Reflection in the Mathematical Sciences

Leon Horsten

Key words and phrases. Truth, Mathematical Knowledge, Justification, Epistemic Entitlement, Acceptance, Implicit Commitment, Reflection Principle, Paradox

Contents

Preface

This book is the result of the emergence of important connections between seemingly disparate research fields. Between the end of my postdoctoral fellowship and about 2014, I divided my research time mostly between thinking about truth theory on the one hand, and thinking about questions in the philosophy of mathematics on the other hand. These seemed to me two distinct areas that are not deeply connected to each other. Conversations with Philip Welch about set theoretic reflection principles on the one hand, and my research collaboration with Graham Leigh—based on prescient observations by Volker Halbach—on the other hand, convinced me otherwise. It gradually dawned on me that there are deep connections between truth theory and proof theoretic reflection principles, and I also came to see how the notion of truth plays a pivotal role in set theoretic reflection principles. Set theoretic reflection principles had already played a modest but not insignificant role in the philosophy of mathematics since the later work of Kurt Gödel. An idea that Feferman explored from a proof theoretical perspective since the 1960s, but which was not really taken up by the philosophical community, suggested that proof theoretic reflection principles are also of profound importance for the philosophy of mathematics. So it all in a sense came together.

In the next five years or so, I had the pleasure to investigate, in collaboration with a number of people, aspects of these connections between truth, reflection, and the philosophy of mathematics. During this time, a group of young, talented researchers joined the research efforts in this area. They have taken the emergent interdisciplinary field "to the next level", as they say. Even though the field is young, it is difficult to keep up.

The different parts of the research area that is gradually being carved out are enriching each other. Philosophical theorising has sparked technical research; mathematical and metamathematical research has sparked philosophical discussions. These are exciting times—at least for me.

Both the philosophical and the more technical in the nascent research area has been growing rather rapidly. In particular, new important metamathematical results on proof theoretic reflection principles are currently produced on more or less a yearly basis. The results and contributions are mostly scattered in the philosophical and logical journals. As a consequence of this, it has become more difficult for everyone involved to keep an overview of the state of the art in the field. This scattering of the results also brought with it, I believe, a threat of fragmentation of the field. Fragmentation would be something that the area at the moment can ill afford: the connections between the different parts are the life-blood of the field. Moreover, due to the rapid developments in the area over the past decade, it has become challenging for interested scholars—young and old—to enter the field.

For all these reasons, I took up, sometime in 2019, the project of writing a research monograph on reflection in the mathematical sciences. The plan was to provide an overview of the research that has been carried out in this area in recent years, to push some parts of it forward, and to explain the connections between the different strands. The aim was especially to concentrate on the fruitful back and forth between philosophical and logical research in this area.

I may have bitten off a bit more than I could chew. We are still in the early stages of this new research program. The research in this field is challenging, both from a philosophical and from a technical point of view. It is not easy to achieve conceptual clarity and to trace conceptual connections across disciplines, and it is not easy to gauge the philosophical import of technical results and developments. As a consequence, this book ended up being more heterogeneous that I had wished, and contains numerous loose ends that I wish I'd know how to pursue further or tie together. Moreover, I undoubtedly have taken wrong turns at several junctures.

But this is not the place to make excuses. One just cannot expect a smooth and polished monograph about a research area that is evolving as quickly as the current logico-philosophical debate about reflection principles. I can only hope that this book nonetheless turns out to be a useful resource and guide for philosophers who are actually working on the relation between truth, reflection, and the philosophy of mathematics as well as for those who are merely interested in knowing more about it.

Without working with others on aspects of the subject matter of this monograph over the past decade, I would not have been able to write this book. I therefore thank the following people with whom I had the privilege to collaborate: Marianna Antonutti, Cezary Cieśliński, Martin Fischer, Volker Halbach, Hannes Leitgeb, Graham Leigh, Guanglong Luo, Carlo Nicolai, Sam Roberts, Daniela Schuster, Joanna van der Veen, Philip Welch, Alexandra (Li) Zhang, and Matteo Zicchetti. Aside for this, conversations with and feedback from the following people has been immensely helpful: Carolin Antos, Neil Barton, Luca Castaldo, Silvia De Toffoli, Maciej Glowacki, Kentaro Fujimoto, Daniel Kuby, Beau Mount, Karl-Georg Niebergall, Johannes Stern, Oliver Tatton-Brown, Claudio Ternullo, Sofie Vaas, Mateusz Łełyk, Bas van Fraassen, Albert Visser, Pascal Wagner, Bartosz Wcisło. No doubt I have forgotten to mention certain people who should be included in this list: I beg them to forgive me for this omission. MORE IN PARTICULAR I AM GRATEFUL TO SOFIE AND PASCAL FOR PROVIDING COMMENTS ON CHAPTERS OF THIS BOOK AND TO ANONYMOUS REFEREES .... One person whom I owe gratitude especially, is my wife Hazel Brickhill. She was always more than willing to help me when I was stuck. Many conceptual (and mathematical) confusions have been avoided simply by asking her what she thinks about a given problem or issue.

The encouragement for the book project that I have received within my little research group at the University of Konstanz has kept me going at times when I might have been ready to throw in the towel. Also, I found inspiration for this project in my teaching here in Konstanz. In particular, I fondly remember MA seminars that I taught on *Philo of Alexandria* and on *Epistemic Entitlement* (2020). Aside from this, the discussions in our online reading group on *Gödel's Philosophical Notebooks* has been important in writing this book.

I gratefully acknowledge institutional support that I have received over the past five years for bringing this book project to a (hopefully successful) end. I am indebted to the University of Konstanz for granting me a *Freisemester* from October 2019 until April 2020, as well as for granting me research leave in the Winter Semester of 2023–2024 for completing this monograph. A research stay by Cezary Cieśliński in Konstanz from 15 April to 30 May 2022 contributed much, and so did the research stays in Konstanz of Philip Welch in the Spring of 2023 (sponsored by the *Zukunftkolleg* of the University of Konstanz) and in the Spring of 2024.

Konstanz,
Easter 2024

Introduction

For a long time, reflection was not much of a research theme in analytical philosophy. This situation is now changing. In recent years, reflection has become a fledgling research topic in epistemology. But above all, reflection has become the subject of intensive research in the foundations and philosophy of the mathematical sciences, where I take the mathematical sciences not only to include pure mathematics, but also probability theory. Reflection in the mathematical sciences is the subject of the present research monograph.

In the mathematical sciences, reflection is mainly investigated in proof theory, set theory, and probability theory. It has become clear that the results about reflection that have been obtained in these mathematical disciplines are of considerable importance for the philosophy of mathematics. For this reason, we will see a continuous back and forth between philosophical and 'foundational' discussions about reflection. As such, it is a sustained exercise in mathematical philosophy.

What this book is about

Reflection is in a sense like symmetry. There is a very general and diffuse sense of the term 'symmetry'. Because of the generality and vagueness of this sense of the word, it is not particularly rewarding to take it as the focus of theoretical research. But there are also several more specific and precise concepts of symmetry that have been employed fruitfully in the physical and mathematical sciences. These concepts have long been, and thoroughly deserve to be, the focus of scientific research. Moreover, these more precise concepts of symmetry are of philosophical relevance. They play an important role, for instance, in contemporary discussions in the philosophy of science about the nature and justification of scientific laws and regularities. So it is with reflection. The words 'reflection' and 'reflecting' are often employed simply as synonyms for the words 'thought' and 'thinking'. Such unspecific uses of the word do not express a distinctive concept of reflection. But 'reflection' is also often employed in certain more specific senses. Some of these senses, at least, are of clear philosophical and scientific importance. They are what this book is about.

When it is used in a more specific meaning, the word 'reflection' often expresses some relation between entities. There are several more specific senses of the word 'reflection', depending on the kinds of entities that are related by the reflection relation that the uses express. They occur in the physical world, but also in the abstract world, and in the mental world. We will see that they also occur *between* the mental world and the abstract world, for instance. In sum, reflection phenomena occur in and between different realms. It then comes as no surprise that the investigation of reflection is not confined to any single intellectual discipline.

Physical reflection phenomena are, from a scientific point of view, well understood. Physical reflection (for instance light reflection) is undoubtedly of philosophical relevance, but the investigation of physical reflection remains outside the scope of this book. So we will not be concerned with *all* more specific senses of the word 'reflection'. We will be concerned with reflection relations in and between the abstract world and the mental realm. More specifically, our focus will be on reflection phenomena in the mathematical universe, in the idealised (mathematical) human mind, and between the mathematical universe and the human mind. This is then a book about the philosophy of mathematics, albeit not primarily about the subjects that have dominated the philosophy of mathematics over the past decades (such as structuralism, logicism,...). More specifically, it is mostly a book about *mathematical epistemology*.

Reflection plays particular argumentative roles in intellectual disciplines. Just as use is made of symmetry arguments in various contexts, one also finds what can be called 'reflection arguments' in set theory, philosophy, and theology. Such reflection arguments are (often implicitly) taken to be supported by law-like regularities by which reflection relations are governed. Since the middle of the twentieth century, parts of mathematical logic have been explicitly concerned with uncovering and investigating these *laws of reflection*. Such laws of reflection in the mathematical sciences are commonly called *reflection principles*.

These reflection principles are grounded in the nature of the reflection relations in question. The task of describing the nature of these reflection relations, and discussing our warrant for the reflection principles that govern them, falls to philosophy as special chapters of metaphysics and epistemology, respectively. In particular, they are chapters in the philosophy of mathematics. Perhaps surprisingly, philosophy has until now not fully taken on this responsibility. It is to this task that this book intends to make a contribution: it is devoted to the investigation of (1) *the nature of reflection phenomena in the mathematical sciences*, and (2) *our warrant for what we take to be laws that govern them*.

I distinguish between two types of reflection: (1) *ontological reflection*, and (2) *epistemic reflection*. Ontological reflection is a world-to-world relation; epistemic reflection has a mental component: it encompasses mind-world and mind-mind relations. Accordingly, I distinguish between ontological reflection principles and epistemic reflection principles.

Ontological reflection principles in mathematics are variations on the thought that there are *small* parts of the mathematical universe that are similar to the mathematical universe V as a whole (which itself is too large to be a set). This is often succinctly expressed by saying that the mathematical universe is *reflected* in certain sets. Since similarity is a vague notion, this thought has to be made more precise before it can be mathematically tested. One way of doing this is by formulating the postulate that there are *sets* x such that the structure $\langle x, \epsilon \rangle$ makes exactly the same set theoretic sentences true as the structure $\langle V, \epsilon \rangle$ (where ϵ is the membership relation). Such principles are known as *set theoretic reflection principles*.

A typical example of an epistemic reflection principle is the statement that everything that is formally provable in a certain formal mathematical theory (Peano Arithmetic, for instance), is true. Epistemic reflection principles thus make explicit reference to a background theory, whereas ontological reflections are in a

sense absolute. Variations on this principle are known as *proof theoretic reflection principles*.

In view of Gödel's incompleteness theorems, as long as the background theory S to which a given proof-theoretical reflection principle refers is sound, a reflection principle R_S for S will typically be proof-theoretically independent of S. Similarly, set theoretic reflection principles—except certain weak ones—are also proof-theoretically *independent* of the core basic principles of mathematics, i.e., the axioms of Zermelo-Fraenkel set theory with the Axiom of Choice.

Set theoretic reflection principles and proof-theoretic reflection principles for accepted theories enjoy wide support in the mathematical community. One obvious question, which was explicitly raised by Kreisel and Levy in the 1960s, is whether there are deep conceptual relations between proof theoretic and set theoretic reflection principles. I will address this question, but will devote more time to the epistemological question:

Wherein consists our epistemic warrant for mathematical reflection principles?

Concerning set theoretic reflection, this question received some attention in the philosophy of mathematics community in the wake of Gödel's *post*-World War II views about the *continuum*, i.e., the set of the real numbers. The focus has been on so-called *richness arguments*. The basic thought here is that V is "structurally rich" in the sense that it contains an astounding variety of isomorphism types. It is supposed to be so rich in this sense, that for every collection of sentences S that are true in V, there are *sets* x in V such that the sentences S are also true in x. In this way, richness considerations are intended to give us *reasons* for thinking that reflection principles are true.

We will see that *strong* set theoretic reflection principles are statements not just about sets, but also about *classes*. There is therefore a connection between set theoretic reflection principles and the laws that govern classes. There is also a link between conceptions of classes—or the *nature* of classes, if you will—and the laws that govern them. Thus we will see that there is a tight link between set theoretic reflection principles on the one hand, and the nature of classes on the other hand.

Set theoretic reflection principles have not only been formulated against an *actualist* and *universist* background of set theory. Recently, set theoretic reflection has also been studied against the background of *potentialist* and *multiversist* theories of sets. In this book, these alternative framings of reflection will not be discussed. The reason is not that they are not important. Rather, it is for me a matter of priority. I make the methodological proposal first to try to understand set theoretic reflection in a "classical" framework and only afterwards to extend it to more exotic surroundings.

That being said, I confess that I have always had difficulties with understanding the modality (or modalities) involved in the relevant notion(s) of potentiality that is supposed to be applicable to the mathematical world. Here it bears keeping in mind that potentiality in biology and in physics, as cornerstones of Aristotelian research programs that were pursued for a millennium or so, were ultimately abandoned. Also—the pervasiveness of forcing arguments in set theory notwithstanding,—I have never been convinced that the conception of the mathematical world as a multiverse is well supported by set theoretic practice or by mathematical practice in general. I am well aware, of course, that these are regarded as strong and contentious statements, perhaps not so much so by working mathematicians, but

certainly in contemporary philosophy of mathematics. Moreover, I am well aware that I am making these claims without arguing for them. But it would take us too far if I were to do so here. So I just leave it there: the reader is free just to ignore these remarks.

Concerning proof theoretic reflection, the situation is importantly different. Feferman suggested since the 1960s that there is a tight connection between our epistemic warrant for mathematical theories on the one hand, and our epistemic warrant for proof theoretic reflection principles for them on the other hand. If this is the case, then the correct epistemic account of proof theoretic reflection will be importantly different from the correct epistemic story of set theoretic reflection. Feferman argues that our epistemic warrant for proof theoretic reflection principles for a mathematical theory S is in a sense "implicitly contained" in our warrant for S. Because of the fact that (for reasonable S) reflection principles are typically logically independent of S, this view seems *prima facie* puzzling. Normally, having epistemic warrant for a theory S does not automatically generate warrant for principles that are independent of S: more epistemic work needs to be done. So Feferman's suggestion is in need of further clarification. Unfortunately for us, Feferman did not spell out his thoughts on this in philosophical detail, and his suggestion was not immediately taken up by the philosophical community.

