quad [\langle \tfrac{\mathrm{Mars}}{\ulcorner x \urcorner}\rangle]$$

Hence both conjuncts must hold at w_2. The first conjunct expresses that x does not exist. Because Mars fails to be in $\mathcal{D}_{w_2}$, the first conjunct does hold indeed.

The second conjunct is an existentially quantified sentence. The only witness that cold make the second conjunct true is Venus, because the witness must be a planet and Venus is the only planet that exists in $\mathcal{D}_{w_2}$. Therefore the following must obtain:

$$\langle W, R, I, B\rangle \models_{w_2} \mathrm{Pla}\, y \land \mathsf{Pos}(\ulcorner \exists z\, z{=}x \land \exists z\, z{=}y \urcorner, \langle \tfrac{x}{\ulcorner x \urcorner}, \tfrac{y}{\ulcorner y \urcorner}\rangle) \quad [\langle \tfrac{\mathrm{Mars}}{\ulcorner x \urcorner}, \tfrac{\mathrm{Venus}}{\ulcorner y \urcorner}\rangle]$$

This implies the following:

$$\langle W, R, I, B\rangle \models_{w_2} \exists z\, z = \langle \tfrac{x}{\ulcorner x \urcorner}, \tfrac{y}{\ulcorner y \urcorner}\rangle \quad [\langle \tfrac{\mathrm{Mars}}{\ulcorner x \urcorner}, \tfrac{\mathrm{Venus}}{\ulcorner y \urcorner}\rangle]$$

That is, at w_2 there is a finite sequence containing Mars and Venus, contrary to our assumption that finite sequences in a world can contain only objects existing in that world. Therefore, the initial assumption is refuted and, therefore, (N) fails at w_1, while the operator version (PLA□) holds at w_1 in the corresponding operator model $\mathcal{M}$. Intuitively, the operator version yields the expected result, while the predicate version does not.

The problem for actualism is not caused by any restrictions on the accessibility relation. It cannot be avoided by replacing the accessibility relation R above with its transitive closure. We cannot have a total accessibility relation because of Theorem 9.3.4; but even if diagonalization were banned and total accessibility relations permitted, the problem for actualism would remain.

Various strategies to address the problem were discussed by Bealer (1993) and Halbach and Sturm (2004). I do not provide a thorough discussion, but only sketch some of them.

First, an actuality operator or predicate cannot overcome the problem: The variable assignment assigning Mars to x and Venus to y could be claimed to exist in the actual world w_1 instead of the world w_2; but the variable assignment cannot exist at w_1 either, because Venus is not in w_1. Bealer (1993) made this observation already for universals rather than variable assignments.

The second strategy sounds desperate: Variable assignment assigning non-existing objects to variables could be permitted; variable assignments would exist *ante rem*. Such entities would be true creatures of darkness. I certainly could not think of them as functions from the set of variables into the domain of the world or as any mathematical entities. However, stranger entities have been dreamt in philosophy. If the point argument is made about universals instead of variable assignment, as Bealer (1993) does, an argument for *ante rem* universals is obtained.

Thirdly, actualism could be rejected in favour of possibilism, that is, constant domain semantics. The same objects exist in all worlds, but only some are 'instantiated'. Variable assignments could be formed using uninstantiated objects.

Finally, one could use a proxy *b* in *w* for an object *a* that does not exist at *w* and form variable assignments with proxies. All the worlds of a pw-model have an infinite domain, so this is not obviously ruled by cardinality constraints.[7]

I do not take a stance here. The problem just outlined is supposed to demonstrate how the additional expressive strength permits a discussion of issues of metaphysics in the object language. In first-order quantified modal logic the problem does not arise, because it is moved entirely into the metalanguage and there problems are solved by using a purely extensional, non-modal metatheory. One may even consider replacing modal talk with purely extensional discourse about possible worlds, as Lewis (1968, 1986) did. Of course, I cannot analyze such alternative approaches here, but one of the criteria would be whether they reach the expressive strength of the fourth grade of modal involvement.

9.6 The Road Ahead

I have outlined one way to capture the strength of the fourth grade of modal involvement in a formal language containing a binary predicate Nec and with a possible worlds semantics. But this is only one way to capture the full strength of the fourth grade of modal involvement. There are various ways to deviate from the approach in this chapter, and I have only briefly sketched some reasons for my choices. However, I hope I have succeeded in conveying a taste of how the fourth grade of modal involvement can help to shed light on issues such as the discussion between actualists and possibilists. The fourth grade of modal involvement may not open a Cantorian paradise, but at least it provides an expansive playground not only for metaphysics, but also for the analyses of various epistemic, alethic, logical, and further modalities.

Acknowledgments

I am grateful to Beau Mount, Simon Nagler, Carlo Nicolai, Johannes Stern, Philip Welch, and Timothy Williamson for useful hints and suggestions.

Notes

1. Quine used 'Nec' (uppercase) for the predicate and 'nec' (lowercase) for the operator. Here the latter is replaced with the more familiar $\Box$.
2. For a recent detailed discussion of modal predicates see Stern (2016). It has been argued that the full force of modal discourse can be restored by using propositional or 'substitutional' quantification or truth predicates (Halbach and Welch, 2009). Some remarks on such extensions are given in the next section.
3. Here I refer to "Three Grades of Modal Involvement" only. Later, in *Word and Object* (1960, §41) he considers *de re* modalities as predicates; and earlier in §35 and in (1956) he discusses propositional attitudes *de re*.
4. Instead of predicates of sentences we could use predicates of Russellian propositions. If we are able to express the operation of applying the property P to the object o, we can use a predicate to say that there is a property P and an object o such that the proposition resulting of applying P to o is necessary. This approach has its own problems, because now *de dicto* modalities are not directly expressible. I will not pursue this strategy any further without claiming that it is not viable. A more detailed discussion would require a formalized theory of Russellian propositions.
5. When Quine talks about a ordering of grades of modal involvement, he does not have an ordering according to reducibility in our sense in mind. As mentioned above, a semantic predicate is the lowest form of modal involvement for Quine.
6. Carnap's translations is different and there are several variations. What caused problems especially in early variants were problems with the iteration of Nec.
7. I thank Beau Mount for bringing this point to my attention.

References

Asher, N. and Kamp, H. (1989). Self-reference, attitudes, and paradox. In Chierchia, G., Partee, B. H., and Turner, R., editors, *Properties, Types and Meaning*, volume 1, pages 85–158. Kluwer.

Bealer, G. (1982). *Quality and Concept*. Clarendon Press.

Bealer, G. (1993). Universals. *Journal of Philosophy*, 90: 5–32.

Bealer, G. (1998). Universals and properties. In Laurence, S. and Macdonald, C., editors, *Contemporary Readings in the Foundations of Metaphysics*, pages 131–147. Blackwell Publishers.

Bull, R. (1969). On modal logic with propositional quantifiers. *Journal of Symbolic Logic*, 34: 257–263.

Carnap, R. (1934). *Logische Syntax der Sprache*. Springer.

Fine, K. (1970). Propositional quantifiers in modal logic. *Theoria*, 36: 336–346.

Forster, T. (2005). The modal aether. In Kahle, R., editor, *Intensionality*, pages 20–41, A K Peters.

Fujimoto, K. and Halbach, V. (2019). *Classical Determinate Truth*. Draft.

Garson, J. (2001). Quantification in modal logic. In *Handbook of Philosophical Logic*, pages 267–323. Kluwer Academic Publishers.

Halbach, V. and Leigh, G. (2021). *The Road to Paradox: A Guide to Syntax, Truth, and Modality*. Cambridge University Press, reprinted in Martin.

Halbach, V., Leitgeb, H., and Welch, P. (2003). Possible worlds semantics for modal notions conceived as predicates. *Journal of Philosophical Logic*, 32: 179–223.

Halbach, V. and Sturm, H. (2004). Bealers Masterargument: Ein Lehrstück zum Verhältnis von Metaphysik und Semantik. *Facta Philosophica*, 6: 97–110.

Halbach, V. and Welch, P. (2009). Necessities and necessary truths: A prolegomenon to the metaphysics of modality. *Mind*, 118: 71–100.

Halbach, V. and Zhang, S. (2016). Yablo without Gödel. *Analysis*, 76: 53–59.

Kaplan, D. (1970). S5 with quantifiable propositional variables. *Journal of Symbolic Logic*, 35: 355. Abstract.

Kripke, S. A. (1959). A completeness theorem in modal logic. *Journal of Symbolic Logic*, 24: 1–14.

Kripke, S. A. (1975). Outline of a theory of truth. *Journal of Philosophy*, 72: 690–716. reprinted in Martin (1984).

Lewis, D. (1968). Counterpart theory and quantified modal logic. *Journal of Philosophy*, 65: 113–126.

Lewis, D. (1986). *On the Plurality of Worlds*. Blackwell Publishers. Malden, MA.

Martin, R. L., editor (1984). *Recent Essays on Truth and the Liar Paradox*. Clarendon Press and Oxford University Press.

McGee, V. (1985). How truthlike can a predicate be? A negative result. *Journal of Philosophical Logic*, 14: 399–410.

Montague, R. (1963). Syntactical treatments of modality, with corollaries on reflexion principles and finite axiomatizability. *Acta Philosophica Fennica*, 16: 153–167. Reprinted in (Montague, 1974, 286–302).

Montague, R. (1974). *Formal Philosophy: Selected Papers of Richard Montague*. Yale University Press. Edited and with an introduction by Richmond H. Thomason.

Quine, W. V. O. (1956). Quantifiers and propositional attitudes. *Journal of Philosophy*, 53: 177–187.

Quine, W. V. O. (1960). *Word and Object*. MIT Press. Cambridge, MA.

Quine, W. V. O. (1976). Three grades of modal involvement. In *The Ways of Paradox*, pages 158–176. Harvard University Press, Cambridge, MA, revised and enlarged edition.

Stern, J. (2014a). Modality and axiomatic theories of truth I: Friedman-Sheard. *Review of Symbolic Logic*, 7: 273–298.

Stern, J. (2014b). Modality and axiomatic theories of truth II: Kripke-Feferman. *Review of Symbolic Logic*, 7: 299–318.

Stern, J. (2016). *Toward Predicate Approaches to Modality*, volume 44 of Trend in Logic. Springer.

Tarski, A. (1935). Der Wahrheitsbegriff in den formalisierten Sprachen. *Studia Philosophica Commentarii Societatis Philosophicae Polonorum*, 1: 261–405. Translated as 'The Concept of Truth in Formalized Languages' in (Tarski, 1956, 152–278); page references are given for the translation.

Tarski, A. (1956). *Logic, Semantics, Metamathematics: Papers from 1923 to 1938*. Clarendon Press.

Visser, A. (1989). Semantics and the liar paradox. In Gabbay, D. and Günthner, F., editors, *Handbook of Philosophical Logic*, volume 4, pages 617–706. Reidel, Dordrecht.

Williamson, T. (2013). *Modal Logic as Metaphysics*. Oxford University Press.

Yablo, S. (1985). Truth and reflection. *Journal of Philosophical Logic*, 14: 297–349.

Yablo, S. (1993). Paradox without self-reference. *Analysis*, 53: 251–252.

10 Opacity and Paradox

Andrew Bacon

In 1961 Prior proved a theorem that places surprising constraints on the logic of intentional attitudes, like ‘thinks that’, ‘hopes that’, ‘says that’ and ‘fears that’. Paraphrasing it in English, and applying it to ‘thinks’, it states: If, at *t*, I thought that I didn’t think a truth at *t*, then there is both a truth and a falsehood I thought at *t*.

In this paper I explore a response to this paradox that exploits the opacity of attitude verbs, exemplified in this case by the operator ‘I thought at *t* that’, to block Prior’s derivation. According to this picture, both Leibniz’s law and existential generalization fail in opaque contexts. In particular, one cannot infer from the fact that I’m thinking at *t* that I’m not thinking a truth at *t*, that there is a particular proposition such that I am thinking it at *t*. Moreover, unlike some approaches to this paradox (see Bacon et al. [4]) the failure of existential generalization is not motivated by the idea that certain paradoxical propositions do not exist, for this view maintains that there is a proposition that I’m not thinking a truth at *t*. Several advantages of this approach over the non-existence approach are discussed, and models demonstrating the consistency of this theory are provided. Finally, the resulting considerations are applied to the liar paradox, and are used to provide a non-standard justification of a classical gap theory of truth. One of the main challenges for this sort of theory—to explain the point of assertion, if not to assert truths—can be met within this framework.

10.1 Prior’s Paradox

Prior’s result may be formalized as follows. Let Q represent a unary propositional operator and X be a sentential variable:

Prior’s Theorem $Q\forall X(QX \to \neg X)\to\exists X(X \land QX) \land \exists X(\neg X \land QX)$

The result is general, in that no special principles about the operator Q are assumed in its proof. Thus it has instances in which Q is replaced with negation or metaphysical necessity. But these instances are unsurprising. Prior is primarily concerned with instances in which *Q* is substituted for various intentional attitudinal verbs, such as '*S* fears that', '*S* hopes that', '*S* said that', and so on. Reading *Q* as 'Simon said at *t* that', Prior would paraphrase his theorem as follows:

> If Simon said at *t* that Simon didn't say anything true at *t*, then Simon has said a truth and a falsehood at *t*.

The puzzle is this: it seems evident that Simon could have said, at *t*, that he didn't say anything true at *t*, and not have said anything else. In which case he would not have said both a truth and a falsehood, he would have said at most one thing, contradicting the theorem.

Prior's proof follows reasoning that should by now be familiar from the liar, and related paradoxes. Suppose Simon had said, at *t*, that he didn't say anything true at *t*. Then either he didn't say anything true at *t*, or he did. Suppose the former. Then, since the proposition that he didn't say anything true at *t* is a proposition he said at *t*, it must be false. So it's not the case that he didn't say a truth at *t*. Contradiction. Suppose the latter: he did say a truth at *t*. Then the proposition that he didn't say a truth at *t*, is itself false, and thus a falsehood he said at *t*. So he has said a truth and a falsehood at *t*.

The informal reasoning above belies the inevitability of Prior's result. In order to appreciate this, some points about the relation between Prior's theorem, and the subsequent paraphrase in English, are in order. (Prior's formal proof will be stated in the next section.) First, Prior's theorem is stated in propositionally quantified logic: a language containing the familiar truth functional connectives, propositional letters, the unary operator *Q*, but also variables that can occupy the position of sentences, and quantifiers that can bind them. The English paraphrase replaces quantification into sentence position with singular quantification over propositions. Since a first-order variable cannot grammatically occupy the position of a sentence, occurrences of bound sentential variables are replaced by first-order variables as arguments to a propositional truth predicate.

Second, because of the use of the propositional truth and falsity predicates, certain avenues for resisting Prior's theorem that are suggested from inspection of the informal argument, are really illusory. This includes, for example, co-opting any of the familiar non-disquotational accounts of sentential truth to play the role of propositional truth, allowing propositions to be neither true nor false, or both true and false, but without relinquishing classical logic. For Prior's paradox is not really stated in terms of propositional truth or falsity. Where, in the paraphrase, we would assert that the proposition that *P* is true, the target of the paraphrase would simply assert *P*. And where, in the informal argument, we seem to move freely between 'the proposition that *P* is true' and '*P*', no such move is made in Prior's official argument.

Other familiar responses to the liar paradox are also of no use here. For example, after proving that a given language *L*, that can talk about its own sentences, does not contain a predicate that is true of all and only the true sentences in *L*, some will emphasize that the matter is not hopeless, since there might be another, more expressive language, L^+, that *can* express truth in *L*. This goes along with the general idea that everything is in principle expressible, but not all at once in the same language. Yet Prior's theorem purports to tell us that there is a particular proposition—that Simon didn't say anything true at *t*—that Simon cannot say uniquely at *t*. The theorem doesn't have qualifications—that he can't say the proposition in this particular language or that—it says that he can't say it uniquely at *t at all.* Thus he can't say it with a sentence of English, of Russian, a more expressive extension of these languages, or with an elaborate round of charades, unless he also says something else along with it.

Of course, the paraphrase in terms of first-order quantification is necessary because we do not have the equivalent of pronouns for sentences, or quantifiers that can bind them in ordinary English. But quantification into other grammatical positions, aside from first-order quantification, is possible in English, including the positions of plural and prepositional phrases, suggesting this restriction is only accidental to English (Prior, 1971; Rayo and Yablo, 2001; Boolos, 1984). And even if not, why think that quantification into sentence position is intelligible only if it can be translated into a language that already has it? As children we are evidently able to learn first-order quantification in our native tongue, without first translating it into an already understood language, so why should these other modes of quantification be any different (Williamson, 2003)?

Might one nonetheless try to resist Prior's theorem by denying the intelligibility of quantification into sentence position? Those who do so will find the first-order paraphrases on better footing, but might reject the reasoning as illicit, since it involves principles concerning propositional truth that are extremely contentious in the context of sentential truth. For instance, we have freely moved between 'the proposition that *P* is true' and '*P*', in order to be faithful to Prior's reasoning, and this looks similar to the move, that is illicit given classical logic, between 'the sentence '*P*' is true' and '*P*'. But unlike sentential truth, these inferences are unproblematic in the propositional setting (and only become problematic when combined with a structured theory of propositions, or some other theory of propositional granularity in which propositions behave enough like sentences to define things like substitution and diagonalization).[1] Indeed, a minimal theory of first-order propositions and truth can be consistently formulated in which all of Prior's reasoning may be faithfully represented.[2] It is most naturally formulated in a type theory, à la Church, that contains the type of first-order predicates

$e \rightarrow t$, expressions that combine with names to form sentences, and a converse type $t \rightarrow e$ of expressions that combine with sentences to form names. Given two primitives, *true* of type $e \rightarrow t$, and *that* of type $t \rightarrow e$, a minimal theory of first-order propositions can be formalized by an axiom stating that *true* is a left-inverse of *that* (in the type theory, $\lambda X.true(that\ X)\ X = \lambda X.X$). The consistency of such a theory roughly amounts to there not being 'more things' of type t than of type e, as in a full model of type theory *that* can be interpreted by any injection in type $t \rightarrow e$, and *true* by any of its left-inverses.[3] Despite being exceedingly minimal—the theory does not prove that Julius Caesar and the proposition that grass is green are distinct, for example—I think the theory is suitable for almost all of the things one might want from a first-order theory of propositions. In particular, it licenses the intersubstitutivity of 'the proposition that P is true' with 'P'.[4]

10.2 Possible Responses to Prior's Paradox

Having accepted the intelligibility of Prior's language, here, finally, is Prior's proof. It has two assumptions:

> **Classical Propositional Logic** All classical tautologies, Modus Ponens.
>
> **Universal Instantiation** $\forall\, XA \rightarrow A[B/X]$, provided B is substitutable for X in A.

Propositionally quantified logic has further axioms beyond Universal Instantiation, but it is the only principle appealed to in Prior's proof. We also take $\forall$, $\neg$ and $\rightarrow$ as primitive, and define $\exists X\ A$ as $\neg\forall X\neg A$ and $\wedge$ as $\neg(A \rightarrow \neg B)$.[5] Here is Prior's proof, translated from Polish notation, and annotated so as to use only the assumptions above.

1. $\forall X(QX \rightarrow \neg X) \rightarrow (Q\forall X(QX \rightarrow \neg X) \rightarrow \neg\forall X(QX \rightarrow \neg X))$ (UI)
2. $Q\forall X(QX \rightarrow \neg X) \rightarrow (\forall X(QX \rightarrow \neg X) \rightarrow \neg\forall X(QX \rightarrow \neg X))$ (1, CL)
3. $Q\forall X(QX \rightarrow \neg X) \rightarrow \neg\forall X(QX \rightarrow \neg X)$ (2, CL)
4. $Q\forall X(QX \rightarrow \neg X) \rightarrow \exists X(QX \wedge X)$ (3, CL, $\exists$Def)
5. $Q\forall X(QX \rightarrow \neg X) \rightarrow (Q\forall X(QX \rightarrow \neg X) \wedge \neg\forall X(QX \rightarrow \neg X))$ (3, CL)
6. $(Q\forall X(QX \rightarrow \neg X) \wedge \neg\forall X(QX \rightarrow \neg X)) \rightarrow \exists X(QX \wedge \neg X)$ (CL, $\exists$Def, $\wedge$ Def)
7. $Q\forall X(QX \rightarrow \neg X) \rightarrow \exists X(QX\wedge \neg X)$ (5, 6, CL)
8. $Q\forall X(QX \rightarrow \neg X) \rightarrow (\exists X(QX\wedge X)\wedge\exists X(QX\wedge \neg X))$ (4, 7, CL)

If we are to avoid Prior's conclusion, then we must reject one of the two assumptions. That is, one must either weaken classical propositional logic, or reject the principle of Universal Instantiation. (Given the duality of the quantifiers and the first assumption, Universal Instantiation is equivalent to Existential Generalization: $A[B/X] \rightarrow \exists X\, A$ provided B is substitutable for X in A. We will treat these two equivalent formulations interchangeably.)