Feferman's suggestions have foundational implications. In particular, they have motivated much of Feferman's own work on *predicativism* about mathematical analysis. The idea is, roughly, that one starts with a theory that is acceptable from a predicative point of view, such as a standard second-order system of arithmetic with comprehension restricted to first-order formulas (with parameters). Accepting this system *implicitly commits* the predicative mathematician to accepting certain further principles. These principles can then be added explicitly, giving rise to a stronger theory S_1. This process is then repeated, leading to successively stronger systems S_2, S_3, and so on. Thus from a predicatively acceptable starting point one can *bootstrap* to stronger theories that still are predicatively acceptable.

The proof-theoretic and set theoretic reflection principles that are studied today did not simply drop from the sky sometime in the middle of the previous century. Proof-theoretic reflection has distant roots in philosophy, where epistemic reflection has for intermittent periods been an active research theme. Ontological reflection has also been a research theme of sorts, not only in the history of philosophy, but also in the history of the part of theology that is called *rational theology*. So these investigations took place not primarily in the philosophy of mathematics, but in the philosophico-theological investigation of the human mind and of the mind of God. The general concepts of epistemic and ontological reflection, as well as patterns of argumentation that have become standard in contemporary proof theory and set theory, thus have a philosophical history. In this book, we will investigate how theories of ontological and epistemic reflection developed in philosophy and (to a much smaller extent) in rational theology. One of the aims is to re-connect epistemological questions concerning reflection principles in contemporary mathematics to their historical roots.

We will see that set theoretic reflection principles are best regarded as *basic* principles: as axioms. Set theoretic reflection principles are fairly widely (but by no means universally!) accepted in the set theoretic community. A fundamental

question then is: wherein does our epistemic warrant for set theoretic reflection principles consist?

As mentioned above, richness arguments have been adduced in support of ontological reflection principles. In fact, we will see that this strategy has an antecedent of sorts in the history of rational theology, namely in the work of Philo of Alexandria. Richness arguments are often taken to constitute *intrinsic evidence* for reflection principles; they are complexes of *philosophical reasons* for reflection principles.

One fundamental question is why it should be rational in the first place to believe that the set theoretic universe displays the relevant richness. But regardless of this, many mathematicians who accept these reflection principles do not accept them *on the basis of* richness arguments. If reflection principles indeed are *basic* principles, then, as far as *mathematical* reasons go, they are rock bottom. It then seems that mathematicians who accept reflection principles as somehow basic, but do not accept them on the basis of philosophical reasons, do not accept them on the basis of reasons at all. Since justification is a matter of having good reasons, this means that these mathematicians do not have a *justification* for their belief in set theoretic reflection principles.

Nonetheless, such mathematicians might still be epistemically warranted to believe reflection principles. I will argue that mathematicians are mathematically warranted in believing certain basic axioms if these are immediate *epistemically optimal responses* to the mathematical challenges that they are confronted with. These warrants are warrants without reasons. As adumbrated earlier, concepts of epistemic warrant that are not reason-based have been investigated in recent epistemology. They go back to work of Tyler Burge and are called *epistemic entitlements*. I will defend the thesis that the mathematical warrants that mathematicians have not only for believing in reflection principles, but in basic mathematical axioms generally, are epistemic entitlements. In particular, the question of mathematical warrant for basic mathematical axioms will be connected with a variant of Crispin Wright's theory of *entitlement of cognitive project*.

The *main* focus of this book will thus be on the question of *warrant for reflection principles in the mathematical sciences*. This research theme is—belatedly, I would say—beginning to attract considerable attention in the philosophy of mathematics. The number of research articles that are concerned with it has steadily been growing over the past decade, and it is not easy to maintain a clear overview of the *status quaestionis*. I will review and critically appraise the most important contributions that have been made to this subject over the past decade, relate them to each other, and to go beyond them by attempting to make some further contributions of my own.

This research is technically, but above all philosophically, challenging. The contemporary debate about warrant for epistemic reflection principles is not closely connected to the history of philosophical thought in centuries past. Rather, it is deeply intertwined with recent work on other philosophical research themes that is difficult in itself and that is fast developing. To see why and how this is so, we have to return to Feferman's thought that when we *accept* a mathematical theory S, we are *implicitly committed* to accepting variations on the principle stating that everything that is provable in S, is *true*.

It is tempting to assume that acceptance, in this context, simply amounts to (propositional) *belief*. Some of Feferman's own writings point in this direction. Nonetheless, I believe that this temptation should be resisted. It is thus incumbent on me to explain what should be meant by acceptance instead. This is a difficult question. It will be addressed it by drawing on philosophical literature on the relation between acceptance and belief, going back to van Fraassen's work on this question going back to the 1980s.

Another difficult and at present not completely resolved question is how the notion of *commitment* should be understood in this context. Moreover, the sense in which one is *implicitly* committed to proof theoretic reflection principles for mathematical theories that one accepts should be made clear. I will argue that the concept of epistemic commitment is actually closely connected to the concept of acceptance that is at play here. For one thing, whereas—*pace* Freud and his followers—it is not clear that *belief* can be implicit in the relevant sense, it is plausible that, because of the more pragmatic nature of the concept of acceptance, a person can implicitly accept a theory.

I will spend a considerable amount of space to examining recent attempts (*post* 2000) to develop Feferman's suggestive but sparse remarks about implicit commitment inherent in the acceptance of theories into detailed epistemic accounts. I believe that, so far, these attempts are only partly successful. My contention—controversial, to be sure—will be that it is easier to explicate the extent and way in which epistemic warrant for *weak* reflection principles such as consistency statements are implicit in what we accept, than to explicate the nature of implicit commitment to stronger proof theoretic reflection principles.

The prototypical and strong proof theoretical reflection principle "All theorems of theory S are true" ostensibly makes use of the concept of *truth*. Indeed, in this reflection principle, truth plays an essential role. We will see that there are also many proof theoretic reflection principles that somehow approximate the prototypical principle but do not make use of the concept of truth. Nonetheless, it has become clear in recent years that the connection between research on proof theoretic reflection on the one hand, and truth theory on the other hand, is deep and important. It therefore plays a major role in this book.

Since the concept of truth is not definable from more fundamental concepts, it is nowadays treated mostly as a *primitive* notion that is governed by basic axioms. There is no agreement on what the basic truth principles are. Roughly, there are two candidate views. On the first view, the primitive truth predicate obeys a collection of *disquotational axioms*, which are axioms of the form

$$\varphi \text{ is true if and only if } \varphi.$$

On the second view, the basic truth principles express the *compositional nature* of the concept of truth. According to this conception, it is a basic feature of the notion of truth that it commutes with the logical connectives, so that for instance the sentence

For all sentences φ and ψ: $\varphi \wedge \psi$ is true if and only if φ is true and ψ is true.

counts as a basic truth axiom. As a rule of thumb, we can say that compositional truth theories are deductively strictly stronger than disquotational truth theories.

It has long been known that if one adds compositional truth axioms to a (sufficiently strong) mathematical theory S, the reflection principle "All theorems of S are true" can be *proved* in the extended theory. This phenomenon has been used to

argue that our epistemic warrant for accepting proof theoretic reflection principles resides in the basic principles governing the truth predicate. If this is right, then compositional truth is more fundamental than proof theoretic reflection. This is sometimes combined with the further view that if one accepts a mathematical theory S, then one is *implicitly committed* also to accept standard compositional truth axioms for the language in which S is formulated. This train of thought can be seen as a way of trying to clarify Feferman's view about our warrant for proof theoretic reflection principles. At the same time, it is an argument for the thesis that the compositional truth axioms are more fundamental than disquotational truth principles, since extending a theory S by disquotational truth does *not* generally result in a theory in which proof theoretic reflection principles for S can be proved.

On the other hand, it has become clear over the past decades that if we start with a disquotational truth theory D instead, and add certain proof theoretic reflection principles for D to it, then the standard compositional truth principles become provable. So if we start with a disquotational truth theory D, and are implicitly committed to reflection principles, then we can say that the compositionality of truth is implicit in disquotational truth. Thus a similar sort of bootstrapping phenomenon seems to occur in axiomatic truth theory as we encounter in predicativism. If this is the right perspective, then the view in the previous paragraph is wrong-headed, and proof theoretic reflection principles are more fundamental than compositional truth principles!

Both views face challenges. Concerning the first view, there is the question how and why someone who accepts a mathematical theory—Peano arithmetic, say—is implicitly committed to accepting the notion of truth in one's conceptual repertoire and that this concept is governed by compositional axioms. Could one not, for instance, reasonably be sceptical about the concept of truth altogether? The second view faces the question how and why we are implicitly committed to proof theoretic reflection principles for theories that we accept. This makes it difficult to adjudicate between the two views. (Both stories could be wrong!) But all this does show that there is a close relation between axiomatic truth theory on the one hand, and proof theoretic reflection principles on the the other hand.

Probabilistic reflection principles will also be discussed. I will concentrate on variants of *van Fraassen's reflection prinicple*. One version of this principle states that one's subjective probability of φ, *given that* the probability of φ is r, should be r. Over the past decades, principles of this kind have been discussed from a philosophical and from a semi-formal perspective in formal epistemology; but they have not been investigated from a proof-theoretic perspective.

One central philosophical question is how, from a conceptual point of view, such reflection principles relate to proof theoretic and to set theoretic reflection principles. A central logical question is what the proof-theoretic properties of such reflection principles are.

These questions are wide open, and unfortunately I will not have a great deal to say about them. Nonetheless, I believe that relating probabilistic reflection principles to other kinds of reflection principles, and especially studying probabilistic reflection principles from a proof-theoretic perspective, has great potential. So I hope that my brief discussion of these matters in this book encourages others to pursue these questions further.

The structure of this book

All of the foregoing is in a fairly straightforward way reflected in the structure of this book, which consists, aside from this Introduction, of three Parts, each consisting of a small number of chapters.

In Part I, the philosophical context of the subject under investigation is explored. I will take stances here that are controversial, and that play an essential role in positions that are developed in Part III. In chapter 1, I discuss the ways in which the traditional concept of doxastic justification can be applied to philosophical questions about epistemic warrant for mathematical beliefs. In chapter 2, notions of non-justificatory warrant (epistemic entitlement) are introduced and discussed. In chapter 3, we turn our attention to epistemic and ontological concepts of reflection, and their evolution in the history of philosophy.

In Part II, the logical, set theoretic, and metamathematical context of the philosophical discussion about reflection principles is described. Background formal mathematical theories and truth theories are discussed in chapters 4 and 5, respectively. Proof theoretic, set theoretic and probabilistic reflection principles are introduced and explored in chapter 6. We will see how proof theoretic and set theoretic reflection principles are importantly different in nature. Proof theoretic reflections will turn out to be *iterable* in ways that set theoretic reflection principles are not, for instance.

All this leads up to Part III, which constitutes the heart of this book. In this Part, the nature of ontological and epistemological reflection phenomena, and our epistemic warrant for proof theoretic and set theoretic reflection principles, are explored. In chapter 7, recent views on epistemic warrant for ontological and epistemological reflection principles are critically examined. In chapter 8, we focus mainly on the weak proof theoretic reflection principle of *consistency*. Concerning this reflection principle, we propose a systematic and detailed phenomenological account of how one can acquire an epistemically entitled belief in the consistency of a theory S from an epistemically warranted belief of S. Chapter 9 deals with questions of epistemic warrant for proof theoretic and set theoretic reflection principles more generally.

In Part IV, I try to look into the future. This is a difficult task, since not only our knowledge of proof theoretic reflection (and its connection to axiomatic truth), but also the philosophical discussion of the topics that are treated in Part III, is developing very rapidly and in different directions. For this reason, Part IV is very short. The only prediction that I am fairly confident of is that this book will be in some ways dated before it appears in print. (I see this as a good thing.)

Aim of this book

One objective of this book is to provide an *overview* of logico-philosophical work on reflection principles in the mathematical sciences. This entails bringing together work that is carried out in different fields and even in different disciplines: history of philosophy, philosophy, proof theory, set theory, formal epistemology, theology. Hopefully this will help to consolidate what I take to be a nascent research area with lots of potential and which is of much importance to philosophy of mathematics, and even to philosophy more generally.

But the aim is also to develop a *philosophical perspective* on the material that is thus brought together, so as to connect the areas and disciplines involved on

a deeper conceptual level. Due to the disparity of results, research backgrounds, and methods involved, the latter has proved to be no easy task, and I do not know whether I have succeeded in this task. On important questions, I have not been able to reach final conclusions, and concerning many philosophical lines of argumentation, I have not been able to reach definite appraisals. Many threads are followed only to some extent, and many issues are left fairly open. In the present state of research in the area, I fear that this is unavoidable.

One prerequisite for achieving these aims is to achieve *conceptual clarity*. Indeed, it seems to me that many of the basic concepts that are involved in the discussion about reflection in the mathematical sciences are in need of explication. It is my hope that this book contributes to our understanding of the relevant fundamental concepts, and how they hang together.

At any rate, this book intends to make research about reflection in the mathematical sciences more accessible to students and scholars who are not familiar with the area and with the wider context of this research. Moreover, I hope that it provides impulses and suggests directions for further research. I will be very pleased if it encourages logicians and formally inclined philosophers to develop some of the themes further. Above all, it would be great if this book contributes to new collaborations between researchers of different disciplines involved in research about reflection.

Prerequisites

The subject of this book lies at the intersection of a number of distinct logical and philosophical areas: proof theory, set theory, axiomatic truth theory, history of philosophy, and theories of epistemic warrant. The view on the nature and epistemology of reflection principles that will be developed involves taking a stance in various philosophical debates in different areas. This means that there are many moveable philosophical parts in the discussions that we will be concerned with. Aside from the entanglement with deep metamathematical theories of proof theoretic and set theoretic reflection principles and with truth theory, it is this that makes matters complicated and controversial—and exciting!

When all is said and done, this is a research monograph on the epistemology of mathematics. Nonetheless, the aim is for this book to be as self-contained as possible. It is intended to be accessible not only to professional philosophers and logicians, but also to postgraduate philosophy students who have an interest in the philosophy of mathematics.

For one thing, this means that considerable logical and metamathematical background information will be supplied. Despite all this, some technical knowledge is presupposed on the part of the reader:

(1) *basic* knowledge of first-order and second-order number theory, set theory, and probability theory;
(2) an intermediate mathematical logic course (including a treatment of Gödel's incompleteness theorems, and rudimentary notions of modal logic);
(3) basic knowledge of axiomatic truth.

The number of mathematical *proofs* that are included in this book is very modest: they are restricted to simple proofs that play a role in the understanding of substantive philosophical issues. Moreover, I have tried to make the technical parts in this book as readable as possible. In particular, in the notation, I have

dispensed with the details of Gödel coding, as is becoming fairly common in books like the one you are reading now. The reader can, if she wishes, rewrite formulas so as to make it all formally correct, but I doubt that much is gained by doing so.

On the philosophical side, I cannot avoid addressing subtle interpretive questions in the history of philosophy and entering into thorny debates in contemporary epistemology. Therefore it is also expected of the reader that she has the equivalent of an undergraduate degree in philosophy under her belt.

How to read this book

If you have exactly the minimally required background knowledge described earlier, then I recommend you to read the chapters in the order in which they appear, from beginning to end. If you in addition are *au fait* with the state of the art on proof theoretic reflection principles, have a solid background knowledge about set theoretic reflection principles, and background knowledge of axiomatic truth theory, then you *can* skip Part II. If you are familiar with parts of this, then you *can* skip parts of Part II, or consult material in Part II whenever the need arises in reading Part III.

However, you would be mistaken to think that Part II *only* contains technical and philosophical background material. Especially in Chapters 5 and 6, conceptual distinctions are introduced and discussed that go beyond what can be found in the literature. Moreover, these conceptual distinctions are underpinned by philosophical arguments that may be considered far from uncontroversial. So even if you are a philosophically and technically sophisticated reader, I expect that you will find it necessary to pause and carefully consider some passages in Part II.

What I caution against above all, is to skip Part I altogether, even if you are a professional philosopher of mathematics who is not particularly interested in the history of philosophy or in contemporary epistemology. The reason for this is that philosophical choices are made in Part I: these choices essentially inform the view that is developed and argued for in Part III. I will argue for these philosophical choices as best as I can, but cannot claim to establish them beyond reasonable doubt. So the impatient professional philosopher who starts at the beginning of chapter III might find the discussion hard to follow, and will probably either give up on it altogether, or end up reading most of Part I anyway.

The reader will see that the material in this book is strongly interconnected. In an effort to make the reader appreciate the connections and to make this book structurally sound, I have included in this book lots of cross-references (mostly in footnotes), and a detailed index. Also, this book is intended to be a useful way of navigating the extensive relevant literature. I have tried to include as many useful references as possible, and I have tried to be explicit, careful and correct in my attribution of arguments, positions, and results to people. Sometimes this is not easy, for this book contains many ideas and results that I have picked up in the corridors of conferences, and in personal conversations. I hope that the reader will forgive me any mistakes—omissions or misattributions—that I have made.

Part I

CHAPTER 1

Mathematical Justification

In this Chapter, we consider the kinds of reasons that a mathematician has for believing in mathematical statements. Moreover, we investigate some of the epistemic concepts that are connected to these reasons, such as justification, mathematical justification, proof, formal proof, epistemic proof.

This area is the battleground of the disputes between the philosophers of mathematical practice on the one hand, and the 'traditional' philosophers of mathematics on the other hand. I will argue for a middle road in this debate.

From a conceptual point of view, this is a messy area. Often the concepts mentioned above are used in a casual way, without much attention to the sometimes subtle differences between them. It will emerge from our discussion that the distinctions between these concepts matter. Philosophical mistakes are made when they are not kept in mind, and they can guide us to a deeper understanding of the nature of epistemic warrant for mathematical statements.

1.1. Gettier problems

It has long been recognised that believing a true proposition is in general not sufficient for knowing that proposition. What is at a minimum needed in addition, is to be *warranted* to believe the proposition. Here being warranted to believe something roughly means having an epistemic right to believe it.

According to the traditional view, the notion of epistemic warrant can be further explained as *justification*. Moreover, until the early 1960s it was part of the received view that having justification in addition to true belief is *sufficient* for knowing: knowing is true justified belief.

Gettier's famous article made it clear that this view is untenable: true justified belief is in general not sufficient for knowledge [**Get63**]. Gettier essentially provided recipes for generating possible scenarios where a person has true justified belief in a proposition, but where we would be disinclined to say that she knows the proposition.

As an illustration, consider the following hypothetical situation. Sophie arrives at the belief that:

($\star$) It is warmer than 20°C outside or Germany is a federation.

and she arrives at the belief that ($\star$) in the following way. She starts by forming a belief that it is warmer than 20°C outside by checking an outside thermometer, which indicates a temperature of 21.5°C. From this, she infers, by propositional logical reasoning (the rule of "Disjunction Introduction") that ($\star$). Germany is a federation, so ($\star$) is true. Disjunction Introduction is a justification-preserving rule of inference, so Sophie's belief that ($\star$) is justified. However, unbeknownst to

Sophie, the outside thermometer is defective, and it is actually 18°C outside. In this situation, we would not want to say that Sophie *knows* that ($\star$).

This scenario indeed points towards a somewhat general recipe for constructing "Gettier-cases". Start with a false but justified belief; then logically weaken it to a true conclusion. Moreover, it also seems clear *why* we do not want to say, in such scenarios, that the subject knows this conclusion: a *false lemma* is essentially involved in her justification process.

Gettier's arguments generated a cottage industry of trying to identify the missing "Gettier condition" that must be added to true justified belief in order to obtain a satisfactory definition or logical analysis of the concept of knowledge. For instance, in the light of the foregoing, one might propose:

S knows that $\phi \Leftrightarrow$
S has a justified true belief that ϕ,
and no false lemmas essentially occur in S's justification.

However, most epistemologists do not believe that all Gettier cases can be taken care of in this way. The following classical example is taken to illustrate this. Suppose John is driving through a county that is filled with barn facades ("fake barns"). At a certain moment, he looks through his side window, and, based on what he sees, he says to himself: "there is a barn" (q). In fact, John happens to be right: he sees a real barn. But this is sheer luck. The overwhelming structures that look like barns in this county are fake barns. In this situation, many are again disinclined that John knows that q. But there is no false lemma involved, for John bases his belief directly on his perception.

The Gettier problem caused a shift from internalist (or justification-based) accounts of knowledge to *externalist* accounts of knowledge. Many epistemologists abandoned justification as a necessary condition for knowledge altogether, and replaced it with an external condition such as *reliability*. For instance, one might propose the following analysis of knowledge:

S knows that $\phi \Leftrightarrow$
S believes that ϕ,
p is true, and
the belief forming process that S has used on the occasion
reliably produces true beliefs.

Here I take justification more or less *by definition* to be an internal affair having to do with reasons for one's beliefs.[1]

But it is again far from clear that this strategy completely solves the Gettier problem. On the basis of the proposed analysis, one might say that Sophie does not know p because weakening a (false) belief by Disjunction Introduction does not *reliably* result in true conclusions. But the fake barn example remains a problem. Basing one's beliefs directly on one's own visual experience is, after all, a reliable belief forming mechanism.

An old argument against externalist theories of knowledge goes along the following lines. Consider Elisa, who is exceptionally good at predicting the future. Beliefs about future events simply pop into her head for no reason when she looks into her crystal ball, and these beliefs almost always turn out to be true. On one occasion, Elisa in this way comes to believe that it will snow in her home town

[1]More about this below: see Section 1.2 and Chapter 3.

in exactly 312 days (r). Sure enough, r is true. Elisa's belief forming mechanism (looking into her crystal ball) is extremely reliable. Yet we do not want to say that Elisa knows that r. So, the argument continues, in many cases there is more to knowledge than mere reliable belief. The problem, in Elisa's case, seems to be that she lacks good reasons for her beliefs about future happenings.

This worry applies especially to mathematical knowledge. Typically, mathematicians can and do adduce good reasons for their mathematical beliefs. And it is because they have these reasons that their beliefs are said to constitute mathematical knowledge. There may be situations where a mathematician is unable to articulate her reasons for a given mathematical belief, yet can still rightly be said to know the relevant proposition. The case of Ramanujan's extraordinary ability to produce deep true mathematical assertions without being able to support them by convincing reasons is often taken to illustrate this point. But there is a salient difference between having but being unable to articulate reasons for one's beliefs on the one hand (Ramanujan), and not having reasons for them at all on the other hand (Elisa). If Ramanujan's true mathematical claims simply popped into his head, unsupported by subconscious reasons, we would be unwilling to say that he knew these claims. In sum, strong forms of externalism cannot be the final word about mathematical knowledge.

There is, then, a deep connection between mathematical knowledge and having good reasons for (true) mathematical beliefs. Nonetheless, there is a conceptual gap between justified true belief and mathematical knowledge. Suppose that Edward Nelson's fear materialises, and first-order Peano Arithmetic (PA) turns out to be inconsistent. Suppose furthermore that little more than Nelson's sub-theory S_2^1 ("feasible arithmetic") of PA is true. And assume finally that, contrary to what Nelson claimed,[2] the mathematical community is justified in believing PA. Sarah, a competent mathematician, proves from axioms of PA that exceed S_2^1 an arithmetical theorem s. Unbeknown to Sarah and the rest of the mathematical community, s in fact also has a proof in S_2^1, albeit a completely different and much longer one. Then Sarah justifiedly believes the true statement s. But because her justification relies essentially on a false lemma (cfr. supra), she does not know that s. So we conclude that justified true belief-accounts of mathematical knowledge are also vulnerable to Gettier problems.

It is not my intention to *define* mathematical knowledge by isolating the "missing Gettier condition". I am sceptical that this can be done. Instead, we will be concerned with questions about mathematical knowledge that are quite independent of Gettier worries. More often than not, therefore, we will be concerned with scenarios in which Gettier problems are absent. In such scenarios, justification can account for the difference between true belief and knowledge. But, as we will presently see, justification is not the only epistemic property that can transform mere true belief into knowledge.

1.2. Justification

Epistemic warrant is a property of *beliefs*, where beliefs are contents of doxastic attitudes. Doxastic attitudes come in various strengths. A belief can be held more, or less strongly. Nonetheless, gradations of belief will not play a major role in this book. We will mostly work with a qualitative notion of belief, and then for

[2]See for instance [**Nel11**].

every proposition p there are only three possibilities: belief that p, belief that $\neg p$, and agnosticism concerning p. As contents of doxastic attitudes, beliefs have propositional structure: propositions are the kinds of entities that form the objects of belief. As propositions, the contents of belief are at least in part conceptual in nature. And if there are no "Russellian" propositions,[3] then beliefs are wholly conceptual in nature.

A belief is said to be epistemically warranted if, from an epistemological point of view, it is in good standing. Equivalently, a person's belief is epistemically warranted if in forming and maintaining the belief, she rationally believes it. From the foregoing Section, it follows that even true beliefs that are in epistemic good standing, can fall short of knowledge. Indeed, it is perfectly possible to be justified in believing that φ while φ is false.

Traditionally, a person's epistemic warrant for believing a proposition has been identified with her *justification* for believing that proposition. A justification for a belief is a complex of reasons that supports a proposition. Thus we say that a person is justified in her belief if her belief is supported by a complex of *reasons* for her belief. Here is a very simple example. I am justified in believing that there is a mole in my garden (t), for I have a good reason for believing t. New small heaps of soil appear almost daily in my lawn, and the best explanation for t is that they were pushed up by a mole.

For something to be a reason in a justification, it must itself be believed. Moreover, in order for a proposition to function as a justifying reason for a person, she must be epistemically warranted to believe it. For instance, justifying reasons can themselves be supported by further reasons, and thus themselves be justified.

Like belief, justification is a gradual notion: a person's reasons can to a high degree, or to a lesser degree support her belief in a proposition. Suppose I have some good reasons for my belief that my mother is in good health (u). Perhaps I later acquire more justification for u in the form of independent supporting reasons. On the other hand, perhaps I am later left with less justification for my belief, when reasons against u are brought to my attention. As with belief, we will in this book often work with a qualitative notion of justification. Roughly, this means that we speak as if there is some sort of vague threshold of support that a person's reasons have to meet before we say that her reasons fully justify her belief in a given proposition. In this way, belief and justification are unlike knowledge. Even though it may in certain cases not clearly be determined whether a person knows a statement, it would even in those cases be unnatural to say that she knows the statement *to a certain degree*.

It is generally believed that reasons can support beliefs in a variety of ways. In the example of my belief that there is a mole in the garden, the relation between my reason and my belief is *abductive* in nature: the heaps of soil are best explained by the presence of a mole in the garden. Reasons can also provide *logical* support for beliefs. For instance, my belief that ψ may be logically supported by my belief that φ and my belief that $\varphi \to \psi$. Reasons can provide *probabilistic* support for beliefs. My belief that I have thrown at least a 2 is probabilistically supported by

[3]Russell, in contrast to Frege, held the view that propositions can contain objects. For instance, on Russell's view, the proposition expressed by "Socrates is wise" contains the person Socrates. For a discussion of the distinction between Frege's and Russell's theory of propositions, see [**FN18**].

my belief that I have thrown a fair 20-sided die. Reasons can provide *inductive* support for beliefs. I have witnessed the sun rising countless times in the morning. These observational beliefs provide inductive support for my belief that the sun will rise in the future. Reasons can provide *semantic* support for beliefs. My reason for believing that bachelors are unmarried is that the word 'bachelor' is more or less defined as being shorthand for 'unmarried male'.

In each of these cases (induction, abduction, deduction,...) the kind of support that the reason gives is a matter of transition from beliefs to other beliefs. In one justification, of course transitions of different kinds may occur.

Beliefs can also throw doubt on other beliefs, and they can again do so in various degrees. Assessed counter-reasons, for instance, can also be part of a justification. Suppose, for instance, that I believe that Louise is a pilot (v). Then it can be part of my justification of v that even though Peter told me that Louise is an air stewardess, I have reasons to believe that he often gets people's professions slightly wrong. How *strong* the supporting or undermining relation between reasons and beliefs is, is a difficult, multifarious, and important epistemological question. It forms the subject matter of confirmation theory; I will not have much to say about it in the rest of this book.

Every justification is a finite structure. So every chain and every anti-chain of reasons in a justification must be finite. The reason for-relation is transitive. The latter is not intended to be a substantial claim; it only means that we do not focus on the *immediate* reason for-relation, but allow for mediate reasons as well. Often the reason for-relation is also taken to be anti-reflexive. From this it would follow that it is also anti-symmetrical and therefore a strict partial ordering. As a consequence of this, a certain kind of "holistic" picture of justification would be precluded. I do not want to take a stance on this matter, and therefore keep an open mind about the anti-reflexiveness of the reason for-relation. However, since, as we will see, it is part of the view developed in this book that there are epistemically warranted beliefs that are not supported by reasons, the reason for-relation cannot be reflexive.