Both options have precedents. Weakening classical logic is a common response to other paradoxes, including the semantic and set theoretic paradoxes.[6] And there is a tradition of rejecting Universal Instantiation going back to Russell's ban on impredicativity: a quantified proposition cannot belong to the domain over which it quantifies.[7] Since propositional quantification is thus necessarily restricted, and since the logic of restricted quantification does not include Universal Instantiation, Universal Instantiation is not valid.

Of course, another option is to simply accept the logic of Prior's argument and follow it where it leads. Indeed, this is where my own sympathies lie (Bacon, 2019). A radical interpretation of Prior's theorem along these lines is that when Simon attempts to utter the sentence 'Simon didn't say anything true at *t*', a mysterious force prevents him from completing his utterance. But this seems like a far-fetched moral to draw, and luckily one can accept Prior's logic without drawing it. Rather, it is perfectly possible for Simon to make the utterance of the *sentence* 'Simon didn't say anything true at *t*', but in doing so he doesn't succeed in saying that Simon didn't say anything true at *t*. According to Prior's own interpretation, he doesn't succeed in saying anything at all. This interpretation is not that radical, for few people would wish to identify saying that *P* with making the sounds that constitute uttering the sentence '*P*'. As a way of warming up to this idea, note that one might cough in a way that, by fluke, sounds exactly like how one would say that snow is white in Farsi. But few would regard this as a way of saying that snow is white. More realistically, one might utter the sentence 'that dog looks funny' while failing to point out a dog, or even anything, and thus fail to say something with a sentence that, in other circumstances, *could* be used to say something. But Prior's theorem, as we have noted already, is quite general and there are many other instances for which the costs are higher. For instance, consider:

> If Simon tried to say, at *t*, that Simon didn't try to say anything true at *t*, then he tried to say a truth and a falsehood at *t*.

Suppose that at *t* Simon utters the sentence 'I didn't try to say anything true at *t*' (with full understanding of English, the intentions to assert, etc), and does nothing else. If we apply Prior's diagnosis here, we should say that when Simon utters the sentence 'I didn't try to say anything true at *t*' he fails to say anything at all. But this response is of no avail: one must deny that Simon even *tried* to say something. For if we concede that he tried to say that he didn't try to say anything true at *t*, then we must accept the absurd conclusion that there were two things he tried to say at *t*.[8]

In fact, accepting Prior's theorem at face value commits us to one of the following (whether we follow Prior's diagnosis of the original paradox or not): (i) that Simon didn't even try to say that he didn't try to say anything true at *t*, despite strong appearances to the contrary, or (ii) that Simon tried to say two things, one false one true, again, despite strong appearances to the contrary. So the cost of accepting Prior's theorem is high, and a thorough examination of the logic that ensures it seems like the responsible course of action, even if we ultimately choose to accept the conclusion.

10.3 Non-Existent Propositions

What should the Universal Instantiation denier say about Simon's utterance? According to the most plausible theories that accept Prior's theorem, uttering the sentence '*P*' is not sufficient for having said that *P*. Indeed, we have argued that it cannot be sufficient for having even *tried* to say that *P*.

To distance themselves from the pathological results of the fully classical view, a denier of Universal Instantiation ought to endeavour to uphold what I will call the *naïve model of speech*. According to this view, there is an action one can perform freely to a sentence '*P*', which I shall call uttering '*P*', which does suffice for saying that *P*, and moreover when one utters nothing else beyond '*P*', one says nothing apart from that *P*, in normal circumstances.[9] Of course, the naïve view will have to be qualified in various ways to take into account context sensitivity; but we will set that aside by restricting attention to the pared down language described in Section 10.2, and by imposing the working assumption that its expressions are context insensitive.[10] Such a view can accept all of the following: that uttering '*P*' typically suffices for one having tried to say that *P* (bracketing context sensitivity), and that typically, when one tries to say that *P* one says that *P*.

What does this amount to with respect to Prior's paradox? According to the naïve model, when Simon utters the sentence 'I didn't say anything true at *t*' he succeeds in saying that he didn't say anything true at *t*. But we cannot apply Existential Generalization (an equivalent of Universal Instantiation) to conclude that there is something that Simon has said:

1. Simon said, at *t*, that Simon didn't say anything true at *t*.
2. For no *X* did Simon say that *X*.

This, of course, blocks the argument. For instance, at one point of the argument we had supposed that Simon didn't say anything true at *t*, and attempted to show, contrary to the assumption, that this proposition is itself a truth that Simon said. Given 2 we should indeed accept this supposition: it is vacuously true, because Simon hasn't said anything. We can also maintain 1, that Simon said that Simon didn't say anything true at *t*; we just can't conclude from this that Simon said a truth—this is an instance of Existential Generalization that is rejected.

One way of understanding 2 is as an instance of a *no-proposition* view. Like the fully classical logician, this theorist accepts that Simon uttered the sentence 'I didn't say anything true at *t*' but that he didn't succeed in saying anything. But it is a different kind of no-proposition view, because despite this he did succeed in saying that he didn't say anything true at *t*.

There are a couple of precedents for this sort of view. One, mentioned already, stems from Russell's ban on vicious circles, according to which a quantified proposition cannot belong to the domain over which it quantifies.[11] A consequence of this ban is that all quantification into sentence position is restricted quantification, for the range of the quantifier in a quantificational claim like $\forall X\ A$, must be restricted in such a way that it does not range over $\forall X\ A$ itself. Thus an instance of Universal Instantiation that instantiates X with $\forall X\ A$ is not legitimate (the instance: $\forall X\ A \rightarrow A[\forall X\ A/X]$). More generally, we know that the logic of restricted quantification does not include the principle of Universal Instantiation, as, after all, the following inference concerning the restricted quantifier 'every *F*' is clearly invalid:

$$\frac{\text{Every } F \text{ is } G}{\text{Therefore, } a \text{ is } G}$$

Theorists sympathetic to Russell's ban on impredicativity generally pursue ramified approaches to propositional quantification, in which there is a hierarchy of different quantifiers, each quantifier restricted to ranging over propositions involving quantifiers from lower in the hierarchy. Each quantifier is restricted, but each proposition is in the range of some quantifier or other.[12]

Another precedent for rejecting Universal Instantiation arises in the treatment of empty names within free logic.[13] According to common sense, there is no such thing as Pegasus, the winged horse-god of Greek mythology. But we still want to maintain that Pegasus is a mythical winged horse, or that the ancient Greeks told stories about Pegasus. One can also look (unsuccessfully) for Pegasus, and believe (mistakenly) that Pegasus is a horse, or that he would have been a horse-god had the mythology been true. We seem to want to assert that there are no mythological horse-gods, but at the same time assert that Pegasus is a mythological horse-god, that there is nothing you are in fact looking for when you are looking for Pegasus, and so on. But these are counterexamples to Existential Generalization (and thus Universal Instantiation).

On this interpretation we do not conceive of the failures of Universal Instantiation arising because the quantifiers are restricted in some way. For according to a view that takes the above judgments at face value, there is not some more expansive quantifier that ranges more widely over things like Pegasus, Sherlock Holmes, Vulcan, and so on. The counterexamples to Universal Instantiation, from this perspective, involve the *unrestricted* quantifiers.[14]

This perspective on Universal Instantiation can be applied to Prior's paradox. For instance, verbs like *says*, *hopes*, *fears*, and so on, create the sorts of contexts from which one cannot existentially generalize. They are not *existence entailing*: to illustrate, *being a horse* is existence entailing, since we may infer from the (false) assumption that Pegasus is a horse that he exists, whereas *being a mythical horse* is not, for the analogous entailment seems false. It seems like Simon could hope that Pegasus was real, but we wouldn't want to infer that there is something that Simon hopes is real. Similarly, he could have said that Pegasus is real, without there being anything that he said is real. One might therefore postulate things that stand to the grammatical category of sentences as empty names stand to the category of singular terms, and have the same attitude towards them as the free logician has towards empty names. Thus, one can say that *P* without there being anything that one has said. The proposition that Simon didn't say a truth at *t* must be like Pegasus in the sense that it can have certain non-existence entailing properties, like *being said* or being *hoped* (or, in the latter case, *being said/hoped to be real*) without existing.

The former view is less naturally classified as a no-proposition view. For each level of the quantifier hierarchy, *i*, there is a paradoxical sentence Simon can utter for which he can truthfully report, of his paradoxical utterance, both 'Simon hasn't said anything$_i$' and 'there is something$_{i+1}$ that Simon said'.[15] And both views should be distinguished, as noted above, from the classical no-proposition view which maintains that, in addition to not saying anything at *t*, Simon hasn't said that he didn't say anything true at *t*.

Do these views avoid the trappings of the classical no-proposition view? A common problem for adherents of the classical no-proposition theory is that they have trouble stating their own view. For instance, someone like Prior might wish to communicate their view about Simon's utterance at *t* by making their own utterance of the following: 'Simon didn't say anything true at *t*', perhaps by clarifying 'because he didn't say anything at all'. But of course, this is an utterance of the very sentence Simon failed to say anything with. Prior's response is that, while one utterance of a sentence might fail to say anything, another can. Thus it is utterances, not sentences, that are the true couriers of successful assertion. But this response is insufficiently general, as variants can be formulated that do not fair so well: D = 'no utterance of D can be used to say a truth and only truths'. A routine argument, mimicking Prior's, shows that no utterance of D can be used to say truths and only truths. But the theorist might want to communicate this result as well, by uttering the sentence 'no utterance of D can be used to say a truth and only truths'. In virtue of being an utterance of the problematic sentence itself, they will not succeed in saying anything.[16]

In this regard the non-classical no-proposition views do better. For they can maintain that, even though an utterance of 'no utterance of D can be used to say a truth and only truths' will not result in you saying anything, one might still end up saying, with such an utterance, that no utterance of D can be used to say a truth and only truths. What good is this, if one hasn't said anything? Well, the purpose of assertion is presumably to pass on beliefs and knowledge to others which can inform their actions, and so forth. But intentional verbs like belief and knowledge are paradigm instances of the sorts of words that create non-existence entailing contexts: for example, Alice may believe that Pegasus is real, even though there isn't anything that she believes is real.[17] Thus in saying that P one may pass on knowledge that P, or a belief that P, even if there is no proposition you've said and which your intended audience thereby knows or believes. One might worry that the intentionality will eventually peter out. For instance, if the beliefs and knowledge acquired are to be efficacious, they had better inform our actions, but whether one has done something seems to be an existence entailing context, as it appears to be objective and not tangled up with our attitudes. Call a belief existent if, for some P it is a belief that P, and non-existent otherwise. Non-existent beliefs, by impacting your credences, might raise the value of existing propositions which you are in a position to make true. So non-existent beliefs can be efficacious, even if we assume that *making true* is an existence entailing context. That said, the notion of making a proposition true is arguably also a propositional attitude, albeit a factive one. It is not simply a matter of certain things happening, but of them happening as a result of the agent's intentions, and is thus like knowledge, belief, and saying in the relevant respects. For instance, given the naïve model of saying, Simon appears to be in control of what he says and doesn't say, as he is certainly in control of his utterances. When Simon utters the sentence 'nothing I've said at t is true' he doesn't say anything true. Since he was responsible for this, we should maintain that Simon has made it the case that nothing he said at t is true. Thus making true is not existence entailing, given the assumption, currently being explored, that there does not exist a P identical to the proposition that nothing Simon said at t is true.[18]

In order to uphold the naïve model of speech in full generality, we need not only to accommodate the case in which Simon utters 'nothing I said at t is false'. For he could have said any other sentence at t. For instance, if he uttered 'snow is white' then he would have said, instead, that snow is white. Indeed, for any sentence A, 'it is possible that Simon said that A' should be true given the naïve model of saying, since it's possible that Simon utter 'A'. Augmenting the language in Section 10.2 with an operator $\Diamond$, representing possibility, we may formulate this as a schema:

Possible Saying $\Diamond QA$

Note, however, that the naïve model of saying motivates something stronger: for given that it's possible that Simon utter 'snow is white' and nothing else, it should be possible that Simon say that snow is white without saying anything else. If we add a propositional identity connective, =, we might try to formulate this as follows:[19]

$$\Diamond(QA \wedge \forall X(QX \rightarrow X = A))$$

But this seems insufficiently strong without Universal Instantiation. For instance, this principle is consistent with the hypothesis that Simon can't say at t that snow is white, without also saying that nothing he said at t is true. $\forall X(QX \rightarrow X = A)$ entails that any *existing* proposition which Simon has said at t must be identical to the proposition that snow is white. But this is satisfied when Simon says both that snow is white and that nothing Simon said at t is true, since the latter is not an existing proposition. In Bacon et al. (2016) a primitive notion of saying uniquely, $Q!$ is introduced to deal with this: it is not defined in terms of Q, the quantifiers and an identity connective, but is taken as basic, and characterized by a model theoretic stipulation (more on this below), and moreover subject to the laws:

Subsumption $\Box(Q!A \rightarrow QA)$
Uniqueness $\Box(Q!A \wedge Q!B \rightarrow A = B)$

We may thus articulate the desired component of the naïve model as follows:[20]

Possible Saying! $\Diamond Q!A$

It is important to this project to know that Possible Saying!, Subsumption, and Uniqueness are indeed consistent with a free version of propositionally quantified logic that does not contain Universal Instantiation. This is undertaken in Bacon et al. (2016), where various models of this, and related theories, are constructed.

10.4 Opacity

In this section I will be exploring another view that responds to Prior's paradox by rejecting Universal Instantiation. But it is a non-standard version of this response: it is not motivated by any general logical constraint, such as Russell's ban on impredicative quantification. Nor does it follow from considerations of non-existence, as the counterexamples involving names like Pegasus. In fact, our quantifications will in general be interpreted as ranging unrestrictedly, and we will accept a schema, $\exists X(X = A)$, to the effect that all sentences express existing propositions: there are no sentences that stand to the propositional quantifiers as empty names stand to the first-order quantifiers.

The key observation is that the problematic instances of Prior's theorem all involve intentional attitudes—attitudes like saying, fearing, hoping, and so forth—that are commonly supposed to create *opaque* contexts. We have already noted that intentional attitudes create contexts which are not existence entailing. I will argue that opacity also creates contexts in which existential generalization is not permissible, albeit for fundamentally different reasons than in the cases involving non-existence, and restricted quantification.

Suppose that Simon knows that Hesperus is visible in the evening and Phosphorus is visible in the morning, but does not realize that Hesperus and Phosphorus are the same. He is looking at the sky in the evening and forms the belief that Hesperus is bright:

1. Simon believes that Hesperus is bright.

Because he does not realize they are the same, he forms no such belief about Phosphorus:

2. Simon does not believe that Phosphorus is bright.

But, of course, they are the very same planet:

3. Hesperus is Phosphorus.

This seems to be a counterexample to Leibniz's law:

$$\mathrm{L}()\ a = b \to A \to A[a/b]$$

There are many different replies to this puzzle. Most assimilate it to some sort of equivocation. According to one version of this idea, words like 'believes' are context sensitive, and the attitude being ascribed by the word 'believes' are different in 1 and 2.[21] According to another, the equivocation is due to the names 'Hesperus' and 'Phosphorus': in embedded contexts names refer not to their customary referents, but to their customary senses (Frege, 2010).

In Bacon and Russell (2019), Jeff Russell and I have investigated the suggestion that we take the judgments 1–3 at face value, as counterexamples to Leibniz's law. But this is entirely consistent with various quantified variants of Leibniz's law, including:

L(xy) $\forall x \forall y(x = y \rightarrow A \rightarrow A[x/y])$

(Sometimes principles like $L()$ and $L(xy)$ are restricted to instances involving direct predications; in this case, for instance, $\forall xy(x = y \rightarrow Fx \rightarrow Fy)$. Following Bacon and Russell (2019) I shall treat these principles interchangeably in the presence of a device of λ-abstraction that allows one to turn a term t occurring in an arbitrary context $A[t/x]$, into a direct predication $(\lambda x\ A)t$.[22] In our discussion of first-order logic, where λ is not present, the unrestricted principles are stronger, and will be our focus.)

$L(xy)$ has been thought to be plausible, even by those who reject $L()$ on broadly Fregean grounds. For while *names* may be associated with interesting senses, which can have interestingly different cognitive import, bound variables refer directly to the things they denote. Let Bx represent the open sentence 'Simon believes x to be bright'.[23] Then we should accept:

$$a = b \wedge Ba \wedge \neg Bb$$

But given Existential Generalization we would be able to infer:

$$\exists x \exists y(x = y \wedge Bx \wedge \neg By)$$

But this directly contradicts an instance of $L(xy)$. So Existential Generalization cannot be valid, and thus neither can Universal Instantiation. Indeed, this is evident when one observes that $L()$ can be derived from $L(xy)$ by two applications of Universal Instantiation (instatiating x with a and y with b).

It has long been suggested that opaque contexts generate failures of Existential Generalization (Quine, 1956; Kaplan, 1968). For instance, many people believe that a single person, who has since been dubbed 'Jack the Ripper', terrorized Whitechapel with a series of murders in the 1888. Thus we believe that Jack the Ripper committed the Whitechapel murders. But we don't know who committed the murders, so there is no person such that we believe that *they* committed the Whitechapel murders.

It is worth emphasizing how these failures of Existential Generalization differ from the failures due to non-existence. For in the first case, there is something, namely the planet Venus, that is identical to both Hesperus and Phosphorus. Hesperus and Phosphorus *exist*. So, too, does Jack the Ripper, assuming, as commonly supposed, that there was a single individual who committed the murders in 1888. So the failures of Existential Generalization are fundamentally different from the failures instanced in inference from ‘Pegasus is a mythical winged horse’ to ‘something is a mythical winged horse’, since in that case there is no such thing as Pegasus. That is, we have a counterexample to a principle, weaker than Existential Generalization, that is validated in the free logics designed to deal with empty-names (see, e.g., Nolt, 2018):

> **Unrestricted Export** $\exists x\; a = x \wedge A[a/x] \rightarrow \exists xA$

Unrestricted Export is derivable given some pretty uncontroversial quantificational logic, and Leibniz’s law, $L()$.[24] But since Leibniz’s law is part of what is at stake in these cases, this argument has little suasive force.

Let’s summarize this with an important definition.

> **Opacity** A context, A, is *transparent*, relative to a language, iff every instance of the schema $a = b \rightarrow A \rightarrow A[b/a]$ in that language is true. It is *opaque* otherwise.

There are philosophers who will be in verbal agreement with much of the foregoing. They will countenance failures of $L()$ (but not $L(xy)$), and will subsequently agree that there are opaque predicates and predicates that do not license export. But they do not take opacity ‘seriously’ in my sense, for they will maintain that attitude verbs don’t create genuine contexts, whose meanings can be computed compositionally, any more than quotation marks do. What, on the surface, looks like at operator expression—‘Simon believes that …’—in reality functions like a predicate of sentences, ‘Simon believes “…”’, and what looks like a context is merely a typographical constituent of a quotation name for a sentence. Provided such philosophers can make sense of my remarks in their preferred framework, it will not matter much, but I will not make systematic attempts to accommodate them.

Putting all this together, what might a first-order logic of opacity look like? We will assume a first-order language with non-logical predicates and operators that may be opaque. Let’s start with the quantificational fragment:

> **Quantified Instantiation** $\forall x(\forall yA \rightarrow A[x/y])$
> **Normality** $\forall x(A \rightarrow B)\rightarrow\forall xA \rightarrow \forall xB$
> **Quantifier Exchange** $\forall x\forall yA \rightarrow \forall y\forall xA$
> **Vacuous Quantification** $A \rightarrow \forall xA$ when x does not occur free in A

Notice that in place of the principle of Universal Instantiation, $\forall xA \rightarrow A[t/x]$, we have its universal closure, Quantified Instantiation.[25] This corresponds to the idea, implicit in our discussion of $L(xy)$, that bound variables, unlike names, are not susceptible to identity confusions. Thus one ought to be able to instantiate bound variables.[26]

Besides Quantified Instantiation, one should also think that instantiation of arbitrary terms is legitimate provided it is not in an intentional context. The distinction between intentional and non-intentional contexts is not a distinction that can be made prior to the interpretation of the non-logical constants. But we may assume purely logical contexts are transparent, independently of how the non-logical constants are interpreted, and so we should be able to instantiate into purely logical contexts. More generally, we should be able to instantiate even in non-logical contexts, provided the variable is not within the scope of a non-logical operation:[27]

Logical Instantiation $\forall yA \rightarrow A[t/y]$ provided y is not in the scope of a non-logical operation.

To this we must add principles governing identity:

L(xy) $\forall xy(x = y \rightarrow A \rightarrow A[y/x])$
Reflexivity $\forall xx = x$
Existence $\exists xt = x$

From Logical Instantiation and identity we can derive the reflexivity schema, $a = a$. And from $L(xy)$ and Logical Instantiation we can derive every instance of Leibniz's law in which the substituted terms do not occur in non-logical contexts. This allows one to derive the symmetry and transitivity schemas in the usual way: $a = b \rightarrow b = a$ and $a = b \rightarrow b = c \rightarrow a = c$.[28] Existence corresponds to the idea that failures of Universal Instantiation are not due to non-existence. I will not adjudicate on the more general issue of Existence when there are empty names in the language. For now we may simply treat it as a stipulation that the language does not contain any empty names.

It is worth pausing, for a minute, to consider whether the existence of opacity *forces* one to reject Existential Generalization, and Unrestricted Export. Part of our argument for these failures rested on the assumption of $L(xy)$, which someone taking opacity at face value may wish to deny. A more theoretical argument for $L(xy)$ is given in Bacon and Russell (2019) from a higher-order generalization, where X is a variable taking predicate position:

L(xyX) $\forall X \forall x \forall y(x = y \rightarrow Xx \rightarrow Xy)$

$L(xyX)$ can be justified as follows: if it were false, one could find x and y that are identical, but are distinguished by a property. In which case, there would be a relation that *did* guarantee that x and y shared the same properties, namely the relation of *sharing the same properties* (defined: $\lambda xy\forall X(Xx \leftrightarrow Xy)$). This relation satisfies the logical role that identity is customarily thought to satisfy, given Universal Instantiation. For if a and b stand in this relation, then $Ba \leftrightarrow Bb$, instantiating X with B.[29] So this relation satisfies $L()$, whereas identity proper does not, according to this view, making conspicuous the charge that one is simply talking about the wrong relation. If there is a relation that satisfies the logical role of identity, then it is hard to see any other relation having a better claim to being the notion of identity appropriate to logic and metaphysics. Of course, some metaphysicians have postulated notions of 'loose identity' by which we sometimes count and individuate objects in natural language, whilst distinguishing it from the proper identity of logic and metaphysics (see, e.g., Chisholm, 1969 and Lewis, 1976). But to object to the logicians use of Leibniz's law on these grounds seems like an overreach.[30]

With $L(xyX)$ so justified we may may infer $L(xy)$ by universally instantiating X with A in the logical context $\lambda X\forall x\forall y(x = y \rightarrow Xx \rightarrow Xy)$.[31] The critical observation here is that, while Universal Instantiation is not in general valid, one can make such instantiations in logical contexts. For while words like 'belief', 'hope', 'says', and so forth create opaque contexts, words like 'not', 'and', and 'all' appear not to. A similar justification may be made of another quantified version of Leibniz's law, namely:[32]

$$\mathbf{L(X)}\ \forall X(a = b \rightarrow Xa \rightarrow Xb)$$

since one can also infer it from $L(xyX)$ by instantiation into a logical context.

Caie et al. (forthcoming) follow Bacon and Russell (2019) in rejecting the orthodoxy regarding Leibniz's law, but develop a classical account of opacity that keeps Universal Instantiation. Consequently $L(xyX)$, $L(xy)$, and $L(X)$ are all rejected along with $L()$.[33] But in addition to the worries above, these sorts of views are ontologically wild in ways that the nonclassical versions are not. For instance, in order to accommodate 1–3, there exists a property that Hesperus has which Phosphorus doesn't (namely, being believed by Simon to be bright). Call the relation of sharing the same properties, strict identity, and its negation strict distinctness: thus Hesperus and Phosphorus are strictly distinct. Just as we talk about there being multiple planets satisfying some criteria when there exist a distinct x and y that are planets with the criteria, we say there are strictly multiple planets when there are strictly distinct planets, x and y, satisfying the criteria. We will follow similar conventions for words like 'lots', 'several', 'more', 'most', and so on. Since Hesperus and Phosphorus are strictly distinct, there are at least two planets, in the strict sense, colocated with Venus. Of course, there's nothing special about the name 'Venus' either, so we should expect there to be strictly more than the strict two: there should be a 'strict lot' of them, one for each possible identity confusion.[34] So there is a sense in which the view inflates the ontology, even if all these strictly distinct planets are all identical in the loose sense. But there is also an internal worry. Clearly Hesperus is a planet between Mercury and Earth, similarly for Phosphorus. Indeed, given their identity, they must share many properties—all the *transparent* properties, including the property of being colocated with Venus, being visible in the evening and being visible in the morning, orbiting the Sun, and presumably any properties definable in the language of physics. But why is it, then, that Simon has the belief that Hesperus is bright, but not Phosphorus? After all, there are two strictly distinct planets colocated with one another, sharing all the same physical properties: it's not like one can be invisible in the evening while the other isn't—they occupy the same region of spacetime, have the same reflective properties, and so on. What is it about Hesperus, and not the physically indistinguishable Phosphorus, that allows Simon to attach his belief to it, and not a strictly different planet?[35]

The view raises other metaphysical concerns. For instance, many philosophers are attracted to physicalism, according to which every property can be defined from physical properties. The strongest version of this thesis is that every property is strictly identical to a logical combination of physical properties. But all physical and logical properties are plausibly transparent, as, plausibly, are things you can create by combining logical and physical operations. In which case belief must be strictly identical to a transparent property, and thus transparent itself (since strict identity obeys Leibniz's law).

The non-classical opacity theorist, by contrast, avoids the excesses of its classical cousin. It draws no distinction beween identity and strict identity: indeed, according to it one cannot define binary predicates that satisfy Leibniz's law. They can accept that belief is strictly identical to a transparent property, without inferring that belief is transparent, since *being transparent* is itself an opaque property.[36] And they may deny that there are two strictly distinct colocated planets.

10.5 Prior's Paradox and Opacity

We have argued that opaque contexts create a distinctive cluster of counterexamples to Existential Generalization (and Universal Instantiation). We have focused on the case of first-order quantification, for the sake of familiarity, but the considerations generalize to other semantic types. For instance, it seems plausible that the property of being a lawyer, and the property of being an attorney are the very same, but Simon might believe that Susan is a lawyer, without believing that she is an attorney. Similar considerations compel us to reject Existential Generalization for quantification into predicate position.

Of special interest is the case of quantification into sentence position. A solution to Prior's paradox along these lines, would accept the following (with similar things being said about the variants of Prior's paradox stated in terms of hope, fear, and so forth):

1. Simon said, at *t*, that nothing he said at *t* is true.
2. For no *P* did Simon say that *P* at *t*.
3. For some *P*, the proposition that *P* just is the proposition that nothing Simon said at *t* is true.

It is 3, in particular, that distinguishes this solution from the solutions based on non-existence, and bans on impredicativity.

But despite this difference, many of the goodmaking features of the non-existence approach apply here too. For instance, unlike the classical no-proposition view, there is a point to uttering sentences that do not result in you saying anything. For one can maintain that Simon didn't say anything true at *t*, even if one hasn't thereby said anything,.

According to the non-existence view, the reason Simon hasn't said anything with his utterance is that the proposition that nothing Simon said at *t* is true simply doesn't exist. Thus no one can say anything by uttering the sentence 'nothing Simon said at *t* is true' (even though one will nonetheless say that nothing Simon said at *t* is true). This non-existence makes general theorizing difficult. For instance, the symmetry of disjunction is sometimes captured with a generalization:

$$\forall XY(X \vee Y = Y \vee X)$$

But according to the view under consideration, this is insufficiently general. It will not entail the instance in which *X* is instantiated with 'nothing Simon said at *t* is true' and *Y* with 'snow is white'.

By contrast, even though we have established that there is no *P* such that Simon said at *t* that *P*, we have not established that someone else couldn't say that nothing Simon said *t* is true and thereby have said something. Nor have we established that Simon couldn't say something with the same utterance at another time. By analogy with the first-order case, while there's no-one who I know to have committed the Whitechapel murders, there might be someone else for which there is someone they know to have committed the Whitechapel murders (Jack the Ripper himself, for instance). Whereas, there can't be anything that that is believed (by *anybody*) to be a winged horse-god on the basis of believing that Pegasus is a winged horse-god.

Similarly, our generalization above is sufficient to prove all its instances, given the sentential analogue of Logical Instantiation. Since we can instantiate *X* with *P* into a purely logical context to get

$\forall Y\, P \vee Y = Y \vee P$. And we may then instantiate again to get $P \vee Q = Q \vee P$, as Y was not in the scope of a non-logical operation.

Of course, not all theorizing is insulated from opacity. For instance, Bacon et al. (2016) object that the universal generalization $\forall X(KX \to X)$, expressing the factivity of knowledge, is not general enough. An appeal to Logical Instantiation is of no avail here, since X is in the scope of an intentional operator, K. One must instead be content with a schema: $KA \to A$. But there is an extent to which schemas are unavoidable. For instance, in classical propositionally quantified logic, to ensure the that the quantified claim $\forall X(KX \to X)$ really does imply all of its instances, one needs a different schema, namely Universal Instantiation, $\forall XA \to A[B/X]$. It cannot be substituted with its universally quantified variant, $\forall Y(\forall XA \to A[Y/X])$, what we earlier called Quantified Instantiation, because it doesn't imply its instances. See note 26.

Another issue raised in Bacon et al. (2016) is that in some of the models of Prior's paradox considered there (motivated by the free logic approach to empty names, and not opacity), the domain of existing propositions weren't closed under certain logical operations. In some models, negations, conjunctions, and disjunctions of existing propositions didn't exist. In the most plausible models, propositions were closed under the finitary Boolean operations, but you would either have a proposition P that existed, without the proposition that Simon said that P existing, or you would have a context such that the proposition defined by $A[P]$ existed for each P, but $\forall X\ A$ didn't. These limitations were no accident, as one could show in that context that the following closure principles would lead to contradiction:

CL$\top$ $\exists p(p = \top)$
CL$\bot$ $\exists p(p = \bot)$
CL$\neg$ $\forall p \exists q(q = \neg p)$
CL$\Box$ $\forall p \exists q(q = \Box p)$
CL*Q* $\forall p \exists q(q = Qp)$
CL$\wedge$ $\forall p \forall q \exists r(r = (p \wedge q))$
CL = $\forall p \forall q \exists r(r = (p = q))$
CL$\forall$ $\forall p \forall q \exists r(q = \varphi \to (r = \forall p \varphi))$

That derivation made essential use, however, of Unrestricted Export: $(\exists X(X = B) \wedge A[B/X]) \to \exists X\ A$. The informal idea was to show, on the basis of these closure conditions, that Prior's proposition, $\forall X(QX \to \neg X)$, existed. Unrestricted Export tells us we may universally instantiate and existentially generalize on propositions we know to exist—in effect, classical reasoning is legitimate provided the relevant propositions exist.[37]

But the logic of opacity, in propositionally quantified logic, should include analogues of all of the principles we discussed in Section 10.4, including the schema Existence:

Existence $\exists X(X = A)$

Thus each of the closure principles listed above is thus trivially secured. And as for the derivation in Bacon et al. (2016), Unrestricted Export must be relinquished. But inspection of the argument reveals that one must apply Unrestricted Export to variables under the scope of Q, a move we have rejected on independent grounds, from considerations of opacity.

So an opacity-inspired response to Prior's paradox is desirable. But is it consistent? That is to say, can we have a model of Possible Saying within the logic of opacity. For clarity, I will restate all the principles we wish to show to be jointly consistent, converting the first-order logic of opacity of Section 10.4 to the context of propositionally quantified logic (the principles from Quantified Instantiation onwards)[38]

Possible Saying! $\Diamond Q!A$
Subsumption $\Box(Q!A \to QA)$
Uniqueness $\Box(Q!A \wedge Q!B \to (C \leftrightarrow C[A/B])$
Quantified Instantiation $\forall X\ (\forall Y\ A \to A[X/Y])$
Normality $\forall X\ (A \to B)\to \forall X\ A \to \forall X\ B$
Quantifier Exchange $\forall X \forall Y\ A \to \forall Y \forall X\ A$
Vacuous Quantification $A \to \forall X\ A$ when X does not occur free in A
Logical Instantiation $\forall Y\ A \to A[B/Y]$ provided Y is not in the scope of a non-logical operation (i.e. Q or $Q!$)
Reflexivity $\forall X\ (X = X)$
L(xy) $\forall XY\ (X = Y \to A \to A[Y/X])$
Existence $\exists X\ (X = A)$

For the sake of simciterplicity, we shall assume that propositions are individuated by necessary equivalence. Thus we may alternatively take $\Box$ as primitive, and defined $A = B$ by $\Box(A \leftrightarrow B)$, or take $=$ as primitive, and define $\Box A$ as $A = \top$. I will take the latter option, as it strikes me that identity is the more basic notion. Our language will thus be that of propositionally quantified logic, with a binary connective, =, and two unary connectives Q and $Q!$.

On this last point, one thing that should be emphasized is that our approach to opacity, despite capitulating much to the thought that attitudes are capable of drawing fine distinctions, does not take sides on propositional granularity. Indeed, everything we have said so far is consistent with *Booleanism*, the thesis that propositions are governed by the Boolean identity, as encoded by generalizations, statable in our propositionally quantified language with identity, like the commutativity of disjunction discussed earlier.

Theorem 10.5.1. The listed principles are consistent, and indeed have a possible worlds model securing Booleanism.

The full proof of this may be found in the appendix. But here is an informal explanation of the construction. Firstly, following Bacon et al. (2016), a model will consist of:

- A set of worlds, W.
- An accessibility relation, R, on W, for interpreting = (and consequently $\Box$). It will be an equivalence relation.
- A mapping, $|Q|: W \to P(P(W))$, that provides the extension of Q at each world w. $|Q|(w)$ is the set of propositions Qed at w.
- A mapping $D : W \to P(P(W))$ where $D(w)$ represents the domain of the propositional quantifiers at w.

A sentence of the form $A = B$ is true at a world w iff A and B are true at the same worlds accessible to w. Provided R is not the universal relation, an identity like this can be true even when A and B are true at different worlds. QA is true at a world w if the set of worlds where A is true is in the extension of Q at w (i.e. in $|Q|(w)$). $Q!A$ is true at w if, moreover, that set is the unique member of $|Q|(w)$. The remaining clauses are the obvious ones, and may be found in the appendix.

In our specific model, W will consist of two copies of the natural numbers, where worlds in the first copy see only worlds in the first copy (along the accessibility relation R), and similarly, worlds in the second copy see only worlds in the second copy.

The domain of quantification in a world in the first copy will consist of finite and cofinite sets of worlds in the first copy. Likewise the domain of a world in the second copy will consist of finite and cofinite sets of worlds from the second copy. (Cofiniteness is being computed relative to the copy in question, not relative the whole set of worlds.) To interpret Q we fix a bijection σ between the naturals, $\mathbb{N}$, and the finite and cofinite subsets of $\mathbb{N}$. The interpretation of Q at the world n, in either copy, is the same. It contains exactly one set of worlds—the union of the two copies of $\sigma(n)$. Note that this proposition is not in the domain of any world, either in the first or the second copy. But it is nonetheless necessarily equivalent to a proposition that is in the domain, in either case, for it coincides with a finite or cofinite set from the first (or second) copy respectively on the accessible worlds at that copy. Given the coincidence of identity with necessary equivalence in this model, this secures Existence.

Say that a set of worlds is symmetric if it is the same in both copies. Given the symmetry of the way the semantics is set up, a world in the first copy is indistinguishable from its sibling in the second copy, and vice versa. Thus one ought to expect a closed sentence could only be true at a world in the first copy if it was also true in its sibling in the second copy, and vice versa. Thus every closed sentence expresses a symmetric proposition. Moreover, if things work out nicely, every closed sentence should express a symmetric proposition that is finite or cofinite relative to either (and thus both) copies. In which case, for any closed sentence A, the proposition it expresses in the model is the unique proposition in the extension of Q at the two sibling worlds the bijection σ associates to that proposition. Thus, $\Diamond Q!A$ will be true at any world in the first or second copy. The tricky part of this argument is establishing that every closed sentence expresses a finite or cofinite set, relative to both copies; the full details are in the appendix.

10.6 Truth and Meaning

The liar paradox poses familiar problems for many attempts at systematic semantic theorizing. Could these considerations on intentionality and opacity shed any light on it? In this section I will attempt to leverage our remarks on opacity and Prior's paradox to motivate a novel theory of truth and a diagnosis of the liar paradox.

While much work on the liar has focused on the notion of truth, there is a large amount of work in semantics that is best thought of as characterizing the notion of a sentence meaning a proposition. According to this paradigm, lexical items more generally are assigned meanings, and the meanings of complex phrases are a computed in terms of the meanings of their constituents, in accordance with the principle of compositionality. The meaning of a sentence is a proposition. The notion of sentential truth is then a derivative notion, to be understood as meaning a true proposition. We might formalize this by introducing a device, M, formalizing 'means that', that combines with the name of a sentence and a sentence to form a sentence, just as verbs like 'says that', 'hopes that', or 'fears that' combine with the name of a person and a sentence to form a sentence.

We can then introduce truth, in propositionally quantified first-order logic, via a definition:

Truth $Tx := \exists Z\,(M\,x\,Z \wedge Z)$
Falsehood $Fx := \exists Z\,(M\,x\,Z \wedge \neg Z)$

Suppose, moreover, that for each sentence of the language in question, A, there is a name $\langle A \rangle$, that on its intended interpretation denotes the sentence A. A *prima facie* compelling disquotational principle says:

The Meaning Schema The sentence 'P' means that P

which we may formalize:

$M\,\langle A \rangle\,A$

But now suppose that there is a name, c, identical to the representative name for the sentence, $\forall X\ (M\ c\ X \rightarrow \neg X)$, i.e. $c = \langle \forall X\ (M\ c\ X \rightarrow \neg X) \rangle$. An instance of the Meaning Schema is:

$$M\ \langle \forall X\ (M\ c\ X \rightarrow \neg X) \rangle\ \forall X\ (M\ c\ X \rightarrow \neg X)$$

Replacing $\langle \forall X\ (M\ c\ X \rightarrow \neg X) \rangle$ for c, we get:

$$M\ c\ \forall X\ (M\ c\ X \rightarrow \neg X)$$

Now interpret $Q\ X$ in Prior's theorem $M\ c\ X$. Applying Modus Ponens to Prior's theorem we get:

$$\exists X\ (M\ c\ X \wedge X) \wedge \exists X\ (M\ c\ X \wedge \neg X)$$

In other words c means a truth and a falsehood. According to our definitions this means c is both true and false. Note that this is not strictly speaking inconsistent. But it contradicts the assumption, seemingly just as important for doing semantics compositionally, that a sentence means at most one thing. For if A meant several propositions, and B meant several propositions, how would we compute the meaning of $A \vee B$? The obvious choice of letting it mean any proposition you can get by disjoining something A means with something B means doesn't work, for otherwise $A \vee \neg A$ can mean falsehoods. We need some way of coordinating the meanings of the two disjuncts, and we can't do this with M as our only primitive.[39]

What about the more traditional liar sentence? Suppose we had a name, c such that $c = \langle \neg Tc \rangle$. Observe that expanding out the definition of $\neg Tc$ you get $\neg\exists X\ (M\ c\ X \wedge X)$, and applying some logic this is equivalent to $\forall X\ (M\ c\ X \rightarrow \neg X)$. Thus there is a straightforward sense in which our paradox considered above is just another version of the liar, modulo our definition of truth.