A justification is often modelled as a finite transitive directed graph of reasons. The foregoing considerations show—and this is old news—that modelling a justification this way is in general rather simple-minded. Perhaps a more realistic way of modelling justifications is as neural networks, with weights (real numbers between 0 and 1, say) attached to nodes as measures of degree of belief, and weights (real numbers between -1 and 1, say) attached to the arrows as measures of (dis-)confirmation. Nonetheless, in a variety of circumstances, modelling justifications as finite transitive directed graphs of reasons can be useful, and we will sometimes do so in what follows.

The case of Ramanujan (cfr. supra) reminds us that, for a variety of reasons, people are not always capable of fully articulating the reasons for their beliefs. So room must be left for unconscious beliefs and for unconscious reasons. Nevertheless, as Lewis Caroll and Quine have taught us,[4] there must be an end to postulating sub-conscious reasons. For instance, it seems wrong to say that a sub-conscious belief that $[\varphi \wedge (\varphi \to \psi)] \to \psi$ is responsible for the support that my reasons φ and $\varphi \to \psi$ lend to my belief that ψ. At some level, we simply "blindly" follow rules when proceeding from reasons to further beliefs.

[4]See [**Qui36**].

Reasons that a person has for believing a proposition can be articulated. Suppose a person is asked: 'Why do you believe that there are tigers on Sumatra?' Then she will typically reply to this question in the form of an *argument* [**Hel92**, Section I]. Her argument usually does not contain *all* her reasons that bear on this question: there are many such, and time is short. Her argument will typically contain those of her reasons that are particularly salient, and will contain information about how they are connected. When pressed, she can usually offer more relevant considerations, or explain in more detail how her reasons bear on her belief. Some of her reasons may be sub-conscious, and therefore will not be part of her argument. Moreover, the strength of the various beliefs that function as reasons regarding the why-question above are relevant. They will at best only partially be indicated in her argument, which is usually only to a very limited extent probabilistic in nature. In sum, the argument that she gives constitutes only an imperfect approximation of her justification.

I have excluded the possibility of justifications consisting of *infinite* complexes of reasons. In particular, a justification cannot contain a *non-wellfounded chain* of reasons. Also, if a given justification contains no unjustified reasons, then this means that it must contain at least one *cycle* of the reason for-relation, which is of course also a form of non-wellfoundedness. (Here I count a self-supporting reason as a very small such cycle.)

Foundationalism in epistemology is the doctrine that claims that the reason for-relation is well-founded. If Foundationalism is correct, then every justification contains reasons that are not themselves supported by reasons. Let us call such reasons *basic reasons*, where whether a reason is basic then depends on the particular justification that we are considering. Recall that we have not committed ourselves to anti-reflexiveness of the reason for-relation. So we have not committed ourselves to (or against) epistemic Foundationalism. But suppose for a moment that there are basic reasons, which is of course a far weaker claim than Foundationalism. We know that a reason in a justification is a *warranted* belief. So the question arises: *How are basic beliefs epistemically warranted?*

One way to delay this question is the following. A basic belief b in one justification j_1 may be supported by reasons $r_1, \ldots, r_n$ in another justification j_2. Perhaps one person or community can have both justifications (j_1 and j_2) in an "unconnected" manner at one point in time. Then perhaps j_1 and j_2 can somehow be "strung together" to a justification in an extended sense.[5] If that is the case, then at a given time t_0 we can consider the *total* "extended justification" for a given proposition p. Indeed, we have an epistemic obligation to combine all our reasons that bear on p, when forming a doxastic attitude towards that proposition p. Otherwise we might have a complex of reasons that support p, and thus be said to be justified in believing that p, whereas we also possess, in an unconnected way and at the same time, reasons that speak against p. In such a situation, we would not be considered fully rational.

Nevertheless, at every point in time, the number of justifications that a person or community possesses, remains finite. So even the total justification for p at t_0 may contain basic beliefs. For any such p (and t_0), we may ask the *harder* question:

[5]If j_1 and j_2 can be pictured as finite transitive directed graphs, then it is straightforward how this is done. In more realistic cases, however, it is not always easy to see how this goes: recall that one justification can *undermine* another justification.

How are the basic beliefs in this total justification for p epistemically warranted at t_0?

This is of one of the cardinal problems of epistemology. However, I won't be able to side-step it. It is my hope that recent developments in epistemology that will be discussed in the following chapters contain elements that will prove relevant to the eventual solution of this deep epistemological problem.

Beside being in harmony with the way in which they support other beliefs (inductive support, deductive support,...), reasons can be classified according to the kind of warrant that underpins them. A belief is *a priori* if its warrant does not rest on sense experience; if a belief is not a priori, then it is *a posteriori*. This means that even if sense experience is needed to acquire some of the concepts in the belief, it may still be a priori. For instance, perhaps in order to learn the concepts 'red' and 'green', one needs sense experience. But one can still know a priori that an object cannot be red and green all over at any given point in time. Similarly, an application of a rule of inference preserves aprioricity if its epistemic warrant does not depend on sense experience; otherwise it is again a posteriori. A *justification* is a priori if its reasons and inferences are a priori. In this sense, to use Paseau's terms, a posteriori reasons in a justification are *dominant*, and a priori reasons in a justification are *recessive* [**Pas15**, p. 790]. Further, someone can have both an a priori and an a posteriori justification for a belief that she holds. To conclude, a person *knows* a proposition a priori if she has an a priori epistemic warrant for it.

1.3. Justification of mathematical beliefs

Suppose that some mathematician—or even a community of mathematicians—holds some mathematical belief. In order for her belief to be in good epistemic standing, it is often said, she must have justification for her belief, i.e., her belief must be based on good reasons. Wherein do these good reasons consist?

Quine and Putnam would point to the fact that mathematical beliefs play a central role in our explanations of natural phenomena [**Qui54**, p. 251]:[6]

> A self-contained theory which we can check with experience includes, in point of fact, not only its various theoretical hypotheses of so-called natural sciences but also such portions of logic and mathematics as it makes use of.

Mathematical reasons play a role in our explanation of why a certain bridge will not collapse; mathematical reasons play a role in our explanation of why an airplane does not fall from the sky. In other words, mathematical reasons are part and parcel of our *best explanations* of the phenomena that we observe, and of the empirical success of the empirical predictions that we make about observable phenomena.

Paseau has described how mathematical propositions can even be justified in much more direct, inductive ways.[7] He gives the example of the theorem that a triangle's perpendicular bisectors always meet at a point [**Pas15**, Section 7]. You might draw a triangle on a piece of paper, construct its perpendicular bisections, and notice that they meet at a point. You then do the same for many different drawn triangles: always the same phenomenon occurs. On the basis of this, you inductively draw the conclusion that this property holds for all triangles drawn on

[6] See also [**Put71**].

[7] Actually, Paseau talks about knowledge of mathematical propositions in this context. But what he writes applies equally to justification.

paper. Furthermore, you may believe that the triangles that can be drawn on a piece of paper are 'isomorphic', in relevant respects, to the triangles that exist in Euclidean space. Thus you draw the conclusion that the property also holds for all triangles in Euclidean space. The upshot is that you have good, direct, empirical reasons for a belief in a mathematical proposition.

Another example of this is the *Kepler Conjecture*, which says that no arrangement of equally sized spheres filling Euclidean space has higher average density than cubic close packing and hexagonal close packing (which are easily seen to be equivalent). Experimenting with stacking oranges will quickly convince you of the truth of this conjecture. But proving the Kepler Conjecture turned out to be very difficult: this was achieved only in the 1990s.[8]

The inference to the best explanation-account of our justification of mathematical beliefs forms part of Quine's *epistemological naturalism* [**Qui69**]. According to this view, our best reasons are the ones that natural science provides, even if these reasons are always defeasible. Further support for this view might be sought in the history of mathematics. For instance, the development of the calculus, and later of mathematical analysis, took place in close interaction with developments in theoretical physics (the discovery and development of Newtonian mechanics).

On Quine's account, the mathematician's support for her mathematical beliefs is ultimately not distinctively mathematical in kind, but at least in part empirical. Thus the mathematician has non-mathematical reasons for her mathematical beliefs. Maddy has objected to Quine's account that it does not correctly describe how *mathematicians* typically justify their mathematical beliefs. Indeed, a typical mathematician is simply not interested in whether, and, if so, how, mathematical propositions form an essential part of empirical theories. Instead, she takes herself to have *mathematical* justifications for her mathematical beliefs.

Continuing in this vein, one might argue that a *mathematical* justification of a mathematical belief usually takes the form of a proof from axioms. Quine might try to give proofs their due by applying his account only to the mathematical axioms. Then the justification that a mathematician has for her mathematical belief that φ consists of her proof of φ from axioms *plus* her justification of the axioms by an abductive argument along the lines sketched above.

But this reply does not get to the heart of Maddy's objection. Today, a significant number of philosophers appeal to inference to the best explanation-considerations as main reasons for their mathematical beliefs. Perhaps this even also holds true, in a less explicitly articulated manner, for non-mathematicians who are not philosophers. But whatever reasons a mathematician may have for believing mathematical axioms, they do not concern the role they play in successful empirical theories.

So we must take a closer look at what counts as *mathematical* reasons for mathematical beliefs. Mathematical justification for a proposition φ is often identified with a *proof of* φ *from axioms*. What it means to *have* a proof of a mathematical statement is a contentious issue. The Four Colour Theorem say that any map on a bounded plane can be coloured with four colours in such a way that adjacent[9] regions have different colours. This theorem was established in 1976 by Appel and

[8] See [**Hal05**].

[9] Adjacent means that two regions share a common boundary curve segment, not merely a corner where three or more regions meet.

Haken by extensive use of computer assistance.[10] Their proof relied on a computer checking 1834 cases ("reducible configurations"), which took the computer 1000 hours. At the time at least, they could only be said to "have" a proof of the Four Colour Theorem if the computer verified cases were seen as part of their proof.[11] But let us be relaxed about this, and count such computer-assisted proofs as genuine mathematical proofs. Even then it is clear that mathematical justification of a statement φ can also take other forms than mathematical proofs of φ.

A *probabilistic proof* of a given proposition φ is a proof that has information about the probability of φ, rather than φ itself, as its conclusion. Instead, such a proof establishes that the probability of φ is smaller than $1 - \varepsilon$, where ε can be made very very small.[12]

Another example of a mathematical justification of a mathematical statement which falls short of being a mathematical proof might be the following. The teacher of my first undergraduate mathematical logic course, Jan Denef, once took time out in one of his lectures to talk to us about the state of affairs in number theory at the time—it was 1988, I believe. He explained to us how it had recently become clear that a collection of conjectures are tightly connected to each other by a net of total and partial implication relations. For instance, he explained, Fermat's Last Theorem seemed closely connected to the theory of elliptic curves, in particular with the Taniyama-Shimura-Weil Conjecture. So, if one of these conjectures could be proved (and they did not all appear completely out of reach), then a bunch of others would also be proved or partially proved. Moreover, these conjectures seemed to be converging on an intelligible 'picture'. They seemed to articulate a deep understanding of the situation: if they would turn out to be false, then we would have no idea what was going on. Of course Fermat's Last Theorem was proved not much later,[13] and the full Taniyama-Shimura-Weil Conjecture was proved in the wake of that. I submit that around 1990, leading number theorists were *mathematically justified* in believing Fermat's Last Theorem. It would have been irrational, for example, if one of these experts, knowing what she did at the time, would decide to spend all her time and efforts on trying to refute Fermat's Last Theorem. Perhaps the current situation concerning the Goldbach Conjecture is in relevant respects similar.[14] In sum, *abductive* arguments can also give *mathematical* justification.

There is a third way in which a mathematician might be justified in a mathematical belief in the absence of a correct proof. Suppose a mathematician has a very complicated argument for a mathematical proposition φ. Suppose hers is a high level argument, resting on many lemmas, some of which she has argued for herself, while others are taken from the literature. Moreover, suppose that she has not verified all these lemmas in detail. To conclude, suppose that there are even errors in some of the proofs of the lemmas, but that these errors can (and in due time will) be corrected. Our mathematician might be painfully aware of the possibility that the details of the proofs of some of the lemmas might turn out to be

[10] See [**AH78**].

[11] I am abstracting here from some errors in their 1976 proof, which were by 1989 all corrected.

[12] An epistemological assessment of probabilistic proofs is given in [**Eas09**].

[13] See [**Wil95**].

[14] For a discussion of the epistemological status of Goldbach's Conjecture, see [**Pas15**, Section 3]. For an overview of Goldbach conjectures until around 2015, see [**Vau16**].

thornier than she expected, and that some of her arguments might contain errors. She might reflect on her predicament and say to herself: this sort of situation has occurred many times before in the recent history of mathematics, and in *most* such situations, the argument turned out to be basically correct or at least in the right direction. In this case, I would again be inclined to say that our mathematician is mathematically justified in her belief of φ. Moreover, the *inductive* considerations that she brings to bear on the situation somewhat strengthen her grounds for believing φ.

To conclude, you can have a proof that there is a proof of a mathematical statement φ while a proof of φ is forever beyond your reach. For any sufficiently strong formal theory T, there are mathematical theorems φ of T that only have astronomically long proofs even though there are short proofs in T of the provability of φ in T. Let $\mathrm{Bew}_T x$ be defined as $\exists y \mathrm{Proof}_T(y, x)$, where $\mathrm{Proof}_T(y, x)$ is an arithmetical predicate that codes the relation "y is a proof of x' in T' in a standard way. Then we have:

THEOREM 1.1. *For any sound $T \supseteq PA$, there is a sentence ϕ_T, which is provable in T, and which is such that the shortest T-proof of* $\mathrm{Bew}_T(\phi_T)$ *is much shorter than the shortest T-proof of ϕ_T.*

PROOF. [**Par71**, Theorem 1.3]. □

Propositions that only have very long proofs in one theory (PA, say), may have much shorter proofs in a stronger theory (ZFC, say). But our theorem 1.1 also holds for the strongest theory that we currently believe. Suppose, for the sake of argument, that this strongest theory is ZFC. Then there are proofs of ϕ_{ZFC} in ZFC, but we can never obtain one by working in ZFC. At the same time, we have a short proof (in ZFC) of $\mathrm{Bew}_T(\phi_{ZFC})$. Does not this proof provide good mathematical reasons to believe ϕ_{ZFC}?

Despite all this, the fact remains that mathematical proof remains somehow the mainstay of mathematical justification. Let us therefore now scrutinise the notion of mathematical proof, and investigate how it is connected to mathematical justification.