How might opacity shed light on this paradox? I have suggested in the previous section that 'says that' is opaque. What things sentences mean supervenes in large part, on what propositions people have said. One could take this to suggest that 'means' is opaque. According to one theory, to say that P is just to utter a sentence that means that P. If 'means' was not opaque, then neither would 'says' be. But this does not seem like a plausible account of saying to me. Merely uttering a sentence is not usually sufficient for saying anything. And it seems plausible that one can say things using *ad hoc* conventions, as when one is in a foreign country and doesn't speak the language. Indeed, the fact that one can say things in the absence of a language suggests a very appealing theory of how sentences can acquire meanings in the first place: that one uses certain sentences when saying things to establish the the more general convention that the sentence be used to say that thing.[40] On this view, *saying* is a primitive propositional attitude, and meaning is to be explained in terms of it. In this case too, we should expect meaning to be opaque, for it is explained in terms of an opaque attitude. This appears to be no accident: other reductive theories of meaning reduce meaning to other intentional attitudes. For example, Lewis, in his book *Convention* (1969), attempts to reduce meaning to the desires and beliefs of the language speakers. Davidson's account of radical interpretation involves the attitude of belief in a similarly central way (Davidson, 1973).

Even if we do not make this reductive move, there are independent reasons for thinking that meaning is opaque. In Bacon and Russell (2019) we are, among other things, concerned with upholding the Millian theory of names. According to this theory, the meanings of 'Hesperus' and 'Phosphorus' are the same: the planet Venus. We might formalize this $[\![\langle a\rangle]\!] = [\![\langle b\rangle]\!]$. But a principle of compositionality then appears to entail that the meaning of 'Hesperus is believed to be bright' is the same as the meaning of 'Phosphorus is believed to be bright', i.e. that $[\![\langle Ba\rangle\ ]\!] = [\![\langle Bb\rangle]\!]$. For the principle of compositionality says: $[\![\langle Ba\rangle]\!] = [\![\langle B\rangle]\!]\ [\![\langle a\rangle]\!]$, and $[\![\langle Bb\rangle]\!] = [\![\langle B\rangle]\!]\ [\![\langle b\rangle]\!]$. Leibniz's law allows us to derive the undesired conclusion. Undesired, because these sentences appeared to have different truth values in ancient times, and so, presumably different semantic values. We draw the following moral: opacity exists in the metalanguage—meaning is opaque.

Many consistent theories of truth have been proposed that avoid the liar paradox: the logical landscape is pretty well-explored at this point.[41] However, much work is still needed in the interpretation of these theories. In the following I want to motivate a theory of truth that falls under the umbrella of the so-called *classical gap* theories of truth, of which prominent proponents include Feferman (1991) and Maudlin (2004). These theories say of the liar sentence, $c = \langle\neg Tc\rangle$, that it is neither true nor false. But many have argued that such theories are self-undermining. For in virtue of the liar being neither true nor false, it follows that the liar isn't true. But this is exactly the liar sentence. Such theorists have to make seemingly absurd speeches like:

1. c is not true.
2. The sentence I just uttered, 1, is not true.

Scharp (2013), for instance, calls this the 'self-refutation problem'. Priest (2005) puts it forcefully as follows: 'truth is the aim of assertion. Once this connection is broken, the notion of assertion comes free from its mooring, and it is not clear why we should assert anything' (p. 485).

Thus, we must do something to explain what we are doing when we assert, if the aim is no longer to assert truths. I think an opacity friendly theory of meaning can meet this demand.

Let us rehearse what we must say about the liar paradox, drawing on our solution to Prior's paradox in the logic of opacity.

- *c* means that nothing *c* means is true.
- For no *X* is it the case that *c* means that *X*.

Given our definition of truth, as meaning a truth, it follows that *c* is not true: there is no proposition it means, and so no truth it means. Similarly, it is not false either, as there is no false proposition it means. Thus we have a version of the classical gap theory.[42] In particular, we must make the seemingly absurd sequence of assertions listed in 1 and 2.

We began by suggesting that it is meaning, not truth, that is the more theoretically central notion. This opens up the option, not available to someone just theorizing in terms of truth, of giving a meaning-theoretic explanation of why we might assertively utter *c*—the sentence '*c* is not true'—even though it is not itself true. The explanation is this: because by uttering that sentence, *one thereby says that* c *is not true*. And by saying that *c* is not true, one may pass on the knowledge that *c* is not true to ones audience, and they can act accordingly. Of course, these sayings and knowledge are opacity laden: there is no proposition one has thereby said, or that ones audience thereby knows. But, as we argued in Section 10.3, in the context of the non-existence approach, this is no barrier to the knowledge playing the relevant role in informing your actions.

10.7 Logic First or Semantics First?

Let me end by discussing one aspect of the our general approach to opacity that bears special emphasis. The explanation of opacity — failures of Leibniz's law—is often thought to be a burden of the philosophy of language. It is usually given a semantic explanation. Perhaps 'Hesperus' and 'Phosphorus' do not refer to a single planet, but to a pair of distinct individual concepts, or perhaps the identity symbol or attitude verbs semantically behave differently than we might otherwise have thought. By contrast, on the present view the opacity of attitude verbs is not explained semantically—any putative explanation along these lines would be dizzying at best given the opacity of semantic words themselves.

As one might expect from a disquotational theory, the semantic facts here are uncomplicated. 'Hesperus' straightforwardly means Hesperus, 'Phosphorus' straightforwardly means Phosphorus, 'is identical to' straightforwardly means is identical to, and so on. A false instance of Leibniz's law, such as 'if Hesperus is Phosphorus and Hesperus was believed to be bright then Phosphorus was too' simply means that if Hesperus is Phosphorus and Hesperus was believed to be bright then

Phosphorus was too. The semantic facts provide no clue as to the reason for the falsity of this last sentence. It is false, rather, because of the following fact about planets and our past astronomical beliefs: that Hesperus *is* Phospherus, Hesperus *was* believed to be bright, and Phosphorus *wasn't*. Further explanation of this fact might be achieved by examining the astronomy of the time, the psychology of the relevant astronomers, and so on; someone looking for a more profound explanation will be disappointed.

Our preceding example also highlights the fact that failures of Leibniz's law in the object language give rise to failures of Leibniz's law in the metalanguage. It is hardly surprising, given a disquotational theory of meaning, that the logic of the metalanguage and object language coincide. But it's also worth emphasizing that a failure of the object and metalanguage harmonize in this way is puzzling. The semantics-first approach to opacity, for instance, purports to have its cake and eat it too: although one can accommodate failures of Leibniz's law in the object language using their semantic apparatus, they claim they aren't *really* denying Leibniz's law. By which they mean, presumably, that they uphold Leibniz's law in the metalanguage in which their semantics is formulated. Apart from precluding disquotational accounts of identity and names[43], this line of reasoning rests on a spurious distinction between the logic of the object language and metalanguage. It would be surprising if one could formulate an adequate semantic account of words like 'belief' without having, in the metalanguage, the means to express the attitude of *believing*. (Indeed, most papers on the semantics of belief are written in English, which evidently does have this word.) And if the metalanguage has opaque words—perhaps even invoking them when formulating the semantics of words like 'belief'—then we shouldn't expect Leibniz's law to hold when reasoning about the semantics of propositional attitudes than when reasoning directly about propositional attitudes.

10.8 Conclusion

We have argued that by taking opacity seriously one may provide a philosophically appealing solution to Prior's paradox. The resulting account of intentionality naturally extends to an opacity-laden theory of meaning that vindicates a classical gap theory of truth that is able to meet the challenges usually directed at that account.

10.A Appendix

We shall work in the language of propositionally quantified logic. The wffs consist of an infinite stock of propositional variables, X_1, X_2, … , a wff $\neg A$, QA, $Q!A$, $A \wedge B$, $A = B$, and $\forall XA$, whenever A and B are wffs, and X a propositional variable.

Following Bacon et al. (2016), a model of this language consists of:

- A set W
- An equivalence relation R on W
- A map $|Q| : W \to P(P(W))$
- A map $D : W \to P(P(W))$ such that $D(w) \subseteq P(P(R(w)))$

Here we follow the usual convention of writing $R(w)$ for $\{x \in W \mid Rwx\}$. A variable assignment is a function $g : Var \to P(W)$, assigning each propositional variable to a set of worlds. Write $g \sim_X h$ iff g and h agree on every variable except, possibly, X. We may interpret an arbitrary wff relative to an assignment as follows:

$\llbracket X \rrbracket^g = g(X)$
$\llbracket A \wedge B \rrbracket^g = \llbracket A \rrbracket^g \cap \llbracket B \rrbracket^g$
$\llbracket \neg A \rrbracket^g = W \setminus \llbracket A \rrbracket^g$
$\llbracket \forall XA \rrbracket^g = \{w \mid w \in \llbracket A \rrbracket^h$ for every h such that $h(X) \in D\,(w)$ and $g \sim_X h\}$
$\llbracket QA \rrbracket^g = \{w \mid \llbracket A \rrbracket^g \in |Q|(w)\}$
$\llbracket Q!A \rrbracket^g = \{w \mid \llbracket A \rrbracket^g$ is the only member of $|Q|\,(w)\}$
$\llbracket A = B \rrbracket^g = \{w \mid \llbracket A \rrbracket^g \cap R(w) = \llbracket B \rrbracket^g \cap R(w)\}$

This model theory is sound for the logic of opacity, except for Existence and Logical Instantiation: for any $w \in W$ of such a model, and any closed theorem of the logic of opacity without Existence and Logical Instantiation, A, $w \in \llbracket A \rrbracket$.[44] In order to validate both Existence and Logical Instantiation a model must satisfy the further constraint:

> For any wff, A, $w \in W$ and assignment g whose range contained in $D\,(w)$, $\llbracket A \rrbracket^g \cap R(w) \in D(w)$.

I'll now spell out the model described in Section 10.5 in a little more precision. Let σ be a bijection between $\mathbb{N}$ and the finite/cofinite subsets of $\mathbb{N}$. We define a model as follows:

- $W = \mathbb{N} \times 2$
- $R = (\mathbb{N} \times \{0\})^2 \cup (\mathbb{N} \times \{1\})^2$
- $|Q|_{(n,i)} = \{\sigma(n) \times 2\}$
- $D((n, i)) = \mathcal{P}_{f/cf}(\mathbb{N} \times \{i\})$

Here are some useful definitions. In what follows X denotes a subset of W

Definition 10.A.1. A *left set* is a finite or cofinite subset of $\mathbb{N} \times \{0\}$, and a *right set* is a finite or cofinite subset of $\mathbb{N} \times \{1\}$. Denote the set of left sets L, and the set of sets R.

Note: a left (right) set must be finite or cofinite *relative to* $\mathbb{N} \times \{0\}$ ($\mathbb{N} \times \{1\}$). A cofinite subset of $\mathbb{N} \times \{0\}$ will not be cofinite relative to W.

Definition 10.A.2. The set $X \subseteq W$ is *left finite or cofinite* iff $X \bigcap \mathbb{N} \times \{0\}$ is a left set (it is *right finite or cofinite* iff $X \bigcap \mathbb{N} \times \{1\}$ is a right set).

Definition 10.A.3. The set X is *symmetric* iff $(n, 0) \in X \Leftrightarrow (n, 1) \in X$ for every $n \in \mathbb{N}$.

Recall that our conjecture was that all closed sentences express sets that are both (i) symmetric and (ii) finite or cofinite (in W). We will call the set of subsets of W satisfying both (i) and (ii), $\mathcal{S}$.

Definition 10.A.4. A *left assignment* is a function g from propositional variables to left sets (a *right assignment* maps variables to right sets). Call the set of all left and right assignments A_L and A_R respectively.

Suppose that As free variables lie in some finite set V of size n. Then $[\![A]\!] \; : L^n \rightarrow \mathcal{P}(W)$ defines an n-ary function from left sets to subsets of W (and similarly defines a function from right sets).

Definition 10.A.5. An n-ary function $f : L^n \rightarrow \mathcal{P}(W)$ is *well behaved* if and only if

a. f only takes right finite or cofinite sets as values.
b. $\{f(\bar{x}) \bigcap \mathbb{N} \times \{1\} \mid \bar{x} \in L^n\}$ is finite. In other words, as we run through the set of possible values for $\bar{x}$, the right half of $f(\bar{x})$ takes at most finitely many different values.

A function from $f : R^n \rightarrow \mathcal{P}(W)$ is well behaved if the natural parallel conditions obtain.

Fix V to be a finite set of variables, and A a formula whose variables all lie in V. I will write $[\![A]\!]$, without the assignment superscript, to denote the n-ary function on left sets, that takes $P_1, \dots P_n \in L$ to $[\![A]\!]^{g[X_1 \mapsto P_1 \dots X_n \mapsto P_n]}$ (for any assignment g).

Theorem 10.A.6. Let V be a finite set of variables, and A a formula whose variables all lie in V. Then $[\![A]\!]$ defines an n-ary function on left sets that is well-behaved.

Proof. For $X_i \in V$, $[\![X_i]\!]$ is constant if we vary X_j for $j \neq i$. If we vary the value of X_i clearly condition (a) holds, and the only value $[\![H_i]\!] \cap \mathbb{N} \times \{1\}$ takes on is $\emptyset$.

If $[\![A]\!]$ and $[\![B]\!]$ are well-behaved, so are $[\![\neg A]\!]$ and $[\![A \wedge B]\!]$ (straightforward).

$\Box$ maps every intension to either $\emptyset$, $\mathbb{N} \times \{0\}, \mathbb{N} \times \{1\}$ or W so $[\![\Box A]\!]$ is clearly well-behaved.

Q and $Q!$ behave the same way: they map every member of $\mathcal{S}$ to a symmetric pair of the form $\{(n,0),(n,1)\} \in \mathcal{S}$, and maps every other set to $\emptyset$. Thus:

1. $[\![QA]\!]$ takes only the emptyset and symmetric pairs as values, so it satisfies condition (a) for well-behavedness.
2. If $[\![A]\!] \cap \mathbb{N} \times \{1\}$ takes only k many different values, then $[\![A]\!]$ can take at most k different values in $\mathcal{S}$. So $[\![QA]\!]$ can take on the emptyset and at most k different symmetric pairs as values, thus it satisfies condition (b).

Suppose $[\![A]\!]$ is well-behaved. Note by assumption all As variables lie in V, so if $X_i \notin V$ $[\![\forall X_i A]\!] = [\![A]\!]$ which is well-behaved.

If $X_i \in V$ then $[\![\forall X_i A]\!] \cap \mathbb{N} \times \{1\}$ maps a tuple to a finite intersection of the different values of $[\![A]\!] \cap \mathbb{N} \times \{1\}$ (we know $[\![A]\!] \cap \mathbb{N} \times \{1\}$ takes on finitely many different values because it's well-behaved). So if the values of $[\![A]\!] \cap \mathbb{N} \times \{1\}$ are all right finite/cofinite so are $[\![\forall X_i A]\!] \cap \mathbb{N} \times \{1\}$.

Note that if $f(x_0, \dots, x_n)$ takes on finitely many different set values as we vary x_i the function $h(x_1, \dots, x_n) = \bigcap_{x_0} f(x_0, \dots, x_n)$ can only take on finitely many different values. Thus $[\![\forall X_i A]\!]$ satisfies condition (b) if $[\![A]\!]$ does.

Theorem 10.A.7. The logic of opacity, Possible Saying!, Subsumption, and Uniqueness are all true in this model at any world in the left half of W. (And by a symmetrical argument, also true at any world in the right half of W.)

Here is a proof sketch.

Proof. Evidently every closed formula expresses a symmetric set (by the symmetry of the semantics). Closed sentences express constant functions from $L^n \rightarrow \mathcal{P}(W)$. By Theorem 10.A.6, every closed sentence expresses (constantly) a right-finite/cofinite set.

Putting 1 and 2 together, closed sentences express sets in $\mathcal{S}$, and every member of $\mathcal{S}$ is in the extension of Q and $Q!$ at some world. This secures the validity of Possible Saying!

Also, since every closed formula expresses a left finite/cofinite set, every proposition expresses a proposition necessarily equivalent to a left set securing Existence and Logical Instantiation.□

Notes

1. The sentential T-schema is problematic because there is a direct Gödel-Tarski style argument against it resting on the fact that substitution (and hence diagonalization) are clearly in good-standing for sentences. The notion of substituting one thing for another in a proposition is not so clearly in good-standing—it is ill-defined if you thought propositions were sets of worlds, for instance. One might thus consider the inconsistency of the propositional T-schema in conjunction with the notion of substitution and the accompanying laws of substitution to be an argument against propositional structure. Those *antecedently* committed to a structured theory of propositions that permits diagonalization cannot accept the propositional T-schema. But such views face hard questions of their own. For instance, while one might hope to explain the failure of the sentential T-schema in terms of some failure of sentences and propositions to interface properly (failure to express a proposition being one such explanation), no similar explanation could be given for the failure of the propositional T-schema.
2. See Bacon (2019, Section 2.1).
3. I do not make the assumption, sometimes made, there are only two elements in the domain of type t, the truth values, or even that the domain of type t consists of sets of worlds. A suitably general class of models free of these assumptions is described in Benzmüller et al. (2004). The matter is slightly more subtle in non-full Henkin models of type theory, as it turns out that having a left-inverse and being injective are not provably equivalent in standard axiomatizations of higher-order logic: see the appendix of Bacon (2018).
4. In type theory one may define a binary connective, =, that takes two sentences, ***A*** and ***B***, and produces another sentence informally corresponding to the identity of ***A*** and ***B***. The principle of β-equivalence, that $(\lambda p.M)N = M[N/p]$ provided N is substitutable for P, ensures the following identities:

    1. *true that* $\boldsymbol{P} = (\lambda X.$*true that* $\boldsymbol{X})P$ by β.
    2. $(\lambda X.$*true that* $X)P = (\lambda X.X)P$ by the minimal theory.
    3. $(\lambda X.X)P = P$ by β.

5. In order for the proof to go through on these assumptions, the choice of primitives matters. Since neither assumption involves existential quantification, for example, it cannot be taken as primitive, or we would not be able to prove sentences involving it (such as Prior's theorem).
6. Priest, for example, recommends weaking classical propositional logic in response to Prior's paradox in Priest (1991).

7. Ramification is spelled out as a response to Prior's paradox in Tucker and Thomason (2011), and Kaplan (1995) and Kripke (2011) express sympathy to this line of response in the context of some related paradoxes. See Tucker (2018) for a non-standard approach to the ban on impredicativity, that does not involve explicit ramification.
8. Cases related to this are discussed in Bacon (2019), and explained in the context of a different interpretation of Prior's result.
9. Thus, according to our understanding of utterance, a cough that sounds like a sentence in Farsi does not count as an utterance of that sentence, but the noise Simon makes in our envisioned situation, does count as an utterance.
10. Of course, many people have argued that attitude verbs like 'say that' are context sensitive, and that quantifiers, including the propositional quantifiers appearing in Prior's language, are context sensitive. Some have even suggested that the context sensitivity of these things are key to understanding the paradoxes, and so should not be bracketed (for instance, Parsons, 1974; Glanzberg, 2001; Burge, 1979; Simmons, 1993). But we are exploring these views, we are exploring the idea that the paradoxes can be solved without postulating context sensitivity, so bracketing the context sensitivity of these expressions is indeed the proper methodology.
11. There are some substantial assumptions that must be taken on in order to even formulate this sort of ban. For instance, one cannot accept a theory of propositional granularity, such as the possible worlds theory, in which a quantified proposition, like $\forall x\ x = x$, and a non-quantified claim, like $F\,a \rightarrow F\,a$, are identified, for otherwise one would not be able to make the distinction between quantified and non-quantified propositions.
12. Although my remark here is representative of what a ramified theorist would like to say, it is strictly speaking nonsense. For in order for it to express a truth, I must have quantified over quantifiers unrestrictedly. But if I could do that, I could introduce an unrestricted quantifier: everything$_U$ is F if and only if, for every quantifier, everything$_i$, everything$_i$ is F. Whether this is the sort of Wittgensteinian nonsense one can simply kick away is a contentious matter, as is the status of heuristically useful nonsense more generally, and I will not attempt to adjudicate here. But see Williamson (2003).
13. This motivation is explored in Bacon et al. (2016), alongside the ramificationist response.
14. Free logicians take failures of Existential Generalization with varying degrees of seriousness. For instance, some explain the failures by saying that atomic sentences involving empty names are by default false, and that empty names do not in general make an important semantic contribution to a sentence in which they occur. In the cases where they appear to, such as in intentional verbs, special mechanisms are invoked (see, for instance, Sainsbury, 2005). I have defended a view that takes failures of Existential Generalization seriously, indeed, which takes them as primitive and not to be explained by any general property of the name 'Pegasus', but in terms of what properties Pegasus in fact has (see Bacon, 2013).
15. The slightly awkward metalinguistic formulation here is to avoid the Wittgensteinian nonsense alluded to in note 12.
16. Variants of this paradox are discussed in Hazen (1987), Zardini (2008), and Bacon (2019).
17. And while Alice cannot know that Pegasus is real, she can know that Pegasus *isn't* real, from which one still can't infer that there's something Alice knows isn't real (or else one could infer that there's something that isn't real).