1.4. Mathematical justification

We have seen that someone may have inductive *non-mathematical* reasons for a mathematical belief (p. 19). But a person may also have inductive *mathematical* reasons for a mathematical belief. As an example of this, consider the Goldbach Conjecture again. The *Goldbach number* $G(n)$ of a natural number n is defined as the number of different ways in which n can be written as the sum of two primes. This means that the Goldbach Conjecture holds if and only if for all even natural numbers ≥ 2, $G(n) \geq 1$. Computer evidence indicates that the function G tends to increase as n increases, such that, for instance, for even numbers $n \approx 10^5$, $G(n) \geq 500$. Despite the fact that it is generated by computers, this is widely taken to be inductive *mathematical* evidence for the Goldbach Conjecture [**Pas15**, p. 779].

Inductive considerations *by themselves* are in many cases insufficient to warrant belief in mathematical propositions. Let us consider a few examples that illustrate this. A perfect number is a natural number that is the sum of its proper divisors. (The number 6 is an example.) One open question is whether there are odd perfect

numbers. Inductive evidence for a negative answer is the fact that the least odd perfect number would have to be larger than 10^{1500}. But this is only one reason, which by itself is not taken as sufficient, for mathematician's beliefs that there probably aren't any odd perfect numbers. Another example of an open question is the following: are there natural numbers of the form $2^p - 1$, with p prime, that are not square-free, i.e., that are such that their prime factorisation contains some factor twice? On this question, Guy writes [**Guy04**, p. 14]: "This seems to be another unanswerable question. It is safe to conjecture that the answer is "No!". This *could* be settled by a computer, if you are lucky." In other words, even though we know for sure, on inductive grounds, that non-square-free numbers of the form $2^p - 1$ (with p prime) would have to be rather large, mathematicians believe that this does not amount, by itself, to good reasons for believing that there are no such numbers.

A cursory look at notebooks of mathematicians of the past shows that inductive evidence has always played an important role in mathematical research. Today, a reputed mathematical journal, *Experimental Mathematics* (founded in 1992), is devoted to this activity. The founding editors describe the "philosophy" behind their journal as follows:[15]

> *Experimental Mathematics* was founded in the belief that theory and experiment feed on each other, and that the mathematical community stands to benefit from a more complete exposure to the experimental process. The early sharing of insights increases the possibility that they will lead to theorems: An interesting conjecture is often formulated by a researcher who lacks the technique to formalise the proof, while those who have the techniques at their fingertips have been looking elsewhere. Even when the person who had the initial insight goes on to find a proof, a discussion of the heuristic process can be of help, or at least of interest, to other researchers. There is value not only in the discovery itself, but also in the road that leads to it.

This suggests that inductive mathematical evidence is taken by the mathematical community to be primarily a matter of *heuristics*: it belongs to the context of *discovery* rather than to the context of justification of mathematical belief.

It is not clear, however, that inductive mathematical evidence cannot do any justificatory work. Especially if there are *several* bodies of inductive mathematical evidence for one mathematical proposition p, where these bodies of evidence appear to be totally independent of each other, the inductive evidence can rightly be taken to provide good reasons for believing p. This can be so despite the contrast between the finiteness of inductive evidence and the infinite number of instances that are covered by the mathematical statement on which it bears.

Inductive evidence is not the only kind of mathematical evidence. *Consilience* is another. When ever stronger (and often ever more complicated) results seem to converge on a conjecture, this is seen as warrant for believing the conjecture to be true. As an example, consider the Goldbach conjecture once again. Progressively, mathematicians seem to be "approaching" this conjecture. The following two results give just a small indication of this. In 1966, Chen Jingrun proved a theorem that is closer to the Goldbach conjecture than the ones that came before [**Che73**]:

[15] https://www.emis.de/EM/expmath/philosophy.html

> Every sufficiently large even number can be written as the sum of a prime and a number with ≤ 2 prime factors.

In 2013, Harald Helfgott proved the *weak Goldbach conjecture*, which is of course also immediately implied by the full or "strong" Goldbach conjecture [**Hel15**]:

> Every odd number greater than 5 can be expressed as the sum of three primes.

Of course this does not mean that mathematicians now believe that the full Goldbach conjecture will be established soon.

1.5. Mathematical proof

We have seen that there can be good non-mathematical reasons, and mathematical reasons that do not constitute a proof, for holding a mathematical belief. So such reasons can provide justification for mathematical beliefs. There is in mathematics a *gold standard* for the justification of mathematical belief, and this standard is mathematical proof. Indeed, mathematical proof is widely seen as an end in itself of mathematical activity [**Pas15**, p. 795]. Let us therefore turn our attention to the concept of mathematical proof.

1.5.1. From mathematical justification to mathematical proof. It is tempting to think that mathematical proof is a species of mathematical justification.

Like a mathematical justification of a mathematical proposition φ, a mathematical proof of φ can be viewed as an *argument* for φ. The reasons in a mathematical proof are *mathematical* reasons, and propositions that express mathematical reasons only contain mathematical concepts. Beside mathematical proofs, there are also non-mathematical proofs. Fitch's argument, for instance, is a *proof* that not every truth is knowable.[16] Moreover, truth is a *philosophical* rather than a mathematical concept. Yet we can prove the consistency of Peano Arithmetic using the concept of truth as follows [**Myh60**, p. 463]:

> The axioms of elementary arithmetic are true, and the rules of inference are truth-preserving. Therefore every theorem of elementary arithmetic is true. Therefore '0=1' is not a theorem of arithmetic.

Indeed, this is a perfectly acceptable proof. But since it crucially relies on basic logical principles of *truth*, it is a non-mathematical proof. We will see later that the distinction between proof and mathematical proof is a significant one.

Unlike mathematical justification in general, mathematical proofs are in the final analysis purely *deductive* arguments [**dT21a**, p. 11]. With logical consequence, we mean *first-order* logical consequence. (We symbolise this consequence relation as $\models$.) Gödel's *completeness theorem* shows that there is a positive test for first-order consequence. So if a mathematical proof contains a claim of the form $\phi_1, \ldots, \phi_k \models \psi$, then if it is correct, a mathematician can find a fully spelled-out *logical derivation* of ψ from $\phi_1, \ldots, \phi_k$. Indeed, a computer can be programmed to find such a logical derivation for *any* given $\phi_1, \ldots, \phi_k, \psi$ such that $\phi_1, \ldots, \phi_k \models \psi$.

There are two objections against the claim that mathematical arguments are purely deductive, neither of which I find convincing.

[16] See [**Wil00**, Chapter 12].

The first objection points to the fact that *actions* are taken to play a major role in mathematical proof, and actions are not logical deductions.[17] For instance, one might read in a proof: "Now construct a perpendicular bisector of line segment AB." But such construction-talk is an anthropomorphising way of speaking that should not be taken literally. In the example under consideration, the construction-instruction can be seen as a vivid way of adducing the reason that *there is* a perpendicular bisector of the line segment AB, and this reason is warranted by fundamental principles of geometry.

The second objection points to the fact that when one considers a transition of the form "φ; therefore ψ" in a mathematical proof, it turns out on closer inspection almost never to be the case that φ logically implies ψ. But this is because, like justifications in general, the argument that a person actually gives for a belief that ψ almost never contains all the reasons that a person has for ψ. In other words, in such a situation, there are typically reasons $\gamma_1, \ldots, \gamma_k$ that are also needed to warrant the transition from φ to ψ. But the writer of the proof may, in the context of the proof, assume that the reader can work this out herself once the key reason φ is given. So this second objection is also found wanting.

The process of interpolating "auxiliary" reasons in mathematical proofs is often highly non-trivial.[18] An article is received by a reputed mathematical journal. The referees closely inspect the proofs in the article. In this process, they work out for their own understanding certain parts in more detail. The manuscript is eventually accepted for publication. It appears as an article in the journal, with some more detail in proofs here and there than in the original submitted manuscript. Professional mathematicians from the relevant sub-discipline are expected to be able to fill in the gaps in the proofs, but most graduate students or even PhD students in the sub-field are not able to do this. A few years later, a graduate textbook appears that contains a proof of the theorem. Typically, this proof contains significantly more detail (more detailed reasons) than the proof in the journal, and therefore the textbook proof will typically be significantly longer. For undergraduate mathematics students, even the graduate textbook proof will be impenetrable. It may be that a dramatic simplification of the original proof is found, and that this simplified proof is accessible even to undergraduate students. But this is not always the case.

Rav has objected to this kind of story that the process of interpolation of reasons in a mathematical proof until a logically correct and non-elliptical first-order derivation is reached, can be an *infinite* process [**Rav99**, p.14–15]. The hypothesis that, on the contrary, such a process of interpolation is always finite, is known as *Hilbert's Thesis* [**Rav99**, p. 11].

As a concrete example, consider a recent discussion that centres around the proof of *Alexander's lemma* [**Ale23**]. This fundamental theorem in knot theory states that every tame knot is equivalent to one with a diagram that winds along an axis. De Toffoli and Giardino claim that parts of this proof are very difficult to formalise, and impossible to formalise without completely altering the structure of the original proof [**dTG16**]. But these claims appear to be unfounded. Alexander's original proof is rigorous and does not contain appeals to intuition.[19] It is true that

[17] A spirited defence of this view can be found in [**Jon98**].

[18] This process is analysed well in [**TBng**], on which the following description is based.

[19] This is shown in [**TB20**, Chapter II]. De Toffoli replies to Tatton-Brown's critique in [**DT21b**].

there are convincing intuitive arguments for mathematical propositions that are not fully rigorous. Jones' intuitive argument for Alexander's lemma,[20] for instance, does not meet standards of rigour. But it *needs* to be made more rigorous to be more than a plausibility argument; moreover, this can be done.[21] In sum, as far as I am aware, Hilbert's Thesis seems to be holding its ground so far. [22] We have no examples of perfectly acceptable mathematical proofs for which we have reasons to believe that the process of transforming them into first-order proofs from basic reasons is an interminable task.

In the end, this leads to the hypothesis that we can *model* a mathematical proof of a statement φ as a logical derivation of φ from mathematical principles that can somehow be considered to be basic. Here what is basic is a relative matter. Certainly the axioms of the system in which the logical derivation takes place, counts as basic. But in a mathematical proof, theorems that are proved elsewhere are virtually always appealed to (and when one looks up their proofs in a textbook, one sees that these appeal to other theorems, and so on). The formalisations of these theorems that are appealed to will appear as basic reasons in the formalisation.

Modelling mathematical proofs in this way certainly has its uses. Given Gödel's completeness theorem, when a first-order derivation of φ from premises is elliptical but logically sound, we can mechanically find a correct logical derivation of φ. This is the idea behind *automated proof verification*. But formal derivations in first-order logic should not be confused with what mathematical proofs really are.

In sum, so far mathematical proof indeed looks like a special kind of mathematical justification. However, there is one essential property of mathematical proofs that is not shared by mathematical justification in general. We saw earlier (p. 16) that justification does not entail truth. But proof, and therefore in particular mathematical proof, does entail truth. In epistemological terms: proof is *factive*. If someone professes to know a mathematical proposition φ because she has proved it, and φ later turns out to be false, then we say that she did not have a proof of φ to begin with: she merely *thought* she did. Williamson puts it as follows [**Wil00**, p. 265]:[23]

> A way of having a warrant to assert p is factive just in case a necessary condition of having warrant to assert p in that way is that p is true. Grasping a proof of a mathematical proposition is a factive way of having warrant to assert it: a necessary condition of grasping a proof of p is that p is true.

Facticity is one of the properties that distinguish proof from proof in a formal system, which is not a factive notion. So 'wrong proofs' are not (informal) proofs at all. A wrong proof may contain logical errors. But it may also be impeccable from a logical point of view but depend on some false basic mathematical assumption φ. In the latter case, the argument will still be a correct proof in some formal system, namely a system that has φ as one of its axioms. But it will not be a proof in the mathematician's sense of the word.

---

[20] See [**Jon98**].

[21] This is shown in [**TBng**].

[22] For an extended argumentation for this thesis, see [**TBng**].

[23] See also e.g., [**Art08**, p. 492], [**Hor94**, p. 286].

Factivity is a property that proof has in common with knowledge. It may be because of this conceptual link with truth that mathematical proof is almost universally taken to be sufficient for mathematical knowledge.[24] When a person has a mathematical proof that ψ, and believes ψ on the basis of this proof, then she knows ψ. It is also because of this conceptual link with truth that mathematical proof is impervious to Gettier challenges.

The claim that mathematical proof entails truth is not universally accepted. De Toffoli, for instance, defines mathematical proof thus [**dT21a**, Section II.3]:

> A *mathematical proof* is a correct deductive argument for a mathematical conclusion from accepted premises that is shareable.

According to this definition, mathematical proof does not have a conceptual connection with truth. A correct argument from accepted premises is not the same as a *sound* argument: a deduction from premises may be correct even though some of the premises are false.[25] Therefore (and also because her definition classifies non-mathematical proofs of mathematical statements as mathematical proofs), her definition does not get the extension of the concept of mathematical proof right.

Perhaps the reason why some refrain from asserting that there is a conceptual connection between proof and truth is that it is felt that claiming such a connection commits one to a platonist view about mathematics. But if the content of the concept of truth is simply given by some natural collection of Tarski-biconditionals,[26] then such worries are unfounded. For then, to say for example "if it has been proved that there are infinitely many prime numbers, then it is true that there are infinitely prime numbers", is little more than to say "if it has been proved that there are infinitely many prime numbers, then there are infinitely many prime numbers", where the acceptance of the latter sentence need not commit one to mathematical platonism. In other words, the concept of truth does not have to be laden with a 'correspondence theory'. But more about this later.

1.5.2. Informal mathematical proof. From the foregoing, it is clear that proofs in formal systems can be contrasted with the proofs that are constructed by mathematicians. The latter are often called *informal proofs*.[27] The distinction formal / informal applies as much to proofs in general as it does to mathematical proofs. So we will use the term 'informal mathematical proof' as standing for a different concept than the term 'informal proof', and I will take 'informal mathematical proof' to be synonymous with 'mathematical proof'. In the literature, no clear distinction is often made between the concepts of informal proof and informal mathematical proof. But we will see that this distinction matters.

There is a sense in which informal mathematical proof stands to formal proof as truth stands to truth in a model. Like truth, informal mathematical provability

---

[24]We have seen earlier (Section 1.3) that mathematical proof is probably not *necessary* for mathematical knowledge.

[25]When commenting on the need for the shareability requirement in her definition, De Toffoli equates correct deduction with the (to logicians) more familiar concept of valid deduction [**dT21a**, Section II.3]: "If proofs are equated with valid deductive arguments in a formal system, then not all proofs would be shareable since proofs could be so long that they could never be grasped..."

[26]For one specific elaboration of such a deflationist account of truth, see [**Hor98**].

[27]See a.o. [**Myh60**], [**Hor05a**], [**Lei09**], [**AM10**].

is taken to be an absolute notion.[28] Indeed, Myhill used the term *absolute* proof instead of informal mathematical proof [**Myh60**]. Like truth in a model, formal proof is a relative notion: it is relative to a formal language, and a proof system (i.e., a system of formal axioms and rules of inference) for that language.

I expect readers to be familiar with the definition of formal proof: we need not go into it here. Defining the notion of informal mathematical proof is anything but straightforward. We have already discussed one definition that is not satisfactory (p. 27), and we will encounter a few more that are not correct. Informal mathematical proof, like knowledge, appears to be a very fundamental epistemological concept. Therefore I am sceptical that a satisfactory definition of mathematical proof can be given. I believe that it is more fruitful to look for and investigate fundamental properties of informal proof that somehow flow from the nature of the concept. Many questions of this sort are very difficult, and little is known. In order to learn more about this notion, let us explore the comparison with formal proof further.

The question whether, for a given formal system S, a sequence of formulas constitutes a formal proof in S, is decidable. Is the notion of informal mathematical proof similarly decidable? Of course decidable here has to be taken not in the formal sense of the word, i.e., as *recursive*. For informal mathematical proof to be decidable, means that for any argument, we can eventually come to know whether or not it is an informal mathematical proof. We can even go further, and ask whether there is, for any argument, a (non-mathematical!) proof that the argument constitutes a proof, or a proof that the argument does not constitute a proof.

This question is related to what is known in epistemology as positive and negative introspection for informal proofs. In contemporary epistemology, these are rightly regarded with suspicion [**Wil00**, Chapter 4], and negative introspection even more so than positive introspection. A mathematician can have a mathematical proof of a mathematical statement without knowing that she does (she may not even believe that she has one), may think she has a mathematical proof of it while she doesn't. But these considerations do not quite decide our question. The problem lies with the '*can* come to know' part of our question. This modal component induces an idealisation from the actual finite, bounded, mortal mathematician (and mathematical community). This makes our question somewhat unclear. If we idealise as far as God, then the answer to our question is trivially and uninterestingly yes, because then knowledge collapses into truth. So we do not want to go that far in our idealisation. But it is not clear how far it is reasonable to take our idealisation to extend [**Kri80**, p. 34–35]. In sum, our question is completely open, and it is not even clear whether it is sufficiently precise to admit for a clear answer.

Informal mathematical proofs consist of *interpreted* statements. Formal proofs, in contrast, are often considered to consist of meaningless formulas [**Rav99**, p. 12]:

> Hilbert's Thesis is just a one-way bridge: from a formalised version of a given proof, there is no way to restore the original proof with all its semantic elements, contextual relations and technical meanings. Once we have crossed the *Hilbert Bridge* into the land of meaningless symbolic manipulations, we find ourselves on the shuffleboard of symbolic manipulations and [...] these symbols do not encode meanings... [I]it is the very purpose of

[28] But more about this later: see Section 1.6.

> formalization to squeeze out the sap of meanings in order not to blur focusing only on the logico-structural properties of proofs.

But this is an exaggeration. Any student of first-order logic will tell you that we can, and often do, keep track of the meanings of symbols when we formalise arguments (R means 'red', Gxy means 'x is greater than y',...). Nothing prevents us from focusing *both* on the logical structure and the meaning of non-logical concepts in an argument.

It is often said that whereas formal systems have axioms, axioms are not presupposed in informal mathematical proofs. As an example, let us look at number theory, which is listed by Rav as a non-axiomatised theory.[29] Number theory freely draws upon other areas of mathematics, such as topology, for instance. But topology itself is based on agreed first principles (axioms): this, in the final analysis, is what allows mathematicians to decide whether a given topological argument in number theory is correct. Moreover, it is alleged that the logic that is used by mathematicians is not first-order logic [**Rav99**, p. 16]:

> ...the standard theorems of group theory [...] are not even expressible in first-order predicate calculus. One just has to think about such fundamental concepts as normal subgroup, torsion group, finite group, composition series, or such famous theorems such as the Sylow theorems about p-groups, the Jordan-Hölder theorem and the like, to realise that the implicit underlying logic of mainstream group theory is second-order logic.

But this just means that group theory contains lots of theorems involving *sets* of groups. Set theory is also axiomatised (in first-order logic), so group theorists can use accepted methods of set theory and still agree whether a proof is correct. At the level of proofs, a principled distinction between first- and second-order does not even make much sense, since quantification over 'second-order' entities can perfectly be treated in a two-sorted *first-order* calculus.[30]

1.5.3. Informal provability. It is very hard to find a satisfactory definition of the concept of informal mathematical proof. We can, however, try to find basic principles that govern it. We may even express the outcome of this investigation as a formal system.

Mathematical proof, like knowledge, is from a logical point of view a *weak* notion. Factivity is about the only 'pure' logical principle that clearly governs the notion of informal proof. Therefore the 'logic of informal mathematical proof' is likely to be uninteresting. For this reason, philosophers have instead focused their attention on the logic of informal prova*bility*. Despite the aforementioned unclarity of this notion (p. 28), the logical laws that (clearly) govern it may turn out to be somewhat interesting.

---

[29] "*Number Theory.* Once more, a non-axiomatized theory! Notice that I am talking about number theory as the term understood by the mathematical community—not to be confused with first-order fragments of second-order Peano Arithmetic, which is a branch of mathematical logic" [**Rav99**, p. 16]. See also [**Lei09**, Section 2].

[30] For an elaboration of this point, see [**Sha99**]. See also [**Azz04**, Section 3].

This idea goes back to early work of Gödel [**Göd33a**]. To a classical propositional background logic, he adds an intensional sentential operator, which is intended to express provability. He then goes on to claim that this operator is governed by the laws of S4 propositional modal logic. (This idea was later extended to quantified and higher-order settings: see [**Sha85**].)

Already at this early stage, somewhat subtle considerations arise. In his discussion of Gödel's proposal, in a setting where the background theory is Peano Arithmetic, Myhill rejects the 4 axiom ($\Box A \to \Box\Box A$) [**Myh60**, p. 469]:

> We have decided not to include Gödel's [Axiom 4]. Our axioms are intended to be added to some underlying formalism. If we take this to be Peano Arithmetic, we intend [the provability operator] to be interpreted as (absolute) provability of an *arithmetical* sentence (i.e., one which does not contain [the modal operator]. The iterated [modal operator in the 4 axiom] does not accord with this interpretation.

Moreover, he restricts the Necessitation rule of modal logic to sentences that do not contain occurrences of the modal operator [**Myh60**, p. 470].

What is behind this disagreement between Myhill and Gödel?

Myhill rightly states that for the concept of *mathematical* provability, the 4 axiom does not hold. Take the sentence

(†) It is informally provable that φ.

where φ is some mathematical claim. Sentence (†) is not a mathematical statement, since it contains the *philosophical* concept of informal provability. So sentence (†) is not even the *sort of thing* that can be mathematically proved. (Something similar can be said for the unrestricted necessitation rule.)

This does not mean that Gödel was wrong. In his [**Göd33a**], Gödel interprets his operator as (informal) *provability*, not as mathematical provability [**Göd33a**, p. 300]. For this notion, the unrestricted Necessitation rule should be expected to hold. If the axioms of Gödel's formal system S of provability can be seen to follow from the content of the concept of proof (and of possibility), then the theorems of S have informal (but not always purely mathematical) proofs. In the light of our earlier discussion on p. 88, the question of the validity of the 4 axiom for informal proof relates to the question whether some form of positive introspection holds for the concept of proof. If it does in a sufficiently strong sense, then the 4 axiom holds for the informal notion of provability.

So the distinction between informal proof and informal mathematical proof matters! Myhill saw that the distinction between informal proof and informal mathematical proof is significant in another sense.

The logical treatment of informal provability and of informal mathematical provability as sentential operators is ultimately unsatisfying. We formalise the notion of truth as a *predicate* because we want to be able to quantify over the entities that are true. For the same reason, we should formalise informal provability and informal mathematical provability as predicates.

If, in an arithmetical framework,[31] we treat informal provability as a predicate, then liar-like paradoxes ensue:

[31] We need to work minimally in an arithmetical framework to enable self-reference via coding.

THEOREM 1.2 ([**Myh60**], [**KM60**]). *Let $\mathcal{L}_B$ be the language of Peano Arithmetic plus a primitive (provability) predicate B. Let the system M consist of Peano Arithmetic, formulated in the extended language, plus the necessitation rule (Nec) and the Factivity axiom ($Fact$) for B. Then M is inconsistent.*

PROOF. By the diagonal lemma, take a sentence K such that

$$(\ddagger) \qquad PA \vdash K \leftrightarrow \neg B(K).$$

Reasoning in M, we suppose (for a reductio) that $\neg K$. Then by the $\leftarrow$-direction of ($\ddagger$), we get $B(K)$, from which by $Fact$ we get K. So we reject our assumption, and have $M \vdash K$. By Nec this gives us $M \vdash B(K)$, so by the $\rightarrow$-direction of ($\ddagger$) we also get $M \vdash \neg K$. Contradiction. □

However, if we consider the system M^-, which results from M by restricting Nec and $Fact$ to arithmetical sentences. M^- is easily seen to be consistent: it has models in the natural numbers. Moreover, the same holds for the system for informal *mathematical* provability that Myhill proposes in [**Myh60**, p. 469], which is obtained from M^- by adding the closure of the extension of B under Modus Ponens as an extra axiom.

In sum, the notion of informal provability is prone to paradox in a way that the concept of mathematical provability is not. This is because the notion of mathematical provability is naturally a *typed* notion, whereas the notion of provability is 'reflexive'.[32]

1.6. The standard of proof

So far, we have taken informal mathematical proof, like truth, to be an absolute notion. Over the past decades, truth theory has developed into a very successful research area, both in philosophy[33] and in philosophical logic.[34] But the investigation of informal proof and informal mathematical proof can hardly be said to have taken off. Why is that?

Let us go back to one of the early attempts to "define" the notion of proof [**Göd53**, p. 341, footnote 20]:[35]

> "proof" means [...] a sequence of propositions convincing a sound mind.

Myhill talks about the dangers of "circularities or vagueness" [**Myh60**, p. 462] in attempts to clarify the notion of proof, and writes [**Myh60**, p. 463]:

> I cannot use any of these apodictic notions [such as 'irrational', 'compelled', 'committed',...] to define the apodictic notion of proof, without falling into one of the two classical fallacies of circular definition or of defining the unknown by the more unknown.

[32] For the same reason, the notion of truth is paradoxical, whereas the concept of mathematical truth is not.

[33] See e.g. the work by Horwich's defence ([**Hor98**]) of deflationism.

[34] See for instance [**Hal11**].

[35] Observe that, for reasons given earlier, this cannot count as an explication of the notion of *mathematical* proof.

Gödel's characterisation of the concept of proof seems particularly vulnerable to the latter charge. In particular, one wonders: what is a 'sound mind'? whose mind? how does it get convinced? One gets the feeling that something like Kant's transcendental subject is appealed to here.

Moreover, many philosophers worry that early conceptions of proof, such as that of Gödel, are very (overly?) platonistic. One is inclined to interpret the term 'thought' in Gödel's characterisation of proof in a Fregean, platonic way. Since we have a good theoretical grip on them, it may not be absurd to see the natural numbers, for instance, as somehow removed from human activity. But conceiving of a proof as a platonic entity somehow, regardless of whether one aims at defining the notion of proof or to uncover essential properties of it, risks making it philosophically somewhat intractable. In the end, the proof of the pudding is in the eating. If reflection on mathematical proof as an absolute notion yields powerful and fruitful theories, then this is cause for optimism. But the longer such theories are not forthcoming, the hollower claims about mathematical proof as an absolute notion sound.

The once standard way of investigating mathematical proof is challenged by the *philosophers of mathematical practice*.[36] They advocate a different methodological approach to questions about informal mathematical proof. They contend that we should start from the extension of mathematical proof, rather than from the intension of the concept. This means that our investigation should be firmly rooted in mathematical practice. Mathematical proofs are the sort of things that can be found in mathematical journals and mathematics textbooks. They are what mathematics teachers write on blackboards and explain to students, they are the kinds of arguments that mathematicians give on *MathOverflow*,[37]...

Mathematical proofs are products of a social practice. And like for other such products, there is little reason to expect that they share anything like an essence or nature that is there for philosophers to uncover. There is a great variety of forms that mathematical proofs can take: a mostly a priori investigation into the nature of mathematical proof will tend to lose sight of the multifarious aspects of informal proof. Only a more empirical approach to the problem will do justice to the diversity of what are counted as proofs by the mathematical community.

The history of mathematics shows that the *standards of mathematical proof* have evolved over time. What counted as a geometrical proof for Euclid is not the same as what counted as a geometrical proof for Hilbert in his *Foundations of Geometry* [**Hil99**]; what counted as a proof in analysis for Euler is not the same as what counts as a proof in analysis for the editors of the *Annals of Mathematics* today. Moreover, sociological research about mathematics shows that what counts as a proof for one group (professional algebraic geometers, say), is not the same as what counts as a proof for another group (undergraduate mathematics students, say). In short, informal proof is a relative, not an absolute notion [**LM08**, p. 97].

Consider for instance mathematical arguments in analysis. In the absence of a definition of limit, and of the corresponding distinction between convergent and divergent infinite series, calculations of infinite sums could not rigorously be justified. At best, therefore, many mathematical arguments in analysis from the

[36]The origin of this movement in the philosophy of mathematics lies in Lakatos' *Proofs and refutations* [**Lak76**].

[37]https://mathoverflow.net

eighteenth century do not pass contemporary journal standards of mathematical rigour. At best, they are gappy proofs (although calculations of values of integrals are very often not justified in contemporary analysis articles!); in many cases, they contain mistakes (sins against dividing by 0, for instance).[38]

De Toffoli argues that containing mistakes disqualifies a mathematical argument from being a real proof [**dT21a**, p. 12]. She recognises that proofs in mathematical journals do not live up to the standard set by her definition (see p. 27). She calls the mathematical arguments that are labelled 'proofs' in mathematical journals *simil-proofs*, and defines them thus [**dT21a**, p. 13]:

> An argument is a *Simil-Proof* when it is shareable, and some agents who have judged all its parts to be correct as a result of checking accept it as a proof. Moreover, the argument broadly satisfies the standards of acceptability of the mathematical community to which it is addressed.