18. Similarly, Simon can make it true that he said that Pegasus is real, even if there is nothing such that he made it true that he said that *it* is real.
19. This is the principle called K2 in Bacon et al. (2016).
20. This is called $K2^+$ in Bacon et al. (2016).
21. See, e.g., Crimmins and Perry (1989).
22. In doing so, I am thus setting aside approaches that treat restrictions of $L()$ involving direct predications as special. Such approaches typically deny the equivalence between $(\lambda x\ A)t$ and $A[t/x]$ (see Kripke, 2005 and Salmon, 2010).
23. Some philosophers place special significance on simple predications, and will therefore object to this sort of terminological shorthand. They will think that 'Hesperus is believed by Simon to be bright' to be different from 'Simon believes that Hesperus is bright'. By simply replacing the former formulations with the latter in the following discussion, we may satisfy those who object to these manipulations, but at the cost of readability.
24. See Bacon and Russell (2019).
25. Note, also, that our axiomatization includes Quantifier Exchange—a principle that is usually derivable with the help of Universal Instantiation. It is provable, even in free logic, from $L()$. But without either $L()$ or Universal Instantiation it is not, and must be put in by hand: see Fine (1983).
26. One might have thought that the quantified version is stronger, but one quickly sees that to derive Universal Instantiation from Quantified Instantiation, you have to already have Universal Instantiation in order to instantiate y with t.
27. This is where I depart from Bacon and Russell (2019). They restrict instantiation to purely logical contexts, but run into trouble making seemingly unproblematic instantiations in non-logical contexts. For instance, one cannot straightforwardly infer $P \vee Q = Q \vee P$ from $\forall XY\ (X \vee Y = Y \vee X)$. Instantiating X with P yields $\forall Y\ (P \vee Y = Y \vee P)$, but the result is no longer logical, and we thus cannot instantiate Y with Q. The notion of scope is a little tricky in higher-order logic with λ-terms, as we do not want to count X as in the scope of a non-logical operation in the context $\wedge PX$ (even though P is non-logical), but we do want to count it as in the scope of the operator B in $(\lambda YY)\ BX$, as it is β-equivalent to BX. One must instead define the scope for things in β-normal form, which is the smallest set of terms that contain (i) the variables and constants, (ii) $QM_1 \ldots M_n$ (provided this is well-typed) whenever it contains $M_1 \ldots M_n$ and Q is a variable or constant, and (iii) $\lambda x\ M$ whenever it contains M. The free variables in the scope of P in a β-normal form term M, $S(P, M)$, may then be defined inductively:

    1. $S(P, Q) = \emptyset$ when Q is a variable or constant.
    2. $S(P, PM_1 \ldots M_n) = \bigcup_i FV\ (M_i)$ when M_i are in β-normal form.
    3. $S(P, QM_1 \ldots M_n) = \bigcup_i S\ (P, M_i)$ when M_i are in β-normal form, Q a variable or constant, $P \neq. Q$.
    4. $S(P, \lambda xM) = S\ (P, M) \setminus \{x\}$ when M is in β-normal form.

    here $FV(M)$ denotes the free variables of M.
28. $L(xy)$ allows us to derive the quantified versions of these, but without Universal Instantiation we can't derive the instances.
29. This ensures that $L()$ holds for direct predications. One can show that $L()$ holds for an arbitrary context A, in higher-order logic, by using the device of λ abstraction, which can turn an arbitrary context into a predication. Specifically, one must appeal to the principle $(\lambda\ x\ A)t = A[t/x]$, which states that

substituting a (substitutable) term *t* into an arbitrary context, *A* (i.e. a formula with one free variable) is equivalent to a direct predication. Philosophers who have rejected this principle tend to insist that it was only the instances of *L*() involving direct predications that we should have cared about in the first place. My remarks to follow about the logical role of identity, then, should be acceptable to these philosophers as well.

30. Caie et al. (forthcoming), for instance, accept the existence of a strict identity relation that satisfies *L*(), but respond to our argument on the grounds that basic judgments about English seem to directly support the idea that identity does not satisfy *L*(). But if there is indeed a relation that satisfies the logical role that identity is supposed to, then it seems to me the primary matter of substance has been settled. What remains is a question concerning which role takes priority, and whose occupant deserves the name 'identity': the logical role of satisfying *L*(), and another role more closely connected to judgments about identity in English. But this is less substantial than it might seem. Consider, for instance, the dynamic semanticist who takes the fact that '*A* and *B*' and '*B* and *A*' are not always equivalent in English to mean that the principle of commutativity of conjunction is not valid. They might concede that there is a connective satisfying the classical laws of conjunction, but that this connective is not really *conjunction*. I think classical logicians can acknowledge this, but go about their usual business unabated, so long as they are explicit that it is the classical connective that is the target of their logical theorizing.
31. Again, the assumption made at the outset that arbitrary contexts are equivalent to direct predications.
32. There is a complication in deriving this second principle: after instantiating *x* with *a*, the context is no longer logical since it contains *a*. Thus second instantiation of *y* for *b* is not in a purely logical context. This can be circumvented as in Bacon and Russell (2019) by a slightly more complicated argument. However, in the present context, it is straightforwardly derivable given Logical Instantiation, since although the context is no longer logical, it is a context in which *y* doesn't appear in a non-logical context.
33. Clearly, any of $L(xyX)$, $L(xy)$, and $L(X)$ can be used to derive *L*() given Universal Instantiation.
34. Presumably one for every identity confusion formulable in some possible language. Alternatively, one might think that there are only as many strictly distinct planets as actual identity confusions, but this either means that the strict number of planets depends on the practices language speakers in surprising ways, or else means there are surprising coincidences between these practices and the strict number of planets.
35. Might one attempt to turn this into an argument against opacity *tout court*? How can Simon believe that Hesperus is bright, but not Phosphorus, while both Hesperus and Phosphorus are plainly visible to him and bright (irrespective of whether they are strictly identical or not)? The initial reason for believing in opacity is that one can fail to believe that Phosphorus is bright, even when it is plainly visible to you, because believing that Phosphorus is bright is more demanding. You must also think about Venus in the right way—presumably in a different way than is required to have the belief that Hesperus is bright. None of this is explained by the fact that there's a thing out there in the world, Phosphorus, which you have failed to believe to be bright. The thought that opacity is all in the head and not in the world is born out in the theory of Bacon and Russell: there simply isn't a plainly visible planet which Simon fails to believe to be bright on the basis of his observations. By contrast, Caie et al. maintain that there

is. According to them, there are two (identical, but strictly distinct) planets, which are both plainly visible to Simon, but of which he believes one to be bright, and not the other.

36. See the discussion in Bacon and Russell (2019) at the end of Section 4.
37. Actually Bacon et al. (2016) do not appeal explicitly to Unrestricted Export—rather they are implicitly assuming $L()$ which allows them to derive it from Quantified Instantiation. Omitting the assumption of $L()$ seems like an oversight.
38. Observe that, in the present setting, the schema $C \leftrightarrow C[B/A]$ is strictly stronger than the identity $A = B$. The schema includes the instance $A = B \leftrightarrow B = B$, which straightforwardly entails $A = B$. But $A = B$ does not entail $QA \leftrightarrow QB$, an instance of the schema. Thus Uniqueness in the present setting is stated as it is, and not as in Section 10.3.
39. Cian Dorr is developing a view with this form, but the coordination problem is solved by having more complicated primitives.
40. This account is obviously extremely crude as it stands, but see Schiffer (1972) for more elaborate theory along these lines.
41. For an early classification of the options in classical logic, see Friedman and Sheard (1987).
42. That we have ended up with a classical gap theory, and not a classical glut theory, follows from our choice to define truth as 'means a true proposition', as opposed to 'means only true propositions', and falsehood as 'means a false proposition', instead of 'means only falsehoods'. In symbols, the alternative definition of truth would be $T\,x := \forall X(M\,x\,X \rightarrow X)$, and falsehood $F\,x := \forall X(M\,x\,X \rightarrow \neg\,X)$. The alternative definitions would vindicate a classical glut theory—*c* would be both true and false—subject to the dual worry that to endorse it one must assert falsehoods, instead of untruths. The choice to pursue this as a gap or glut theory strikes me as a matter of taste, and not a matter of substance, as it is meaning, not truth and falsehood, that is the theoretically central notion.
43. See the discussion of opaque semantics in §2 of Bacon and Russell (2019).
44. One has to be a little more careful with open theorems, by restricting attention to assignments that assign variables to propositions in the domain of the relevant world.

References

Bacon, A. (2013). Quantificational logic and empty names. *Philosophers' Imprint*, 13.

Bacon, A. (2018). The broadest necessity. *Journal of Philosophical Logic*, 47(5): 733–783.

Bacon, A. (2019). Radical anti-disquotationalism. *Philosophical Perspectives*.

Bacon, A., Hawthorne, J., and Uzquiano, G. (2016). Higher-order free logic and the prior-Kaplan paradox. *Canadian Journal of Philosophy*, 46(4–5): 493–541.

Bacon, A. and Russell, J. S. (2019). The logic of opacity. *Philosophical and Phenomenological Research*, 99(1): 81–114.

Benzmüller, C., Brown, C. E., and Kohlhase, M. (2004). Higher-order semantics and extensionality. *Journal of Symbolic Logic*, 69(4): 1027–1088.

Boolos, G. (1984). To be is to be a value of a variable (or to be some values of some variables). *Journal of Philosophy*, 81(8): 430–449.

Burge, T. (1979). Semantical paradox. *Journal of Philosophy*, 76(4): 169–198.

Caie, M., Goodman, J., and Lederman, H. (forthcoming). Classical opacity. *Philosophical and Phenomenological Research*.

Chisholm, R. M. (1969). The loose and popular and the strict and philosophical senses of identity. In Care, N. S. and Grimm, R. H., editors, *Perception and Personal Identity*, pages 82–106. Press of Case Western Reserve University.

Crimmins, M. and Perry, J. (1989). The prince and the phone booth: Reporting puzzling beliefs. *Journal of Philosophy*, 86(12): 685.

Davidson, D. (1973). Radical interpretation. *Dialectica*, 27(1): 314–328.

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

Fine, K. (1983). The permutation principle in quantificational logic. *Journal of Philosophical Logic*, 12(1): 33–37.

Frege, G. (2010). On sense and reference. In Byrne, D. and Kölbel, M., editors, *Arguing About Language*, pages 36–56. Routledge.

Friedman, H. and Sheard, M. (1987). An axiomatic approach to self-referential truth. *Annals of Pure and Applied Logic*, 33(1): 1–21.

Glanzberg, M. (2001). The liar in context. *Philosophical Studies*, 103(3): 217–251.

Hazen, A. (1987). Contra buridanum. *Canadian Journal of Philosophy*, 17(4): 875–880.

Kaplan, D. (1968). Quantifying in. *Synthese*, 19(1–2): 178–214.

Kaplan, D. (1995). A problem in possible worlds semantics. In Asher, W. S. a. D. R. a. N., editor, *Modality, Morality and Belief: Essays in Honor of Ruth Barcan Marcus*, pages 41–52. Cambridge University Press.

Kripke, S. A. (2005). Russell's notion of scope. *Mind*, 114(456): 1005–1037.

Kripke, S. A. (2011). A puzzle about time and thought. In Kripke, S. A., editor, *Philosophical Troubles*. Collected Papers, Vol. I. Oxford University Press.

Lewis, D. (1969). *Convention: A Philosophical Study*. Wiley-Blackwell.

Lewis, D. K. (1976). Survival and identity. In Rorty, A. O., editor, *The Identities of Persons*, pages 17–40. University of California Press.

Maudlin, T. (2004). *Truth and Paradox: Solving the Riddles*. Oxford University Press.

Nolt, J. (2018). Free logic. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2018 edition.

Parsons, C. (1974). The liar paradox. *Journal of Philosophical Logic*, 3(4): 381–412.

Priest, G. (1991). Intensional paradoxes. *Notre Dame Journal of Formal Logic*, 32(2): 193–211.

Priest, G. (2005). Truth and paradox. *Journal of Philosophy*, 102(9): 483–486.

Prior, A. N. (1961). On a family of paradoxes. *Notre Dame Journal of Formal Logic*, 2(1): 16–32.

Prior, A. N. (1971). *Objects of Thought*. Clarendon Press.

Quine, W. V. (1956). Quantifiers and propositional attitudes. *Journal of Philosophy*, 53(5): 177–187.

Rayo, A. and Yablo, S. (2001). Nominalism through de-nominalization. *Nouˆs*, 35(1): 74–92.

Sainsbury, M. (2005). *Reference Without Referents*. Clarendon Press.

Salmon, N. (2010). Lambda in sentences with designators. *Journal of Philosophy*, 107(9): 445–468.

Scharp, K. (2013). *Replacing Truth*. Oxford University Press.

Schiffer, S. (1972). *Meaning*. Clarendon Press.

Simmons, K. (1993). *Universality and the Liar: An Essay on Truth and the Diagonal Argument*. Cambridge University Press.

Tucker, D. (2018). Paradoxes and restricted quantification: A nonhierarchical approach. *Thought: A Journal of Philosophy*, 7(3): 190–199.

Tucker, D. and Thomason, R. H. (2011). Paradoxes of intensionality. *Review of Symbolic Logic*, 4(3): 394–411.

Williamson, T. (2003). Everything. *Philosophical Perspectives*, 17(1): 415–465.

Zardini, E. (2008). Truth and what is said. *Philosophical Perspectives*, 22(1): 545–574.

11 Infinite Types and the Principle of Union

J. P. Studd

The theoretical need for objects is well established. Hardly any systematic theorizing avoids the need to deploy first-order quantification ranging over this type of entity. But should we also countenance *other types* of quantification?

Let us assign *type 0* to anything that falls in the domain of a first-order quantifier. Following Frege, we reserve the term 'object' for entities of this type. The question is then whether we should admit into our ideology quantification over entities of types other than type 0? For example, is the intended semantic value of 'object' itself an object? Or is it an entity of some other type? To employ some Fregean loose-talk:[1] should we take predicates to refer to *type 1 concepts* (under which fall the objects that satisfy the predicate). And should we countenance *second-order* quantification over type 1 concepts. Or should we alternatively draw on plural quantification? Is an arbitrary extension to be encoded as multiple objects—abusing grammar: as a *plurality*? What about other types? Should we accept *third-order* quantification over *type 2 concepts* under which type 1 concepts fall? Or superplural quantification over *superpluralities*, each comprising multiple pluralities? What about type 3, or type 4, and so on? Should we countenance quantification over type ω concepts under which entities of any finite type fall? Or type $\omega + 1$? Where does the type hierarchy give out?

Øystein Linnebo and Agustín Rayo (2012) argue that 'plausible' assumptions lead to a surprising conclusion: one should countenance a proper-class-sized infinity of *sui generis* types (p. 276).[2] This chapter takes up their argument for this thesis—Infinite Types—and argues that one of its assumptions is rather less plausible than Linnebo and Rayo suggest. The assumption that one should countenance any language which 'pools together' the expressive resources drawn from any set of languages already deemed legitimate—the Principle of Union—is the subject of Section 11.5. Before we come to that, Section 11.4 attends to a technical glitch in Linnebo and Rayo's argument, which is regimented in Section 11.3. First of all, two preliminaries are in order: Section 11.1 elaborates on the infinite type hierarchy and lays out

Linnebo and Rayo's assumptions more fully; and Section 11.2 introduces the languages of very high order with which Infinite Types is concerned.

11.1 An Infinite Type Hierarchy

The type hierarchy that Infinite Types demands should be sharply distinguished from the cumulative hierarchy of sets, described by Zermelo–Fraenkel set theory. Set theory posits the familiar 'V'-shaped hierarchy comprising proper-class-many *ranks* of sets. Unlike concepts or pluralities, however, sets fall within the range of *first-order* quantifiers. Set theory's extensive *ontology* may be twinned with an austere Quinean *ideology*, which is unprepared to countenance any type of quantification other than first-order.

The distinction, roughly carved, is between what objects there are (ontology) and what expressive resources are legitimate (ideology). To briefly elaborate, consider the different attitudes a typical *pluralist* and *singularist* may take to the following sentences:

(1) Gödel often worked alone.
(2) Linnebo and Rayo collaborate.

The pluralist characteristically takes plural resources seriously.[3] She may maintain, for instance, that the subject terms in (1) and (2) exhibit *different types* of reference. The singular term 'Gödel' *singularly refers* to exactly one object (i.e. Kurt Gödel); the plural term 'Linnebo and Rayo' *plurally refers* to multiple objects (i.e. Øystein Linnebo and Agustín Rayo). The singularist, on the other hand, rejects the idea of *sui generis* plural reference as an expressive resource distinct from—and irreducible to—singular reference. On one version of her view, the plural term 'Linnebo and Rayo' functions semantically as a singular term, *singularly referring* to exactly one object (e.g. the set {Linnebo, Rayo}, or some other plurality-encoding *object*).

Similarly, the singularist takes the singular quantifier 'some logician' and the plural quantifier 'some logicians' to express the *same type* of quantification, albeit over different kinds of object (logicians and objects encoding pluralities of logicians). In contrast, a pluralist may typically maintain that the singular and plural quantifiers express *different types* of quantification—singular and plural—over the same kind of objects (logicians); on this view, once again, plural quantification is irreducibly plural and not a disguised form of singular quantification.

Pluralism posits two types of quantification. Linnebo and Rayo argue that a sober-seeming package of assumptions should lead their advocates to accept a transfinite *ideological hierarchy* of types in addition to—or instead of—an *ontological hierarchy* of sets. The conclusion of their argument may be provisionally stated as follows (p. 276):[4]

> **Infinite Types.** For every ordinal α, finite or transfinite, one should countenance a language of at least order α, equipped with quantifiers ranging over type β entities, for each $\beta < \alpha$.

Notwithstanding the ‘entity’-talk, each type in the hierarchy is to be understood as a further *sui generis* expressive resource distinct from, and irreducible to, lower types.

The argument for Infinite Types draws on three main assumptions (pp. 274, 276, 294):

> **Absolute Generality.** First-order quantifiers may range over an *absolutely comprehensive domain* (i.e. a domain comprising absolutely everything whatsoever).
>
> **Semantic Optimism.** For any legitimate object language, one should countenance a metalanguage that permits one to frame a *generalized semantic theory* for the object language (i.e. a theory which generalizes over every possible interpretation of the object language).
>
> **Principle of Union (Set Version).** For any set-sized collection of legitimate languages, one should countenance its *union language* (i.e. the language which results from ‘pooling together’ the resources of every language in the collection).[5]

The first assumption has received extensive discussion elsewhere. Absolute Generality has clear *prima facie* appeal. Not least since theories such as the first-order theory of identity, Zermelo–Fraenkel set theory with urelements, physicalism, mereological nihilism, and atheism—to name just a few—seem to cry out for an absolutely general formulation. This assumption, however, remains controversial.[6] Here, we shall simply set this debate aside, and follow Linnebo and Rayo in accepting Absolute Generality as a working assumption; we suppose henceforth that there is an absolutely comprehensive domain.