So a mathematical argument can be a Simil-Proof and contain significant mistakes.

When we take the standpoint of mathematical practice seriously, it is not so clear that mathematical arguments that contain significant mistakes cannot still rightfully be called mathematical proofs. In this context, I am reminded of a story that Yannis Moschovakis shared at a dinner conversation. As a graduate student, he got hold of a copy of Paul Cohen's manuscript about the independence of the Continuum Hypothesis. He found a significant mistake in the argument; this mistake had also been spotted independently by others. Moschovakis went to his supervisor and told him that the proof was wrong. His supervisor, so Moschovakis told us, replied along the following lines: "We *know* that there is a mistake in the proof. But it can be fixed. Go back and try to understand the independence of the Continuum Hypothesis." It was very clear from Moschovakis's story that his supervisor thought that Cohen had *proved* the independence of the Continuum Hypothesis even though it contained a significant mistake—Moschovakis, of course, in retrospect agreed. Cohen proved this theorem because his argument contained the *mathematical reasons why* the Continuum Hypothesis cannot be decided from ZFC.

According to what may be labelled the received view, the relation between informal and formal mathematical proof can be described as follows [**Avi20**, p. 2–3]:

> When someone in the mathematical community makes a mathematical claim, it is generally possible to express that claim formally, in the sense that logically adept and sufficiently motivated mathematicians can come to agreement that the formal claim expresses the relevant theorem. One justifies an informal claim by proving it, and if the proof is correct, with enough work it can be turned into a formal derivation. Conversely, a formal derivation suffices to justify the informal claim. So an informal mathematical statement is a theorem if and only if its formal counterpart has a formal derivation.

[38] De Toffoli mentions Gauss' original argument for the fundamental theorem of algebra as an example [**dT21a**, p. 12].

This claim about the formalisability of mathematical proofs can be seen as an explication of the sense in which informal mathematical proofs are *rigorous*.

In the same vein, Azzouni takes it to be a normative constraint on informal mathematical proofs that they "*indicate* an 'underlying' derivation" [**Azz04**, p. 84]. He furthermore points to the fact that this explains why disagreements about knowledge claims are so infrequent in mathematics [**Azz04**, p. 84]:

> Since (a) derivations are (in principle) mechanically checkable, and since (b) the algorithmic systems that codify which rules may be applied to produce derivations in a given system are (implicitly, or, often nowadays, explicitly) recognised by mathematicians, it follows that if proofs really are devices mathematicians use to convince one another of one or another mechanically checkable derivation, this suffices to explain why mathematicians are so good at agreeing with one another on whether some proof convincingly establishes a theorem.

So the *standard of (mathematical) proof*, one might say, is that the *derivation-indication* that an informal mathematical argument contains, checks out. This checking out is an 'in principle' matter: there is no expectation that any mathematician attempts to work even parts of the 'underlying derivation' in formal detail. Indeed, when evaluating mathematical proofs, mathematicians usually operate on a much higher level of abstraction of general ideas and strategies.[39]

Like knowledge, mathematical proof does not seem to be a matter of degree. Rota expresses this sentiment in his characteristically subtle way as follows [**Rot97**, p. 183]: "The expression 'correct proof' is redundant. Mathematical proof does not admit of degrees. A sequence of steps in an argument is either a proof, or else it is gibberish." Nonetheless, surely mathematical proof is a somewhat *vague* concept.[40] So the last statement in this quotation must be taken with a grain of salt. If the gaps become very large, or the mistakes become very substantial, then one becomes less inclined to say that the author has really proved the mathematical statement, even if it is true. So it can still, in particular cases, be difficult to determine whether the standard of proof has been met.

Let us then assume that formalisability is a normative constraint on informal mathematical proof. And let us then also assume that this formalisation can be seen as a *completion* of the complex of reasons in an informal mathematical proof, rather than a drastic deformation of it.

Now the question arises: in *which* formal system will this formalisation take place? Different mathematical fields have their own *local* formal systems. But all mathematical fields draw freely on other fields. So a formalisation of a proof in one field is likely to draw on axiom systems from other fields. As an example, consider number theory. Mathematical induction is a local axiom that will always play a

---

[39]Thurston puts it as follows [**Thu94**, p. 9]:

> When people are doing mathematics, the flow of ideas and the social standard of validity is much more reliable than formal documents. People are usually not very good in checking *formal correctness* of proofs, but they are quite good at detecting potential weaknesses or flaws in proofs.

[40]I do not commit myself here on the question whether vagueness is always a purely epistemic matter.

central role in number theory. But in recent decades, number theory has made use of category theoretic arguments, so, indirectly, it appeals to basic principles of category theory.

Ultimately, the basic concepts of all mathematical theories can be defined in the theory of pure sets. In this sense, set theory constitutes not a local system but a *foundational* system. Under this set theoretic guise, their basic axioms can be proved from the basic principles of set theory.[41] In this context, we can pursue the example of number theory a bit further. It is not surprising that under familiar definitions of natural numbers in terms of sets, the axiom of mathematical induction can be proved to hold. After all, set theory has induction built in almost at its level of axioms, namely as the principle of transfinite induction on ordinals.

No one seriously claims that the set theoretic definitions of mathematical entities from other mathematical disciplines constitute *ontological reductions* of entities of these disciplines (the natural numbers, graphs,...) to pure sets. The untenability of such a claim was convincingly argued for in [**Ben65**]. Nor does anyone suggest that, for instance, a graph theorist "should really work in set theory": that would be madness. These points have been repeated ad nauseam, so let us not dwell on them further here.

As far as the axioms of set theory go, the principles of Zermelo-Fraenkel set theory with the Axiom of Choice (ZFC) is almost universally accepted by the mathematical community. In fact, the vast majority of mathematicians would be happy to go a bit beyond ZFC, if necessary. If someone were to prove the Riemann Hypothesis from ZFC plus the axiom that there are inaccessible cardinals, for example, then it would be considered *proved*—not that anyone currently expects that a large cardinal axiom (or its consistency) is needed to prove the Riemann Hypothesis. In fact, much less than full ZFC is needed to "reconstruct" virtually all of mainstream mathematics.

1.7. Warrant for axioms

Why are philosophers interested in the question of the standard of proof? Ultimately, the complex of reasons that a mathematician has must support her belief in a mathematical statement. If this complex of reasons is an associate formal proof, then it seems that we may be able to make a case that it does. We should keep in mind that the mathematician will not be conscious of *all* reasons in the formal proof, nor does she need to be. One reason for this is that there usually is not a *unique* way, up to trivial transformations, that an informal mathematical proof can be formalised. So we are giving ourselves some latitude when we say that *her* reasons are the reasons in a formalisation of her informal mathematical proof.

Suppose, then, that a mathematician has an informal mathematical proof of a mathematical statement φ. If the foregoing, is correct, there will then be a formal proof of φ that can be seen as the 'completion' of the complex of reasons contained in the informal proof. The formal proof then constitutes the complex of reasons that justifies our mathematician's belief in φ. Let us look at this formal proof from an epistemological point of view.

[41] For category theory (viz. the 'category of all categories') this is perhaps less straightforwardly so. But everything that can be done in category theory can be done in set theory as long as very mild large cardinal assumptions are used (such as the existence of inaccessible cardinals).

The formal proof will be a proof in some formal system. Let us ignore, for the moment, the question how the logical inference steps in this formal proofs are warranted. Let us also ignore the question, which perhaps only rarely arises, what our warrant for basic *non-logical* inference steps in the formal proof is. Then we are left with the question how the mathematician's belief in the premises of the formal proof is warranted. As said before (p. 26), some of these reasons will be formalisations of theorems that were proved elsewhere. Perhaps the author of the informal proof has not gone through the proofs of some of these theorems. In those cases, the question arises what the mathematician's warrant is for believing them. Let us also postpone this question. Then we are left with the difficult question which cannot be put off forever, namely:

> *Wherein consists the mathematician's warrant for her belief in the basic axioms on which her formal proof rests?*

Much has been written on this fundamental question. Gödel's writings on it[42] have been particularly influential. Gödel focuses on our warrant for believing in the axioms of set theory. We have seen that beside the foundational axioms, there are local axioms to consider (for instance, arithmetical axioms). So we will also have a few words to say about the justification of local axioms. Moreover, I will not follow Gödel closely in what follows.

Gödel distinguishes between *intrinsic* and *extrinsic* evidence for set theoretic axioms. Let us first consider intrinsic warrant for axioms.

It is claimed that many of the basic axioms of set theory can be seen to be true on the *iterative conception of sets*.[43] The basic idea is that sets are somehow formed in *stages*, where at each next stage, the full power set of of the previous stage is formed. Then if certain assumptions on the structure of the stages are made, many basic axioms of set theory can be justified. Let us briefly look at how this works for the Axiom of Choice. The truth of an antecedent of this principle means that there is a stage α where a family of non-overlapping sets has been generated. Then there must be a prior stage β where each element of each of the sets in the family has been formed. Therefore at stage $\beta + 1$ a choice set is formed.

Despite some opinions to the contrary, most philosophers do not believe that *all* axioms of ZFC can be motivated in this way. For instance, it is hard to see how the Replacement Axiom, which says that the image of a set under a functional correspondence also forms a set, can be motivated in this way.[44] But there is a second intrinsic way in which some axioms of set theory can be justified, namely by arguing for them from the *limitation of size* principle, which says that every collection that is not *too large* to form a set, forms a set.[45] This principle can be used, for instance, to justify the Axiom of Replacement. Suppose that an instance of the antecedent of Replacement holds. Since the size of the image is bounded by the set that bounds its pre-image, this image forms a set. A third way of justifying set theoretic axioms is based on the thought that the mathematical universe cannot be *uniquely* characterised in a mathematical manner. This idea can also be formulated

---

[42] See [**Göd47**].

[43] For a description of how this goes, see [**Sho77**].

[44] See [**Boo71**]. For instance, Schoenfield in his description of the iterative conception just *stipulates* that the stages are such that they validate Replacement: see [**Sho77**, p. 324].

[45] See [**Hal84**, Chapters 4 and 5].

in a positive manner: if a mathematical property holds of the mathematical universe, then it also holds of a mathematical object that is not the universe, i.e., of a *set* (REF). Such arguments are called arguments from *reflection*.[46] This argument can be used to argue for the Axiom of Infinity. The mathematical universe contains infinitely many elements. So, by reflection, there is also a *set* that contains infinitely many elements; in particular, the natural numbers therefore have to form a set. To conclude, there is a fourth kind of intrinsic warrant for some basic axioms of set theory: some basic principles of set theory are taken to be *analytically true*. For instance, the Axiom of Extensionality appears true by virtue of the content of the concept of set. At some point, the mathematical community reached a consensus that the *extensional* notion of set is what set theory is about. (Perhaps something similar can be said about the Axiom of Foundation.) Some think that most if not all basic axioms of set theory are analytical;[47] but this is again a minority position. Indeed, this only seems plausible if the meaning of the concept of analyticity is stretched beyond its usual limits.

Of course there is no general agreement over the question to what extent these attempted justifications achieve their aim. However, there is a school of thought that takes these forms of intrinsic evidence, *in combination*—even if it is perhaps overkill—to provide warrant for *all* axioms of ZFC. But there is a fundamental difference between the kind of warrant provided by the iterative conception and the limitation of size conception on the one hand, and the kind of warrant provided by analyticity on the other hand. Warrants from the iterative conception or from limitation of size are complexes of reasons. So they fall under the category of *justifications*. (How strong such justifications are, is of course a matter of dispute.) But it would be an *overintellectualisation* if we were to say that warrants from analyticity are likewise complexes of reasons. We do not have to go through a complex of reasons in order to be warranted to accept the Axiom of Extensionality. It is through understanding the concept of set that we cannot help but take the Axiom of Extensionality to be self-evident, and thus come 'automatically' to accept it. Thus our ordinary warrant for the Axiom of Extensionality does not have propositional structure: facts about relations between meanings here operate at the level of causes rather than at the level of reasons. This is not to suggest that we cannot *reflect* on this type of warrant (as I have just done) and also produce reasons for believing the Axiom of Extensionality. We will have more to say later about the type of warrant of which our ordinary warrant for the Axiom of Extensionality is an instance.

There are also *extrinsic* reasons for believing axioms of set theory. We saw earlier (Section 1.3) that Quine believed that ultimately mathematical axioms are justified by the empirical success of theories of which they are an essential part. In the same vein, Gödel argues that intra-mathematical success can also count as a convincing reason for believing a proposed new axiom [**Göd47**, p. 521]:[48]

[46]See [**Pas07**, Section 2]. For a general overview of the role of reflection in the foundations of set theory, see [**Rob**].

[47]See [**Par90**, Section 7].

[48]The point that mathematics has standards for success that are largely autonomous from empirical science is also strongly emphasised in [**Mad07**].

> Furthermore, however, even disregarding the intrinsic necessity of some new axiom, and even in case it had no intrinsic necessity at all, a decision about its truth is possible also in another way, namely, inductively by studying its "success," that is, its fruitfulness in consequences and in particular in "verifiable" consequences, i.e., consequences demonstrable without the new axiom, whose proofs by means of the new axiom, however, are considerably simpler and easier to discover, and make it possible to condense into one proof many different proofs. [...] There might exist axioms so abundant in their verifiable consequences, shedding so much light upon a whole discipline, and furnishing such powerful methods for solving given problems [...] that quite irrespective of their intrinsic necessity they would have to be assumed at least in the same sense as any well-established physical theory.

In other words, mathematical axioms can be justified by *inference to the best explanation*. In these situations, according to the view under consideration, we have something like a virtuous feedback mechanism, i.e., a virtuous circle: 'good' concrete consequences follow from an abstract axiom, and these consequences in turn abductively support the axiom. So the axiom is a reason for believing the consequences, and the consequences are reasons for believing the axiom. Some philosophers believe that not much more than the axioms of ZFC are intrinsically warranted; for axioms that go significantly beyond what can be proved in ZFC, extrinsic warrant is all we have.[49]

We have seen that beside the foundational axioms, there are local mathematical axioms. Views about our warrant for local mathematical axioms vary widely. I do not propose to give an overview of these views here, but restrict myself to a few remarks. Some local axioms can be regarded as being analytic. For instance, the axiom that 0 is the smallest natural number seems warranted by the content of our concept of natural number. But not all local axioms are analytic. For instance, a "miniature" version of the iterative conception is thought by some to play a role in our warrant for some of the central arithmetical axioms, such as the axiom that every natural number has a successor.[50] After all, there is a sense in which the natural numbers can be regarded as "iteratively generated" as well (by the successor relation).