The second assumption—Semantic Optimism—calls for a slightly longer explanation. It’s a familiar point that generalizing over possible interpretations of an object language plays a central role in Tarski’s account of logical validity and its model-theoretic descendants.[7] Suppose, for example, that the object language is the first-order language of Zermelo–Fraenkel set theory with urelements (whose non-logical predicates are the set-predicate *ß*, and the membership-predicate $\in$). Then, in standard model theory, a possible interpretation of this language is encoded as a set-interpretation—typically, for example, as a pair $\langle M,i\rangle$, comprising a (non-empty) set M to serve as the domain, together with an interpretation-function i, which maps the language’s non-logical predicates to their extensions, such that *i*(*ß*) is a set of members of M and $i(\in)$ is a set of pairs of members of M.

Of course, this is not the only way to capture interpretations of the object language in set theory. A minor variant of this implementation, which will be useful below, instead encodes the structure $\langle M,i\rangle$ as a single function—$\llbracket\cdot\rrbracket^0$—which extends i to map the quantifier-symbol $\forall$ to the domain:

$$\llbracket\forall\rrbracket^0 = M \quad \llbracket\beta\rrbracket^0 = i(\beta) \quad \llbracket\in\rrbracket^0 = i(\in)$$

The superscripts here serve to remind us that each of $\llbracket\forall\rrbracket^0$, $\llbracket\beta\rrbracket^0$, and $\llbracket\in\rrbracket^0$—and indeed the function $\llbracket\cdot\rrbracket^0$ itself—is a type 0 entity (i.e. an object).

Whatever encoding we choose, however, standard model theory does *not* qualify as a generalized semantic theory in the sense operative in Semantic Optimism—at least, not if we assume Absolute Generality. For given this assumption, not every possible interpretation is encoded as a *set-interpretation* of the relevant kind. Take, for instance, the interpretation of the object language, where β and $\in$ receive their *intended extensions* based on the absolutely comprehensive domain. This interpretation is not encoded as a set-interpretation $\langle M,i\rangle$ or $\llbracket\cdot\rrbracket^0$. For, according to standard set theory, no set-domain M comprises everything whatsoever, and no set-extension $\llbracket\beta\rrbracket^0$ or $\llbracket\in\rrbracket^0$ comprises every set or every element–set pair (i.e. every pair whose first coordinate is an element of its second).[8]

In order to generalize about every possible interpretation—including ones with non-set-sized domains—we may instead appeal to further ideology in the metatheory. For example, Rayo and Gabriel Uzquiano (1999) show how to frame a generalized semantic theory for a first-order object language in a metalanguage with two types of quantifier, singular and plural. Assuming pluralism, *any* first-order quantifier's domain may be encoded simply as the one or more objects that the quantifier ranges over—speaking loosely: as a *plurality-domain* or *type 1 entity* (which we may write $\llbracket\forall\rrbracket^1$).[9] Similarly, any extension for the language's predicates may be captured with a plurality-extension $\llbracket\beta\rrbracket^1$ or $\llbracket\in\rrbracket^1$, comprising zero or more items, or zero or more pairs of items, drawn from the domain.

Moreover, with a little more coding, Rayo and Uzquiano show that these three pluralities may, in turn, be encoded within a single plurality-interpretation—$\llbracket\cdot\rrbracket^1$—that captures an interpretation of the *whole* language. More generally, any function $\llbracket\cdot\rrbracket^1$ mapping each expression e in a given set E to a corresponding plurality $\llbracket e\rrbracket^1$ may be encoded as the plurality comprising every pair of the form $\langle e,o\rangle$ where $e \in E$ and o is a member of $\llbracket e\rrbracket^1$. On this basis, they show that a plural metalanguage suffices to frame a generalized semantic theory for a standard (non-plural) first-order language, with each possible interpretation of the object language captured with a plurality-interpretation $[\![\cdot]\!]^1$ (Rayo and Uzquiano, 1999, pp. 319–320).

As we shall see in Section 11.3, Linnebo and Rayo's case for Infinite Types goes one step further and argues that, given Absolute Generality, a generalized semantic theory for a first-order object language *must* employ an additional tier of ideology. And this is just the beginning. With further applications of Semantic Optimism, they aim to lead us to any ordinal level of their type hierarchy.

But must we accept Semantic Optimism in the first place? Kreisel's famous 'squeezing' argument provides grounds to think that, for a first-order object language, model-theoretic validity (i.e. truth under all set-interpretations) coincides with true validity (i.e. truth under all possible interpretations).[10] And Linnebo and Rayo concede that they have not offered a 'systematic defence' of Semantic Optimism in their 2012 paper (p. 277).

Nonetheless, two motivations for this assumption have been forthcoming.[11] First, Rayo and Timothy Williamson (2003) argue that Kreisel's argument breaks down for richer object languages, such as those that contain Vann McGee's quantifier $\exists^{AI}$ (formalizing 'there are absolutely infinitely many' or 'there are more than set-many').[12]

Second, a widely accepted approach to natural language semantics effectively calls for a generalized semantic theory in order to systematically describe *the intended interpretation* of the object language.[13] The Mostowski–Barwise–Cooper approach to quantifier-semantics implements a model-theoretic version of Frege's idea that quantifiers denote type 2 concepts (under which fall type 1 predicate-denotations). Assuming a set-domain M, a first-order metatheory permits us to encode an arbitrary predicate-extension based on M as a subset of M and an arbitrary quantifier-extension as a set of predicate-extensions based on M—for example:[14]

$$[\![\text{some sets}]\!] = \{A \subseteq M : |\ [\![\text{set}]\!]\ \cap A| > 0\}$$

$$[\![\text{most sets}]\!] = \{A \subseteq M : |\ [\![\text{set}]\!]\ \cap A| > |\ [\![\text{set}]\!]\ - A|\}$$

But, if we are to extend this approach to object languages that achieve absolute generality, we need a metatheory that is equipped to generalize about arbitrary predicate- and quantifier-extensions based on the absolutely comprehensive domain. And adapting Rayo and Uzquiano's trick for encoding interpretation-functions, this is tantamount to a generalized semantic theory for the object language.

Having briefly outlined these motivations, we shall henceforth assume Semantic Optimism, without further argument. My main interest in the remainder of this chapter concerns the ideological commitments of the package that combines Absolute Generality and Semantic Optimism. In particular, should *optimistic absolutists*, as we shall call those who accept these two assumptions, accept Infinite Types, as Linnebo and Rayo argue?

To answer this question calls for us to examine the Principle of Union. This assumption has so far received comparatively little attention,[15] in part, perhaps, because it has every appearance of being a near-truism. After all, following Linnebo and Rayo, we may argue as follows: assuming that one countenances each language in a given set, as per the antecedent of the Principle of Union, one should also countenance the union language on the grounds that it 'would be made up entirely of vocabulary that had been previously deemed legitimate' (p. 276). On closer examination, however, the Principle of Union is far from trivial. First let's see how we can get from Linnebo and Rayo's three assumptions to the conclusion Infinite Types. This calls for one more preliminary.

11.2 Languages: Simply and Ordinally Typed

Infinite Types calls for optimistic absolutists to countenance languages of very high order. This section introduces these languages. For the sake of concreteness, we shall follow Linnebo and Rayo in defaulting to a hierarchy of types of higher-order quantification into predicate position (glossed in broadly Fregean terms). But the demands made by Infinite Types may equally be met by countenancing other types of ideology. For instance, one may instead adopt a generalized version of pluralism that, in addition to singular quantification, also takes seriously plural quantification, superplural quantification, and so on.[16]

The languages that concern us differ from more familiar typed languages in various ways. Let's begin with a brief review of an example of the latter. A well-known relational formulation of the simple theory of types takes the set of *simple types* to be the least inclusive set that contains 0 and contains $(\tau_1, \ldots, \tau_k)$ for any finite sequence of its members $\tau_1, \ldots, \tau_k$. The simple types may be recursively divided into *levels*: the level of type 0 is 0; the level of type $(\tau_1, \ldots, \tau_k)$ is the least ordinal to exceed the level of each of $\tau_1, \ldots, \tau_k$ (i.e. the maximum level plus one). A small sample of simple types from the first few levels is displayed in Table 11.1.

For example, within this type structure, the types of expression available in a standard first-order language (with no function symbols) belong at levels 0 and 1: individual constants and variables correspond to type 0 constants and variables (written with explicit type indices: a^0, b^0, ... and x^0, y^0, ...); similarly a first-order language's monadic predicates correspond to type (0) constants (now written: $a^{(0)}$, $b^{(0)}$, ...); its dyadic predicates to type (0,0) constants ($a^{(0,0)}$, $b^{(0,0)}$, ...); and so on.

More generally, what we may call *simply typed* or ST-*languages* may include type τ variables and constants for any simple type τ (written: x^τ, y^τ, ... and a^τ, b^τ, ...). An atomic ST-formula is then a string of the

Table 11.1 Some simple types

	monadic	*dyadic*	*triadic*	⋯
level 1	(0)	(0,0)	(0,0,0)	
level 2	((0))	((0),(0))	((0),(0),(0))	
	((0,0))	((0,0,0),0)	⋮	
	⋮	⋮		
level 3	(((0)))	(((0)),((0)))		
	(((0,0,0),0))	⋮		
	⋮			

form $t(t_1, \dots, t_k)$ where each t_i is a term (i.e. a variable or a constant) with simple type τ_i, for $i = 1, \dots, k$, and t is a term with simple type $(\tau_1, \dots, \tau_k)$. Complex ST-formulas are then finite strings formed in the standard way using the usual connectives (¬, →, etc.) and the usual quantifiers (∀ and ∃), which may bind variables of any simple type.

At lower levels, the formulas thus obtained are simply notational variants of their more familiar counterparts. For example, the second-order formula on the left is written in the type-indexed notation as displayed on the right:

$$\forall X(Xa \wedge \exists xAxb) \qquad \forall x^{(0)}(x^{(0)}(a^0) \wedge \exists x^0 a^{(0,0)}(x^0, b^0))$$

We may outline a standard Fregean interpretation for an ST-language as follows. A type 0 constant denotes a type 0 entity (i.e. an object). For a simple type $\tau = (\tau_1, \dots, \tau_k)$, a type τ constant denotes a type τ entity (i.e. an extensional relational-concept in which stand zero or more sequences comprising a type τ_1 entity and $\cdots$ and a type τ_k entity, related in that order). For any simple type τ, a type τ variable then ranges over type τ entities.[17]

The typed languages that Linnebo and Rayo consider differ from ST-languages in three main respects. First, for 'reasons of simplicity', Linnebo and Rayo officially do not 'consider types for functions or polyadic relations' (p. 272).[18] In the context of the simple theory of types, this leaves us with a linear type structure, with just one simple type at each level (permitting us to relabel simple types with finite ordinals in the obvious way):

level or type 0:	0
level or type 1:	(0)
level or type 2:	((0))
level or type 3:	(((0)))
⋮	⋮

This very simple type structure, however, retains the ability to encode polyadic level 1 relations provided we have the resources to encode as a single object each pair, and thus each n-tuple, of objects (e.g. in the Kuratowski-fashion). Equipped with pairing resources, each n-adic level 1 relation may be encoded as a monadic type 1 concept under which fall n-tuples of objects. This familiar trick may be extended to higher level polyadic relations by encoding each n-tuple of level p entities $e_1^p, \ldots, e_n^p$ as a single type p entity: $\langle e_1^p, \ldots, e_n^p \rangle^p$.[19]

The second departure from ST-languages is that Linnebo and Rayo extend the type structure into the transfinite. The class of types is taken to be the class of all ordinals. The languages based on this type structure—which we may call *ordinally typed* or *OT-languages*—may then include (monadic) variables and constants of any ordinal type β (x^β, y^β, ... and a^β, b^β, ...).[20]

Third, OT-languages liberalize the formation rule for atomic formulas. An ST-language is *non-cumulative* in the sense that when s and t are monadic terms with simple types n and m respectively, the string $s(t)$ is a well-formed atomic ST-formula just in case m is the greatest finite ordinal with $m < n$ (i.e. $m = n - 1$). An OT-language is permitted to be *cumulative* in the sense that when s and t are monadic terms with ordinal types β and γ, the string $s(t)$ is a well-formed atomic OT-formula just in case γ is any ordinal with $\gamma < \beta$ (greatest or otherwise).[21] Complex OT-formulas are then finite strings formed in the standard way using the usual connectives and quantifiers, which may bind variables of any ordinal type.

Cumulative OT-languages also admit of a broadly Fregean interpretation. Type 0 constants and variables are interpreted as before. When β is a successor ordinal (i.e. $\beta = \gamma + 1$), a type β constant denotes a (*cumulative*) type β entity (i.e. an extensional concept under which fall zero or more entities with type γ, *or lower*).[22] When λ is a limit ordinal, a type λ constant denotes a type λ entity (i.e. an entity with type $\gamma < \lambda$). For any ordinal β, a type β variable then ranges over type β entities.[23]

Note that the ordinal types are cumulative in the sense that an entity of one ordinal type also qualifies as an entity of all higher ordinal types. Moreover, in the case of a limit ordinal λ, the type λ only comprises entities which also have a lower type.[24]

The order of an OT-language is then measured according to the types of its variables and constants. The argument for Infinite Types is sensitive to the exact characterization of this notion. And we shall later find reason to amend Linnebo and Rayo's official definition.[25] But, provisionally, here is how order is characterized in the main text of their article: 'a language is of *order* α when all of its variables have type-indices below α' (p. 272); moreover 'a language of order α may contain constants of type less than or equal to α' (p. 273). For example: 'The language of the [monadic fragment of the] simple theory of types is a language of order ω, as it has variables of all types below ω' (p. 273).

One last comment is in order: although they officially eschew non-monadic types, Linnebo and Rayo do in practice permit OT-languages to contain a limited stock of *polyadic* predicate-constants. In order to encode pairs in the context of a generalized semantic theory, for instance, they deploy atomic formulas such as $\mathrm{OP}^{\beta+1}(x^0, y^\beta, z^\beta)$ ('z^β encodes the pair $\langle x^0, y^\beta \rangle^\beta$').[26] If polyadic constants are to be included in OT-languages, however, it's important to keep track of their level. This is readily achieved by taking the *unofficial types* to enrich the ordinal types with the type $(\beta_1, \ldots, \beta_k)$ for any finite sequence of ordinal types $\beta_1, \ldots, \beta_k$, with $k > 1$. The level of ordinal type β may then be defined to be β; and the level of an unofficial type $(\beta_1, \ldots, \beta_k)$ to be the least ordinal to exceed each of $\beta_1, \ldots, \beta_k$. For example, $\mathrm{OP}^{\beta+1}$ is a constant with unofficial type $(0, \beta, \beta)$ and level $\beta + 1$ (as indicated by its superscript). The syntax and the Fregean interpretation for OT-languages may be naturally extended to encompass these polyadic constants.[27] Following Linnebo and Rayo, however, OT-languages still lack polyadic variables.

11.3 Linnebo and Rayo's Argument

Preliminaries dealt with, turn now to Linnebo and Rayo's argument.[28] As we noted in Section 11.1, their argument proceeds from the assumptions Absolute Generality, Semantic Optimism, and the Principle of Union, and ends with the conclusion Infinite Types, which we shall henceforth regiment as follows:

Infinite Types. For every ordinal α, finite or transfinite, one should countenance an OT-language of order α, or higher.

The argument for Infinite Types takes the form of a transfinite induction. It's helpful to think of the argument taking place in two stages. First, on the basis of their three assumptions, Linnebo and Rayo offer sub-arguments in favour of three intermediate premisses, which we shall label Base, Successor, and Limit. Infinite Types then follows from the three premisses (assuming a suitable background theory of ordinals that sustains transfinite induction).[29] This section regiments the overarching argument. We shall then return to critically assess the details of the non-straightforward sub-arguments in Sections 11.4 and 11.5.

The first premiss concerns OT-languages of order 1:

Base. One should countenance an OT-language of order 1, or higher. This premiss should be uncontroversial given that standard first-order languages count as OT-languages of order 1.[30]

The second premiss is more contentious:

> **Successor.** If one should countenance an OT-language of order α, or higher, one should also countenance an OT-language of order $\alpha + 1$, or higher.

Linnebo and Rayo's case for Successor deploys two auxiliary theses, the second of which draws on Absolute Generality:

> **Positive Thesis.** It is possible to give a generalized semantics for an OT-language of order α in an OT-metalanguage of order $\alpha + 1$, or $\alpha + 2$ in the case when α is a limit ordinal.

> **Negative Thesis.** It is impossible to give a generalized semantics for an OT-language of order α in an OT-metalanguage of order α.

Granted these two theses (which we return to in Section 11.4), Linnebo and Rayo argue that Semantic Optimism 'motivates ascent' from order α to order $\alpha + 1$, or higher (p. 276). For assuming one should accept an object language with order α, Semantic Optimism requires that one should also accept a metalanguage capable of framing its generalized semantics, and an OT-language can only do so if it has higher order.

The premisses Base and Successor (in conjunction with the principle of mathematical induction for finite ordinals) suffice to establish a weaker version of Linnebo and Rayo's eventual conclusion:[31]

> **Finite Types.** For every finite ordinal n, one should countenance an OT-language of order n, or higher.

This is already a very substantial ideological commitment. Quineans may blanch at the thought of countenancing languages of order 2, let alone admitting into their ideology infinitely many further types of *sui generis*, irreducibly type n, quantification.

With the help of the third and final premiss, however, Linnebo and Rayo aim to push the absolutist's ideological commitments arbitrarily high up the sequence of transfinite orders:

> **Limit.** If one should countenance an OT-language of order α, or higher, for each ordinal α less than a limit ordinal λ, one should also countenance an OT-language of order λ, or higher.

Linnebo and Rayo appear to take Limit to be a variant of the Principle of Union, requiring no further motivation beyond the motivation already given for that assumption.[32] And with this premiss, the argument is complete: Infinite Types straightforwardly follows from Base, Successor, and Limit (in conjunction with the principle of transfinite induction).

Aside of the staggering ideological commitment, Linnebo and Rayo show that the resulting transfinite hierarchy of types has some marked similarities with the cumulative hierarchy of sets. The cumulative nature of predication in OT-languages permits them to define a 'type-unrestricted notion of predication': the defined formula $t^{\gamma} \varepsilon s^{\beta}$ is equivalent to $s^{\beta}(t^{\gamma})$ whenever $\beta > \gamma$, but remains a well-formed formula for any β and γ.[33] Drawing on the work of Degen and Johannsen (2000), Linnebo and Rayo further observe that ε can take over much of the work of $\in$: for sufficiently large α, OT-languages with order α—equipped with a suitable infinitary logic:[34] the *pure cumulative logic of order* α—recover a fairly substantial subtheory of Zermelo–Fraenkel set theory (namely ZF less Replacement).[35]

This leads Linnebo and Rayo to conclude that 'there is no deep *mathematical* difference between the ideological hierarchy of type theory and the ontological hierarchy of set theory' (p. 289). They suggest further that this may lend support to an anti-absolutist position they call *liberalism*.[36] This view rejects Absolute Generality in favour of an open-ended, potentialist conception of the cumulative hierarchy. According to liberalism, given *any plurality-domain* of quantification, the cumulative hierarchy can always be extended to encode the domain as a *set-domain* (which must then lie outside the initial plurality-domain). Linnebo and Rayo write:

> The non-liberalist might come to see the connection between type theory and set theory as a reason for moving in the liberalist direction. For one might have thought that a big selling point of non-liberalism was its tidy ontology: there is no need to countenance an open-ended hierarchy of sets, and no reason to doubt the truth of Absolute Generality. But once one notices that Absolute Generality can be used to motivate ascent into higher and higher levels of the ideological hierarchy, one might come to see the supposed tidiness of non-liberalism as an illusion.
>
> (p. 293)

11.4 Order: Generic, Full, and Cofinal

Of course, how illusory the supposed tidiness of non-liberalism really is depends on the extent to which there is no deep mathematical difference between the ontological and ideological hierarchies, as Linnebo and Rayo claim.[37] And this claim, in turn, relies on the types extending into the transfinite. The non-availability of infinite types would mark a clear and substantive mathematical difference between the two hierarchies and would block the mooted interpretation of set theory. Optimistic absolutists may seek to hold on to their tidy world-view by rejecting one or more of the premisses that lead to Infinite Types.

As we noted, we have no grounds to doubt the good standing of first-order languages, as affirmed in the premiss Base. But that still leaves the other two premisses. What should we make of Successor and Limit? The answer to this question depends on how we understand two of the key terms of art Linnebo and Rayo deploy. A proper assessment of the Principle of Union calls for a closer examination of what it is to *countenance* a language. But first we need to iron out a technical glitch with Linnebo and Rayo's definition of *order*.

The characterization of order that Linnebo and Rayo provide in the main text of their article (quoted in Section 11.2) leaves open whether the operative notion is what we may call *full order* or a less demanding notion of *generic order*:[38]

- An OT-language is said to be a *generic language of order α*, or to have *generic order α*, iff (i) each of its variables is a monadic variable of ordinal type β, with $\beta < \alpha$; (ii) for each $\beta < \alpha$, it has a countable stock of variables of type β; and (iii) each of its constants has level γ with $\gamma \leq \alpha$.
- An OT-language is said to be a *full language of order α*, or to have *full order α*, iff it is a generic language of order α, as per (i)–(iii), and moreover (iv) for each $\gamma \leq \alpha$, it has at least one constant of type γ.

For example, an OT-version of a monadic fragment of the language of the simple theory of types, with countably many variables v^n of each finite ordinal type n (with $n < \omega$) attains generic order ω; but this language does not attain full order ω unless—unlike typical formulations—it also has constants of all finite ordinal types, and at least one constant t^ω with transfinite type ω.

In Appendix B, however, Linnebo and Rayo make clear that the relevant notion of order is *full order*.[39] The importance of not deploying the weaker notion of generic order becomes plain when we attend to the details of their Semantic-Optimism-based case for the premiss Successor. Recall that the ascent from α to $\alpha + 1$ demanded by Successor flows from two theses:

Positive Thesis. It is possible to give a generalized semantics for an OT-language of order α in an OT-metalanguage of order $\alpha + 1$, or $\alpha + 2$ in the case when α is a limit ordinal.

Negative Thesis. It is impossible to give a generalized semantics for an OT-language of order α in an OT-metalanguage of order α.

The Positive Thesis holds for both generic and full order. Linnebo and Rayo build on the Rayo–Uzquiano strategy of encoding the interpretation of a first-order language as a plurality of pairs—$[\![\cdot]\!]^1$—lifting this encoding to higher types, and replacing pluralities with concepts. Suppose the object language is an OT-language of order α (full or generic). Then any constant t^γ in the object language has level $\gamma \leq \alpha$ (by (iii)). Consequently, any possible denotation of t^γ may be encoded as a monadic concept of type γ (exploiting higher-level pairing in the polyadic case). Moreover, for $\gamma \leq \alpha$, an $(\alpha + 1)$-order OT-metalanguage is equipped with type γ variables v^γ (by (ii)); and these variables range over all concepts of this type.

As before, we may then further exploit the metatheory's ability to code n-tuples of type γ entities as further type γ entities, in order to capture an arbitrary interpretation of the *whole* object language as a *single* entity of sufficiently high type. Linnebo and Rayo show that, when α is a successor ordinal, each interpretation of the object language may be encoded as a type α entity—$[\![\cdot]\!]^\alpha$—with the help of type $\alpha + 1$ constants to express semantic notions in the metalanguage. Similarly, when α is a limit ordinal, each interpretation may be encoded as a type $\alpha + 1$ entity—$[\![\cdot]\!]^{\alpha+1}$—with the help of type $\alpha + 2$ constants. Consequently, a metalanguage of either full order $\alpha + 1$ or full order $\alpha + 2$ permits us to frame a generalized semantics for the object language.[40]

It is the Negative Thesis where Linnebo and Rayo make use of the fact that the object language is a *full* language of order α.[41] For then (by (iv)) the object language must contain a constant t^α of type α, which denotes a type α entity. In order to generalize about interpretations, however, an order α metalanguage must use a bound variable with type below α (by (i)). Suppose, then, that for some $\beta < \alpha$ interpretations of the object language are implemented as type β entities—i^β—and write $i^\beta(t^\alpha)$ for the *type* α denotation that is (encoded by) the semantic value of the constant under i^β. In order to attain a generalized semantic theory, every possible denotation for t^α must be equal to $i^\beta(t^\alpha)$ for at least one interpretation i^β; in other words, the function $i^\beta \mapsto i^\beta(t^\alpha)$ maps the type β interpretations *onto* every type α entity. The Negative Thesis may then be established by proving a higher-order version of Cantor's theorem which states, on the contrary, that there is no function mapping the entities of type β onto every entity of type α, whenever $\beta < \alpha$.[42]

The same argument cannot be made when the object language in question is a *non-full* language of generic order α. For in this case there is no guarantee that it contains a type α constant t^α. Indeed, Rayo and Uzquiano show that a generic OT-language of order 2 with no type 2 constants (namely, a second-order version of the language of set theory) can have its generalized semantics framed in a metalanguage of generic order 2 which enriches the object language with a suitable level 2 satisfaction predicate.[43] This provides a counterexample to the Negative Thesis when order is taken to be *generic order*.

So far, then, so straightforward. The argument for the Negative Thesis reaffirms what Linnebo and Rayo had already made clear in their characterization of order in Appendix B: the notion of order in play is full order and not generic order. The trouble is that their argument in favour of the premiss Limit pulls in the opposite direction.

Recall that Limit states that one should countenance a language of limit order whenever one should countenance languages of all lower orders. Linnebo and Rayo's considerations about 'pooling together' resources deemed legitimate (outlined in Section 11.1) provide support for the following thesis:

> **Limit-1.** If one should countenance an OT-language $\mathcal{L}_\alpha$ of order α for each $\alpha < \lambda$, one should also countenance the corresponding *union language*—$\mathcal{L}_\lambda$—the OT-language whose constants and variables comprise each term available in any of the previously countenanced languages $\mathcal{L}_\alpha$, with $\alpha < \lambda$.

This thesis is a straightforward consequence of the Principle of Union.[44] But the Principle of Union does not imply Limit unless we draw on further assumptions, such as the following:[45]

> **Limit-2.** The union language $\mathcal{L}_\lambda$ is an OT-language with order λ.

Should we accept Limit-2? The question is again sensitive to the notion of order. Limit-2 is in good standing for generic order, but not for full order. Consider, for instance, a sequence of full OT-languages of finite order n—$\mathcal{L}_n$—for each $n < \omega$. Each full language $\mathcal{L}_n$ is equipped with countably many variables (x^p, y^p, ...) for each $p < n$, together with the constants $c^0, \ldots, c^n$ (but no other variables or constants). Merging the languages together, the union language—$\mathcal{L}_\omega$—is an OT-language which contains countably many variables and a constant c^q for each finite type $q < \omega$. But it fails to attain full order ω because it lacks constants of type ω. The full-order version of Limit-2 fails.

To briefly take stock: neither notion of order is fit for purpose. To sustain Limit-2 we must opt for generic order, which undermines the Negative Thesis used in the argument for Successor. On the other hand, if we switch to full order, the Negative Thesis is restored to good standing at the expense of undermining Limit-2. The argument for Infinite Types teeters on the brink of equivocation.

One response is to tweak Linnebo and Rayo's definition of order. A notion of order between full and generic sustains both Limit-2 and the Positive and Negative Theses:

- An OT-language is said to be a *cofinally full language of order* α or to have *cofinal order α* iff it is a generic language of order α, as per (i)–(iii), and meets the following condition: (iv′) for each $\gamma < \alpha$, it has at least one constant whose type exceeds γ.

In other words, (iv′) requires that the types exemplified by constants in a cofinally full language be cofinal with the ordinals less than its order. For example, $\mathcal{L}_\omega$ is a cofinally full language of order ω even though it fails to attain full order ω.

When order is taken to be cofinal order, it is straightforward to verify that Limit-2 holds. The Positive Thesis also remains in good standing.[46] Linnebo and Rayo's argument for the Negative Thesis may then be adapted as follows. When α is a successor ordinal, we may apply the same argument as before to show that an object language with cofinal order α—which, in the successor case, is still equipped with at least one level α constant, (by (iv′))—cannot have its generalized semantics framed in another OT-language with cofinal order α. When α is a limit ordinal λ, the argument may be adapted as follows. In order to generalize over interpretations, the metalanguage (with cofinal order λ) must deploy a bound variable v^β of some type β, with $\beta < \lambda$. But the type β interpretations that v^β ranges over are unable to encode every possible interpretation of the object language (also assumed to have cofinal order λ). This is because the object language is equipped with a constant t^γ with type $\gamma > \beta$ (by (iv′)). In order to attain a generalized semantics in this way, every type γ denotation for t^γ needs to be encoded within a type β interpretation. And, as before, this conflicts with the version of Cantor's theorem that states that, for $\beta < \gamma$, there is no way to map the type β entities onto every type γ entity.[47]

Linnebo and Rayo's argument for Successor may then proceed as before: an optimistic absolutist who countenances an OT-language with cofinal order α can frame its generalized semantics in an OT-language with cofinal order $\alpha + 1$, or higher, but not in an OT-language with cofinal order α. And presumably, as before, this 'motivates' ascent to a cofinally full OT-language of order $\alpha + 1$.

But do the Positive and Negative Theses *demand* that the optimistic absolutist countenance a metalanguage of this kind, as called for by Successor? The proposed technical patch highlights a more philosophical concern facing Linnebo and Rayo's argument for Successor. For even once the good standing of the Positive and Negative Theses is secure, the absolutist may wonder why he is required to frame the generalized semantics for the cofinally full ordinally typed object language in *another* cofinally full OT-language of some order or other.[48] Indeed, why must the metalanguage be an extensional, polyadic-variable-free OT-language at all?

Linnebo and Rayo present their (official) eschewal of types for polyadic relations as a simplifying assumption. But without further argument, the absolutist may suspect that by setting aside all types of variable other than those monadic ones which fit neatly into the linear hierarchy of ordinal types, Linnebo and Rayo lay down the rails to infinity which they use to drive him up the type hierarchy. Can the argument for Successor, or something like it, go through without a ban on non-OT-metalanguages?

For a polyadic language whose variable types divide into ordinally indexed levels, such as a simply typed ST-language, the analogue of Successor would call for ascent up the *levels* of the hierarchy. The obvious way to develop a Linnebo-and-Rayo-style cardinality argument for an analogue of the Negative Thesis would then be to deploy something like the following thesis:

> **Successor-1.** There are more type τ entities (encoding the semantic values of a type τ constant with level $\alpha + 1$) than there are entities of any type of level α, or lower.

In the case of an ST-language (for $\alpha < \omega$), some elementary cardinal arithmetic shows that Successor-1 holds provided we assume that the underlying domain of type 0 entities has cardinality κ for some infinite set-cardinal κ. Given Absolute Generality, the assumption that the object language's domain may contain infinitely many objects, while essential,[49] seems reasonable. Although to develop the argument in good conscience, it needs to be shown further that Successor-1 also applies to non-set-sized domains.

On the other hand, OT-languages are not the only way to extend simply typed languages to transfinite levels. Rather than play down polyadic relation types, we might embrace types of infinite adicity. As usual, let a γ-sequence be a sequence whose members are indexed by the ordinals less than γ.[50] Then the class of what we may call *infinite polyadic types* is the least inclusive class that contains 0 and contains $(\tau_\beta)_{\beta<\gamma}$ for any γ-sequence of its members (finite or infinite). The notion of level and the extensional Fregean interpretation for simple types may both be generalized to infinite polyadic types in the natural way, allowing for relational-concepts in which infinite sequences of entities are related.

Allowing for infinite polyadic types, Successor-1 fails. Consider, for instance, a level 2 constant t with type ((0)) and a level 1 variable v whose infinite polyadic type is formed from a γ-sequence of 0s:[51]

$$\underbrace{(0, 0, \ldots)}_{\gamma \text{ times}}$$

In this case, when the underlying domain has infinite cardinality κ and $|\gamma| \geq \kappa$, the number of type ((0)) entities that serve as possible denotations for the constant t does *not* exceed the number of level 1 entities ranged over by the variable v.[52] In this case, Linnebo-and-Rayo-style cardinality considerations do not permit us to show that a generalized semantics calls for us to ascend the *levels* of the type hierarchy.

Linnebo and Rayo briefly consider the possibility of infinite polyadic types. They argue that 'strong pragmatic reasons' speak in favour of linearly-ordered ordinal types: infinite polyadic types represent a 'major complication' of type theory, calling in particular for a system that admits infinitely long strings of quantifiers (p. 281).

This makes clear the kind of motivation that Successor is supposed to enjoy. Linnebo and Rayo's intention is not to force the optimistic absolutist up the levels of the type hierarchy on pain of renouncing either Absolute Generality or Semantic Optimism. Instead, it seems, he is to be enticed to countenance OT-languages of higher and higher order on the grounds that this provides an attractive way to make good on these assumptions. Of course, how strong this enticement is depends on how attractive competing non-OT-metalanguages may be. And it's not clear how this is to be judged, except on a case-by-case basis.

11.5 The Principle of Union

Let's set aside concerns about Successor and return to the Principle of Union, which lies behind Limit (and, in particular, Limit-1). The assumption may be restated as follows:[53]

> **Principle of Union (Set Version).** For any collection of legitimate languages $\{\mathcal{L}_i : i \in S\}$, indexed by a set S, one should countenance the *union language*—$\mathcal{L}_S$—the language obtained by 'pooling together' the expressive resources in each $\mathcal{L}_i$ with $i \in S$.[54]

As we noted in Section 11.1, Linnebo and Rayo take the principle to be 'plausible' (p. 276): assuming the antecedent of the Principle of Union is met, they argue, one should also countenance the union language $\mathcal{L}_S$ on the grounds that it 'would be made up entirely of vocabulary that had been previously deemed legitimate' (p. 276). But they also acknowledge that the Principle of Union is 'non-trivial' (p. 276). This final section argues (with Linnebo and Rayo) that this assumption is indeed far from trivial and (against them) that it is either highly implausible or dialectically ineffective, depending on how the Principle of Union is understood.

To begin with, it's important to distinguish two versions of Linnebo and Rayo's argument. The status of Infinite Types, and its various supporting premisses and assumptions, depends on what it is to *countenance* or to *accept the legitimacy* of a language.[55] The theses deployed in the argument admit of thick and thin readings corresponding to thick and thin interpretations of these locutions.

The relevant distinction is related to David Lewis's (1975) well-known distinction between *languages* (count noun) and *language* (mass term). In the former case, an (interpreted) language is nothing more than a suitable correlation between its expressions and their meanings. The thin sense of 'legitimate' (and the correlative sense of 'countenance') only concerns the existence of the relevant correlation. A language is *thinly legitimate* if there is a (suitably encoded) interpretation-function—$[\![\cdot]\!]$—which maps the (non-logical) expressions in its lexicon to their (intended) semantic values.

The legitimacy of a language in the thin sense need not have any connection with language in Lewis's mass-term sense. In this case, language is something we engage in, 'a form of rational, convention-governed human social activity' (Lewis, 1975, p. 7). The thick sense of legitimate demands that the language's interpretation be suitably related to this kind of human activity. A thinly legitimate language is also *thickly legitimate* provided it is a language that we or moderately idealized versions of ourselves—finite beings free from some of the limitative accidents of our biology—are capable of using and understanding.

The difference between thick and thin legitimacy may be illustrated by adapting Jorge Luis Borges's fantastical tale of the Library of Babel. In our version,[56] the library is infinite in extent, and comprises one or more copies of every possible book that can be written with a single English sentence of no more than 80 characters (with a well-defined semantic value). The books are haphazardly arranged but each is shelf-marked with a unique finite ordinal.

The library induces a *Babellian language*—$\mathcal{B}_\omega$—whose only expressions are sentence letters (s_0, s_1, ...), each of which we stipulate to be interpreted with the semantic value of the corresponding English sentence (ordered according to shelf-mark). The thin legitimacy of this language is witnessed by a type 0 object, the set of pairs $[\![\cdot]\!]_\omega$ that encodes the function that maps each Babellian sentence letter to its stipulated semantic value. However, this language is clearly not thickly legitimate. Finite beings like us *are* able to use and understand a language with an infinite number of sentences if, for example, their semantic values are compositionally generated from a finite lexicon. But, in the case of the Babellian language, even allowing for moderate idealization, we are unable to learn the infinitely many arbitrary correlations that would be required to use and understand the full language.[57]

With this distinction in hand, let's return to the argument for Infinite Types. Does Linnebo and Rayo's conclusion call for us to countenance languages of very high order in the thick sense or merely in the thin one? Some of their formulations suggest that the operative sense of 'countenance' is the thick one (where to countenance a language is to accept its thick legitimacy). For example, they state Infinite Types by writing that 'one should *admit use* of α-level languages in *one's theorizing*, for arbitrary α' (p. 276, my emphasis). A thick reading is required moreover for Semantic Optimism to truly deserve its label. If *we* are to engage in generalized semantic theorizing, the metalanguage needs to be a thickly legitimate one that we can use and understand.

However, the Principle of Union has very little plausibility on the thick reading. Consider again the Babellian language. Assuming we're willing to go along with the thick version of the argument as far as Finite Types, it seems hard not to also grant the thick legitimacy of any *finite fragment* of the Babellian language—$\mathcal{B}_n$—whose vocabulary comprises the first n Babellian sentence letters. After all, a moderately idealized speaker capable of mastering the n different types of *sui generis* higher-order quantifier, $\forall x^0, \ldots, \forall x^{n-1}$, required to use and understand an n-th order language would seem to be equally capable of learning how to use and understand the first n sentences of the Babellian language, $s_0, \ldots, s_{n-1}$, via their English translations. According to the Principle of Union, read thickly, we should also therefore thickly countenance the union language whose vocabulary comprises the sentence letters available in each $\mathcal{B}_n$, with $n < \omega$. But the union language is just the full Babellian language $\mathcal{B}_\omega$, which is not thickly legitimate.

In any case, Linnebo and Rayo ultimately shy away from a thick reading of their argument. Notwithstanding their thick-sounding formulation of Infinite Types, they go on to concede that languages of infinite order—governed by an infinitary logic—are 'very different from the sorts of languages that humans are actually capable of using' (p. 277). This leaves us with the thin interpretation of the argument. On this interpretation, its conclusion is substantially weakened: the thin version of Infinite Types calls only for us to accept that *there are* suitably encoded interpretation-functions for very high order languages.

The thin Principle of Union likewise need not have any connection with language as a social practice in which we engage. The thin legitimacy of the union language is simply a question of whether there is a function that specifies its interpretation. The thin Principle of Union is consequently reminiscent of the kind of union principle available in standard set theory:

> **Set-Theoretic Union.** Suppose that I is a set of indices, and A_i a set for each $i \in I$. Then there is also the union set—$\bigcup_{i \in I} A_i$—whose elements are each element of any A_i with $i \in I$.

Read thinly, some instances of the Principle of Union are unwritten by Set-Theoretic Union. Imagine for example that the thin legitimacy of each finite fragment $\mathcal{B}_n$ of the Babellian language is witnessed by a (type 0) interpretation-function $[\![\cdot]\!]_n$ (encoded as a set of expression–semantic-value pairs $\langle s_i, [\![s_i]\!]_n \rangle$, with $i < n$). In this case, the union set—$\bigcup_{n<\omega} [\![\cdot]\!]_n$—also qualifies as a (type 0) interpretation-function and witnesses the thin legitimacy of the full language $\mathcal{B}_\omega$.

But it should now be an all too familiar point that interpretations cannot always be encoded as set-functions or type 0 entities. Consider again the languages $\mathcal{L}_n$ of full order n, equipped with the constants $c^0, \ldots, c^n$, and their cofinally full union language $\mathcal{L}_\omega$ (introduced in Section 11.4). In light of the Positive and Negative Theses, an arbitrary interpretation of $\mathcal{L}_n$ is always encoded as a type n entity—$[\![\cdot]\!]^n$—but may fail to be realized at lower types. Consequently, Set-Theoretic Union, which deals only with type 0 sets, says nothing at all about whether we may merge together interpretations of higher types.

How, then, are Linnebo and Rayo to persuade their opponent to accept the thin Principle of Union? Suppose, for instance, that she endorses Finite Types, and accepts the legitimacy (thick and thin) of each language $\mathcal{L}_n$ $(n < \omega)$, but has yet to see a good reason to admit type α quantification into her ideology (for $\alpha \geq \omega$). Should Linnebo and Rayo's opponent accept the legitimacy of $\mathcal{L}_\omega$ on the grounds that it is made up entirely of vocabulary already deemed legitimate, as they argue?

We've already seen that this provides no grounds for thick legitimacy. When it comes to thin legitimacy, it's true that Linnebo and Rayo's opponent accepts that each expression of $\mathcal{L}_\omega$ is part of a language $\mathcal{L}_n$ whose thin legitimacy is witnessed by a type n interpretation-function $[\![\cdot]\!]^n$. Suppose for concreteness (following the Rayo–Uzquiano coding outlined in Section 11.1) that $[\![\cdot]\!]^n$ is implemented as a type n entity under which fall the following:[58]

(i) one pair of the form $\langle e,o^0\rangle^0$ (where e is c^0 and o^0 is the constant's intended denotation);

(ii) zero or more pairs of the form $\langle e,o^p\rangle^p$, for each $p < n$ (where e is c^{p+1} and o^p is an entity that falls under the constant's intended denotation).

The availability of type n entities of this kind, however, is not yet enough for $\mathcal{L}_\omega$ to be thinly legitimate. For this we need something further, namely an interpretation for the *whole language*, a function which assigns a semantic value to *every $\mathcal{L}_\omega$-constant* $(c_0, c_1, \ldots, c_{n+1}, \ldots)$. And this is not provided by any interpretation-function $[\![\cdot]\!]^n$, which only encodes semantic values for the $\mathcal{L}_n$-constants $(c_0, \ldots, c_n)$.

All the same, might Linnebo and Rayo simply 'pool together' the various *interpretations* previously deemed legitimate to obtain an interpretation for $\mathcal{L}_\omega$—$\bigcup_{n<\omega} [\![\cdot]\!]^n$—under which falls every pair $\langle e,o^p\rangle^p$ that falls under any $[\![\cdot]\!]^n$, for $p < n < \omega$? The difficulty is that since $\langle e, o^p\rangle^p$ may only become available at type p, such a 'union'-entity is liable to have falling under it entities with arbitrarily high finite type. In this case, $\bigcup_{n<\omega} [\![\cdot]\!]^n$ is available at type $\omega + 1$ since every item falling under it has finite type n and thus type ω (which, recall, comprises every entity with finite type). But it is not itself a type n entity, for any $n < \omega$, since some type p entities, with $p \geq n$, fall under the 'union'-entity. Nor, therefore, is $\bigcup_{n<\omega} [\![\cdot]\!]^n$ a type ω entity.

Consequently, the union principle we require to merge together each interpretation $[\![\cdot]\!]^n$ is something along the following lines:

> **Type-Theoretic Union.** Suppose that I is a set of indices, and $o_i^{n_i}$ an entity with ordinal type $n_i < \omega$ for each $i \in I$. Then there is also a type $\omega + 1$ entity—$\bigcup_{i \in I} o_i^{n_i}$—under which falls each entity that falls under any $o_i^{n_i}$ with $i \in I$.

But what reason has Linnebo and Rayo's opponent to accept a principle of this kind? After all, Type-Theoretic Union baldly asserts that *there are* entities of exactly the infinite types of which she is sceptical. To attempt to persuade their opponent to accept Infinite Types on the basis of Type-Theoretic Union is little better than attempting to argue against a sceptic about sets with infinite rank by simply assuming the Axiom of Infinity.

To briefly take stock, deeming each language $\mathcal{L}_n$ thinly legitimate may well call for Linnebo and Rayo's opponent to admit into her ideology entities of arbitrarily high finite type. But she certainly does not *thereby* countenance any type $\omega + 1$ union-entity witnessing the thin legitimacy of the union language. Nor have we yet seen a non-question-begging argument in favour of her doing so.

In fact, when we reflect on the would-be argument's conclusion, it's hard to see how Linnebo and Rayo could *give* such an argument. Sooner or later, we must dispense with loose 'entity'-talk. But if we are genuinely to *state* that there is a type $\omega + 1$ entity $\bigcup_{n<\omega} [\![\cdot]\!]^n$—rather than hoping to pragmatically convey this higher-order thesis with metaphorical 'entity'-talk—we need to *use* a language equipped with type $\omega + 1$ variables. And no argument Linnebo and Rayo might frame in a language of infinite order is apt to persuade their opponent of this conclusion. For their ability to *give* such an argument for the thin legitimacy of an infinite-order language *presupposes the thick legitimacy* of the language in which it is framed. However, to repeat, Linnebo and Rayo acknowledge that finite beings like us are unable to use infinite-order languages of this kind.

Where does this leave the optimistic absolutist? We saw some reasons to contest Linnebo and Rayo's argument for the thesis Finite Types in Section 11.4. Even if we follow them this far, however, their argument for Infinite Types makes essential use of the Principle of Union, and despite its truistic-seeming appearance I've argued that it's far from trivial, in both its thick and thin versions.

Notes

[1] Notoriously, talk of 'concepts' in English, if it speaks about anything, speaks about *objects* of a certain kind. This loose-talk must consequently be taken as elliptical for a suitable paraphrase in a higher-order language. See, for instance, Williamson (2003, pp. 458–459). Analogous remarks apply to 'plurality'-talk. See, for instance, Studd (2019, pp. 77–79).

[2] Page references are to Linnebo and Rayo (2012) unless indicated otherwise.

[3] Pluralist and singularist attitudes may be implemented in various different ways. Compare, for instance, Yi (2006), McKay (2006), and Oliver and Smiley (2013).

[4] A more precise regimentation follows in Section 11.3.

[5] This assumption weakens the 'Principle of Union (Strengthened Version)' stated on p. 294 of Linnebo and Rayo's (2012) article, taking the 'definite collection' it mentions to be a set-sized collection. Linnebo and Rayo's strengthened version also envisages larger definite collections encoded as entities higher up the type hierarchy (e.g. pluralities, superpluralities, and so on). But here we focus exclusively on the weaker set-sized version. As will become clear in Section 11.4, it's important to distinguish both the set and strengthened versions of the Principle of Union from the thesis given this name on p. 276, which we relabel 'Limit'. We henceforth reserve the label 'Principle of Union' for the set version stated here.

[6] See, for instance, Williamson (2003), and the chapters collected in Rayo and Uzquiano (2006); Studd (2019) makes an extended case against Absolute Generality.

[7] See Williamson (2003 pp. 425–426).

[8] Compare, for instance, Linnebo and Rayo (2012, p. 275) and Studd (2019, pp. 69–72).

[9] We use the prefix 'plurality-' to indicate the need for a plural paraphrase.

[10] See Kreisel (1967).

[11] Opponents of Absolute Generality often take arguments against an absolutely comprehensive domain to also rule out a domain comprising absolutely every set or absolutely every interpretation. But they may still endorse a suitably restricted version of Semantic Optimism. See Studd (2019, pp. 80–81).

[12] Rayo and Williamson (2003, pp. 337–338), Rayo (2006, p. 245); see Studd (2019, pp. 83–84) for critical discussion.

[13] See Studd (2019, pp. 84–85).

[14] See Barwise and Cooper (1981).

[15] Florio and Shapiro (2014) offer critical discussion (especially of the strengthened version of the Principle of Union mentioned in note 5); Linnebo and Rayo (2014) reply.

[16] Linnebo and Rayo outline this hierarchy in an appendix to their article, pp. 297–299.

[17] The term 'type' is ambiguous between *expression type* and *entity type*. The intended disambiguation is usually obvious from the context. Occasionally it will be important to remember that constant-symbols and variables (of any expression type) are standardly themselves taken to be objects (with entity type 0).

18. Linnebo and Rayo add that 'with one possible exception' the inclusion of these types would 'not substantially change' their philosophical arguments (p. 272). We return to the exception in Section 11.4.
19. Linnebo and Rayo outline such an encoding in Appendix B, pp. 304–306.
20. See p. 272.
21. See p. 273.
22. When β is a finite ordinal, and τ is the monadic simple type of level β, (cumulative) ordinal type β entities should not be confused with the (non-cumulative) simple type τ entities (deployed in the context of simple type theory). Terms such as 'type β entity' are henceforth used in the cumulative way. And 'type' henceforth means 'ordinal type' unless indicated otherwise.
23. Compare pp. 273, 297.
24. The fact that type λ variables consequently range only over entities with type $\gamma < \lambda$ is important in light of the 'limit rule' deployed in the logic Linnebo and Rayo take to govern OT-languages. See note 34.
25. See Section 11.4.
26. See Appendix B, for instance p. 304.
27. Linnebo and Rayo also treat these cumulatively: for instance, the predicate $\text{OP}^{\beta+1}$, with unofficial type $(0, \beta, \beta)$, yields a well-formed atomic formula when combined with (monadic) terms of respective ordinal types γ_0, γ_1, and γ_2 where $\gamma_0 \leq 0$ and $\gamma_1, \gamma_2 \leq \beta$. See, for instance, p. 304.
28. See pp. 275–276.
29. Throughout we assume a background theory which includes at least Zermelo–Fraenkel set theory with Choice (ZFC).
30. Compare p. 275.
31. Compare p. 275.
32. Indeed, Linnebo and Rayo employ the label 'The Principle of Union' for a minor variant of Limit on p. 276, providing the brief motivating argument outlined in Section 11.1. For the present use of this label, see note 5. We return to Limit in Section 11.4.
33. See pp. 281–283.
34. In addition to extensionality axioms for all ordinals below α and impredicative comprehension axioms for successor ordinals below α, the system includes an infinitary 'limit rule', which permits us to infer $\forall x^{\lambda}\varphi(x^{\lambda})$ from $\{\forall x^{\gamma}\varphi(x^{\gamma}) : \gamma < \lambda\}$ for any limit ordinal λ below α. See pp. 288–289 for details.
35. See Proposition 2, p. 289.
36. See pp. 290–293.
37. The absence of Replacement in Linnebo and Rayo's target set theory is also noteworthy here. See p. 289, note 28.
38. Linnebo and Rayo leave clause (ii) tacit on pp. 272–273, but their statement of the 'Principle of Union' (on p. 276) makes it clear that it is intended. Clause (ii) is necessary to avoid trivializing Infinite Types. Unlike Linnebo and Rayo's characterization, moreover, clause (iii) makes explicit provision for polyadic predicate-constants. The constraint on their level is important when we come to argue in favour of the Positive Thesis.
39. See p. 299.
40. Compare pp. 300–308. The Positive Thesis for generic order immediately follows since the metalanguage also has generic order equal to its full order.
41. The argument is intended to be understood with first-order quantifiers ranging over the absolutely comprehensive domain, and higher-order quantifiers ranging unrestrictedly over all suitably typed entities based on this underlying first-order domain. This appeal to Absolute Generality prevents type α entities based on one first-order domain being encoded as lower-typed entities based on a larger first-order domain.

42. This is a straightforward generalization of the version of Cantor's theorem whose proof is outlined by Linnebo and Rayo on pp. 299–300. In addition to the pure cumulative logic of order α, the proof makes use of the pairing resources they outline in Appendix B.2, pp. 304–306.
43. See Rayo and Uzquiano (1999, pp. 320–322). Compare Rayo and Uzquiano (2006, p. 244).
44. In the ambient background theory—see note 29.
45. We also assume that sublanguages of legitimate languages are legitimate.
46. This is a corollary of the Positive Thesis for full order since the generalized semantics for a full language of order α induces a generalized semantics for any of its cofinally full sublanguages (and the full metalanguage also qualifies as cofinally full).
47. The argument here relies on the fact that Semantic Optimism calls for us to generalize over interpretations of the *whole language*. A more limited optimism fails to motivate Successor. For example, the cofinally full language $\mathcal{L}_\omega$ *is able* to generalize over arbitrary interpretations of any *finite set* of $\mathcal{L}_\omega$-sentences (since this is also a set of $\mathcal{L}_n$-sentences for sufficiently large $n < \omega$).
48. An exactly analogous question arises, of course, if order is taken to be full order, as on Linnebo and Rayo's official characterization in Appendix B.
49. When the underlying domain has cardinality 2, for example, there are sixteen (extensional) level 2 entities of type ((0)) and the same number of level 1 entities of type (0,0).
50. As usual, we may identify a γ-sequence with a function whose domain is $\{\beta : \beta < \gamma\}$ and which maps each ordinal less than γ to the member of the sequence it indexes.
51. In other words, the variable has type $(\tau_\beta)_{\beta < \gamma}$ where $\tau_\beta = 0$ for each $\beta < \gamma$.
52. Working in ZFC, let $\mu = |\gamma|$. Then the number of (extensional) type ((0)) entities is 2^{2^κ} and the number of level 1 entities of the type (0, 0, . . .) formed from a γ-sequence of 0s is 2^{κ^μ}. Moreover, when $\omega \leq \kappa \leq \mu$, we have that $2^{2^\kappa} \leq 2^{\kappa^\mu}$.
53. This is equivalent to the formulation from Section 11.1 in our background theory (which, recall, includes ZFC).
54. How are we to pool together languages which differ on the interpretation of a common expression? One option would be to include both disambiguations in the union language. Here we shall sidestep this question by focusing on cases where the languages to be pooled-together agree on the interpretation of their common expressions.
55. We use these terms interchangeably.
56. Borges's library has a more stringent book format and allows nonsense strings.
57. We assume here that the haphazard arrangement of books is such that there is no effective method—so long as we remain ignorant of the contents of the library—for us to read off the English sentence (or its semantic value) from the corresponding Babellian sentence letter.
58. These pairs are always available at a type below n (as indicated by their type indices). The first coordinate is an expression, and therefore an object with *entity type* 0 (notwithstanding its *expression type*—see note 17). The second coordinate has (entity) type below n.

References

Barwise, J. and Cooper, R. (1981). Generalized quantifiers and natural language. *Linguistics and Philosophy*, 4(2): 159–219.

Degen, W. and Johannsen, J. (2000). Cumulative higher-order logic as a foundation for set theory. *Mathematical Logic Quarterly*, 46(2): 147–170.

Florio, S. and Shapiro, S. (2014). Set theory, type theory, and absolute generality. *Mind*, 123(489): 157–174.

Kreisel, G. (1967). Informal rigour and completeness proofs. In Lakatos, I., editor, *Problems in the Philosophy of Mathematics*, pages 138–171. North Holland.

Lewis, D. (1975). Languages and language. In Gunderson, K., editor, *Minnesota Studies in the Philosophy of Science*, volume VII, pages 3–35. University of Minnesota Press.

Linnebo, Ø. and Rayo, A. (2012). Hierarchies ontological and ideological. *Mind*, 121(482): 269–308.

Linnebo, Ø. and Rayo, A. (2014). Reply to Florio and Shapiro. *Mind*, 123(489): 175–181.

McKay, T. (2006). *Plural Predication*. Oxford University Press.

Oliver, A. and Smiley, T. (2013). *Plural Logic*. Oxford University Press.

Rayo, A. (2006). Beyond plurals. In Rayo and Uzquanio 2006.

Rayo, A. and Uzquiano, G. (1999). Toward a theory of second-order consequence. *Notre Dame Journal of Formal Logic*, 40(3): 315–325.

Rayo, A. and Uzquiano, G. (2006). *Absolute Generality*. Oxford University Press.

Rayo, A. and Williamson, T. (2003). A completeness theorem for unrestricted first-order languages. In Beall, J. C., editor, *Liars and Heaps: New Essays on Paradox*, pages 331–356. Oxford University Press.

Studd, J. P. (2019). *Everything, More or Less: A Defence of Generality Relativism*. Oxford University Press.

Williamson, T. (2003). Everything. *Philosophical Perspectives*, 17(1): 415–465.

Yi, B.-U. (2006). The logic and meaning of plurals. Part II. *Journal of Philosophical Logic*, 35(3): 239–288.

Contributors

Editors

Carlo Nicolai is Lecturer in Philosophy (Logic) at King's College London. He received his DPhil at the University of Oxford in 2014. He was Marie Skłodowska-Curie Individual Fellow at the Munich Center for Mathematical Philosophy, and VENI NWO Research Fellow at the University of Utrecht. Nicolai's works involve theories of truth and consequence, modality, the epistemology of logic and mathematics.

Johannes Stern is a Research fellow and permanent member of staff at the University of Bristol where he directs the ERC Starting Grant *Truth and Semantics*. Johannes specializes mostly in theories of truth and modality, and topics at the intersection of logic and the philosophy of language. His publications include a single-authored monograph entitled *Toward Predicate Approaches to Modality*.

Contributors

Andrew Bacon is Associate Professor at the University of Southern California. He received his DPhil in Philosophy at the University of Oxford. Bacon's main interests are in metaphysics, epistemology, the philosophy of language and philosophical logic. Bacon has recently completed a book on vagueness entitled *Vagueness and Thought*.

Paul Égré is a Senior Researcher at CNRS, based at Institut Jean-Nicod in Paris and a Professor in the Philosophy Department at ENS. His research deals with logic, language, epistemology, and cognitive science. Much of Égré's work concerns vagueness in language and in perception.

Volker Halbach is Professor of Philosophy and Tutorial Fellow of New College, University of Oxford. He received his PhD at LMU, Munich, and was then Assistant Professor at the University of Konstanz. His research interests are in logic, philosophy of mathematics, philosophy of language, and epistemology.

Leon Horsten is Professor for Theoretical Philosophy with special emphasis on Metaphysics, Epistemology, and Logic at the University of Konstanz. Before Konstanz, he was Professor of Philosophy at the University of Bristol. Horsten's research brings formal methods to bear on philosophical problems. The formal methods involved are drawn not only from philosophical and mathematical logic, graph theory, but also from other parts of mathematics and probability theory.

Richard Kimberly Heck is Professor of Philosophy at Brown University. Heck received their PhD from MIT in 1991. Heck works on historical, conceptual, and technical problems emerging from the philosophy of Gottlob Frege, especially Frege's philosophy of arithmetic. Their two books *Frege's Theorem* (Oxford University Press, 2011) and *Reading Frege's Grundgesetze* (Oxford University Press, 2012) are both focused on that topic. Heck has also worked extensively on philosophy of language, philosophy of logic, and philosophy of mind. Most recently, they have been working on a range of issues concerning gender and sexuality.

Julien Murzi is Associate Professor in Philosophy at the Philosophy Department (KGW) at the University of Salzburg, and an external member of the Munich Centre for Mathematical Philosophy. Murzi specializes in the philosophies of language and logic, but also has serious interests in logic proper, metaphysics, epistemology, and the philosophy of mathematics. He is running a four-year FWF project on semantic paradox, titled The Liar and Its Revenge in Context.

Lavinia Picollo is a Lecturer in Philosophy at University College London. She received her PhD from the University of Buenos Aires in 2015. After that she spent a year as a postdoctoral fellow at the Center for Advanced Studies and two years as an Assistant Professor at the Munich Center for Mathematical Philosophy (MCMP), both at LMU Munich. Picollo works on philosophical logic, formal metaphysics, and the philosophy of logic and mathematics.

Lorenzo Rossi is Assistant Professor at the Munich Center for Mathematical Philosophy (MCMP), LMU Munich. His areas of specialization are Logic, the Philosophy of Logic, the Philosophy of Language, and the Philosophy of Mathematics. Before joining the MCMP, he obtained a DPhil at the University of Oxford he was a post-doc at the Department of Philosophy (KGW), University of Salzburg.

Thomas Schindler is Research Associate at the Department of Philosophy, University of Bristol, where he is part of the ERC funded project Truth and Semantics. Schindler obtained his PhD at the Munich Center for Mathematical Philosophy (MCMP) at LMU Munich. He works mostly in philosophical logic, metaphysics, philosophy of language, and philosophy of mathematics.

J. P. Studd is an Associate Professor at the University of Oxford, and a Fellow of Lady Margaret Hall. He works primarily in logic, metaphysics, and the philosophy of mathematics.

Matteo Zicchetti is a PhD student at the University of Bristol. He obtained his Bachelors Degree at the LMU Munich.

Index