1.8. A leaching problem

The mathematical warrant that an *individual mathematician* has for her belief in a mathematical statement φ frequently consists in her having gone through an informal mathematical proof of φ. This proof typically appeals to theorems that have been proved elsewhere, and *their* proofs appeal to other theorems. Thus a chain of proofs is generated that eventually leads back to fundamental axioms.

Our mathematician cannot be expected to have gone through all the proofs in this chain. She simply accepts some of the "auxiliary theorems" on faith (because they are listed as theorems in an authoritative textbook; because Y, a specialist in the field, told her about the theorems;...). But the mathematical warrant for φ of

[49]For an extended philosophical discussion of this theme, see [**Mad88**].

[50]See [**Par07**, p. 173–174].

the whole *mathematical community* consists of the combination of the proofs in the chain:[51] the mathematicall community has no one to defer to. So the mathematical warrant for φ of the mathematical community ultimately rests on its mathematical warrant for the axioms. It is at this point that difficult questions arise.

Suppose that our justification for one of the axioms of φ is given by *intrinsic reasons*. For definiteness, suppose that the proof of φ depends on the Axiom of Choice, and that our only mathematical warrant for the Axiom of Choice is given by an argument from the iterative conception along the lines sketched above (p. 36). Then we would only have a *philosophical* warrant for the Axiom of Choice, and not a mathematical one. After all, some of the concepts in these explanations (such as 'generated at a stage') are not mathematical ones. As a consequence, the mathematical community's warrant for φ would not be purely mathematical. But this seems wrong: mathematics does not answer to a higher epistemic tribunal.[52] A mathematician need not go through the argument for the Axiom of Choice from the iterative conception in order to be in good epistemic standing when she uses the Axiom of Choice in her proofs.

This is a *leaching problem*: non-mathematical reasons "leach" into the overall warrant for a mathematical statement. It is not only a problem for justifications from the iterative conception, but for all justifications by intrinsic reasons (justifications from limitation of size, justifications from reflection). This worry does not extend, however, to our warrants from analyticity, for such warrants are not given by *reasons* (matematical or other). What about *extrinsic* warrants for axioms?

As we have seen earlier, we must distinguish between intra-mathematical and extra-mathematical success arguments. Intra-mathematical variants, such as those of Gödel and Maddy, take mathematics to be epistemically autonomous. On their view, mathematics needs no external epistemic shoring up. In this way, they respect the epistemic autonomy of mathematics better than extra-mathematical success arguments.

Nonetheless, the leaching problem affects not only extra-mathematical, but also intra-mathematical versions of success arguments. According to these views, belief in axioms is warranted by inference to the best explanation arguments. *Inference* means that these warrants are complexes of reasons. *Explanation* is a philosophical concept, so these complexes are not purely mathematical in nature. Therefore these views still do not fully respect the epistemic autonomy of mathematics. A mathematician need not go through an inference to the best explanation argument in order to be warranted in believing the axioms of her field of research. Moreover, inferences to the best explanation are, as we have seen, to some extent *circular* complexes of reasons. As arguments go, they are not particularly good ones. It is therefore not immediately clear how much epistemic force a success argument in mathematics has.

Explanatory power is an epistemic virtue. Simplicity, testability, coherence, fruitfulness,... are other such. Epistemic virtues are (hopefully) truth-conducive. In order to be in good epistemic standing, a mathematician has to be highly sensitive and responsive to these virtues. But at its basic level, this responsiveness is a quasi-causal mechanism that *directly* results in belief-formation and belief-change.

---

[51] I assume here that the mathematical community has no other mathematical warrant for φ.

[52] See [**Mad07**, Part IV, Section 3].

This mechanism does not have to be conceptualised into complexes of reasons. Thus mathematicians can be, and very often are, epistemically warranted in their belief in mathematical axioms without having *reasons* for their belief. In this way, the mathematical community as a whole becomes *mathematically* warranted in its belief that φ, where φ is a mathematical theorem. In particular, it is mathematically warranted in its belief of the axioms without having a justification for them. Thus the mathematical community is unaffected by the leaching problem.

We may hypothesise that, unconsciously, this responsiveness *is* conceptualised by mathematicians in the form of arguments such as inferences to the best explanation. But this is mere speculation, and somewhat doubtful. But above all, it is not *needed* to be mathematically warranted to believe a mathematical axiom.

The philosopher can reflect on the mathematician's responsiveness to epistemic virtues, and produce philosophical reasons for belief in mathematical axioms. This puts the philosopher in a *different* epistemic position vis-à-vis the axioms. If her philosophical reasons are good reasons, then they provide a distinctive kind of warrant for them, namely a *justification*. But the unphilosophical mathematician is no less warranted in her belief in the axioms without having this justification for them. She may, for instance, be deeply sceptical of what are in her view hopelessly vague terms, such as 'explanation', 'simplicity',...

At the outset, I claimed that a strong form of *externalism* is implausible for mathematical knowledge (p. 15). Indeed, in our discussion of mathematical proof, we have dwelled much on the role of reasons in mathematical justification. Nonetheless, the position that we have ended up with has externalist overtones. In particular, the mathematical community's mathematical warrant for axioms is external. A brief comparison with externalism in contemporary epistemology may clarify the position that we have landed on further.

Reliabilism is at the moment probably the most popular view in general epistemology. According to a classical version of reliabilism, Julie has a justified belief in φ if and only if the belief-forming process that was used by her in her formation of the belief that φ is reliable.[53] From our point of view, more or less by definition this cannot be right. I hold that a person's justification is a matter of her reasons. But Julie need not have access to the reliability of her belief-forming process in order for her belief in φ to be epistemically warranted. Similarly, I take the mathematical community to be epistemically warranted but not justified in its belief in mathematical axioms. Nevertheless, the external factors that make Julie's belief warranted *can* become internal reasons for her if she reflects on them.

Moreover, such need not be the only internal grounds for warranted belief. To illustrate this, let us now briefly return to the iterative conception of sets. For all we know, it may be essentially correct: sets may in fact somehow be "generated in stages". In that case, the philosophical argument for the Axiom of Choice given earlier (p. 36) tracks the truth. Then this reasoning can be used to *justify* one's belief in the Axiom of Choice.

So the following is not excluded. Mathematician A is warranted in her belief in the Axiom of Choice by her practical responsiveness to the epistemic virtues of this principle. Philosopher B is justified in believing the Axiom of Choice by motivating this principle from the iterative conception of sets. A and B can even be the same person: she can be a mathematician *and* a philosopher. Then she is

[53] See [**Gol79**].

doubly warranted in her belief in the Axiom of Choice. Something similar can be said for other strategies for justifying axioms on intrinsic grounds.

All this depends on there being a type of epistemic warrant that does not function as a complex of reasons. In the next Chapter, we turn to the question what the nature and properties of such a type of warrant could be.

CHAPTER 2

Epistemic Entitlement

In the previous Chapter, we discussed one particular type of epistemic warrant: justification. Moreover, since the general notion is familiar, and our concern is to a large extent with the epistemology of mathematics, we focussed on types of *mathematical* justification.

In this Chapter, I will introduce a second type of epistemic warrant: epistemic *entitlement*. The discussion of this notion will mostly be held at a more general level: not much will be said about the way in which epistemic entitlement plays a role in contemporary debates in the epistemology of mathematics. In later Chapters I will relate this notion to certain specific debates in mathematical epistemology.

Compared to the concept of justification, the notion of epistemic entitlement has entered the epistemic literature at a late stage in the history of philosophy. There is less agreement about the content, viability, and scope of this concept than there is about the notion of justification.[1] It is not my aim to give a neutral and balanced overview of philosophical accounts of epistemic entitlement that are currently explored, or to trace the history of the concept of entitlement in epistemology. Instead, I focus on what I regard as some of the seminal work on this concept that is, in my opinion, highly relevant for mathematical epistemology.

Specifically, I will chiefly be concerned with the views of epistemic entitlement by Tyler Burge and by Crispin Wright (with important additions made by Robert Audi). Burge's and Wright's accounts are seen as very different, and there is not much overlap between the bodies of epistemological literature to which they have given rise. Nonetheless, I will argue that if we abstract from the irreconcilable differences in their overall epistemological outlook—roughly speaking: Wright is more of an epistemological internalist, whereas Burge is more of an externalist,—their views on epistemic entitlement are not only largely compatible but even complementary. In the last Section of this Chapter, I foreshadow an epistemological account of an epistemological process of reflection that is further developed in Chapter 8: this account rests on and combines elements of Burge's and of Wright's accounts of epistemic entitlement.

It is of course important to know how justification and epistemic entitlement relate to each other. It is my hope that in the course of the present Chapter, the reader will acquire a feeling for the mechanics of the interplay between these two kinds of warrant.

[1] The fact that the notion of epistemic entitlement seems to be making its way to introductory textbooks—see for instance [**Wil01**, Chapter 13]—provides grounds for optimism about the future.

2.1. Two types of epistemic warrant

I will start by giving an account of the *theoretical core* of Burge's theory of epistemic warrant.[2] This theory is often taken to be very subtle, or even somewhat obscure, but undeservedly so. What follows is an attempt to give a *clear* description of Burge's view of epistemic warrant.

An epistemic warrant has a source, and it has an object in the sense that it is a warrant *for something*. In the previous Chapter, we restricted our discussion to cases where the source of a warrant is a complex of reasons, and the object of a warrant is a belief. We have seen how the resulting belief is then called *justified*, and the warrant for it is called a *justification*. Moreover, we mostly restricted our attention to cases where the belief is a *mathematical* belief. In this Chapter, we take a wider perspective. We begin by considering warrants that do not have complexes of reasons as their source. In later Sections, we also consider warrants for things other than beliefs.[3]

Reasoning is one *competence* that humans have for rational belief-formation. But, as we will see, there are other such competences: perception, interlocution, memory, reflection... Because these competences are *rational* competences, we have an epistemic right to apply them in belief-formation. Exercise of these competences results in warranted belief.[4]

Burge has much to say about conditions that need to be satisfied for a competence to be able to deliver warranted beliefs. In brief, the competence in question must serve as a *guide to the truth*. This means that when the competence is *functioning normally*, and is exercised in *normal circumstances*, it generates true beliefs. This means that Burge holds a *reliabilist* position in epistemology, which is a form of externalism. But Burge is a *moderate externalist*, since he reserves in his theory a central place for an internalist notion of justification.

It is a common complaint against reliabilism in epistemology that concepts such as "functioning normally" and "normal circumstances" are left rather vague. This is no different in Burge's writings. On this score, no real progress is made: this is not where the action is. I will not try to solve this problem here. Another standard objection is that reliability is not sufficient for epistemic warrant. This protestation traces back to a fictional example by Bonjour [**Bon85**, p. 41]:

> Norman, under certain conditions which usually obtain, is a completely reliable clairvoyant with respect to certain kinds of subject matter. He possesses no evidence or reasons of any kind for or against the general possibility of such a cognitive power or for or against the thesis that he possesses it. One day Norman comes to believe that the President is in New York City, though he has no evidence either for or against this belief. In fact the belief is true and results from his clairvoyant power under circumstances in which it is completely reliable.

---

[2]The seminal article here is [**Bur93**]. Burge spells out his views on epistemic warrant further, and modifies it in some respects, in a series of articles, most of which can be found in Part II of [**Bur13a**]. In [**Gra20**], Peter Graham gives an excellent overview of Burge's views in epistemic warrant.

[3]See Sections 2.5, 2.7, 2.8, and 2.9.

[4]For a discussion of rational competences for belief-formation, see [**Gra20**, Section 1].

In order to rule out being warranted to believe Norman on the basis of relying of Norman's reliability, Peacocke adds another condition [**Pea04**, p. 11]:

> A fundamental and irreducible part of what makes a transition one to which one is entitled is that the transition tends to lead to true judgments [...] in a distinctive way characteristic of rational transitions.

Then it becomes important to explain what *distinctive ways characteristic of rational transitions* are: this is in part what Peacocke's book is about. Again, I will not take a stance on whether Peacocke's solution is ultimately satisfactory. Rather, I take away from this discussion that the reliability in question need not hold in very abnormal situations, but should nonetheless have some counterfactual strength. Indeed, if Norman's reliability were not accidental, if it were not a statistical fluke, then we might come to be entitled to rely on it. Of course, we have strong prima facie grounds for being sceptical about the counterfactual strength and even the lastingness of the reliability of Norman's belief-forming process. After all, other people do not seem to have the cognitive powers that he allegedly has. So we would have thoroughly to test Norman's powers of clairvoyance before we are epistemically warranted to rely on them. More in general, *if* we have prima facie grounds for doubting the reliability of a cognitive process, then we are not epistemically warranted to rely on it in our belief-formation process before these doubts are assuaged.

A belief-forming competence can be exercised by someone in a situation that is not normal, when she is unaware of the situation's abnormality. Moreover, a normally reliable competence that is exercised can malfunction without her being aware of this. In such situations, her resulting belief is still warranted.[5] In other words, in such a situation our agent is from an epistemic point of view in no way blameworthy. When a normally reliable belief-forming mechanism malfunctions but still produces a true belief, the resulting belief does not qualify as knowledge. Suppose, for instance, that Melissa hallucinates a cloud in the sky, and on the basis of her hallucination forms the belief that there is a cloud in the sky, while there coincidentally is a cloud in the sky. Then Melissa does not *know* that there is a cloud in the sky. This just goes to show that Gettier can strike not only for justified true belief, but for other kinds of warranted belief as well.

The sources of warrants generated by the exercise of perception or interlocution are not complexes of reasons. The source of a perceptual warrant for a belief is a perceptual representation. A perceptual representation does not have propositional structure and therefore is not a putative reason.[6] The source of an interlocutionary warrant for a belief is someone's assertion. Someone's assertion is not a proposition (even though its content is), so it is not a reason.[7] Thus the traditional identification between epistemic warrant and justification is challenged. Burge breaks with epistemological tradition by arguing that we can in certain circumstances be entitled to believe a proposition without having to do justifiying work.

Propositions can be the source of warrants for beliefs without these warrants being justifications. In a person an innate connection may exist, which ensures that

---

[5] See [**Gra20**, Section 2], [**Bur03**, p. 506–507].

[6] See [**Bur03**].

[7] See [**Bur93**], [**Bur97**]. For a somewhat different but highly interesting epistemological account of interlocution, see [**Mor05**].

any occurrent belief in which the concept F occurs, immediately and automatically causes a belief that danger is present (q). This causal pathway may be a reliable route to true beliefs that q. Nonetheless, it would be wrong to say that in such cases, F-beliefs function as *reasons* for believing that q [**Bur11b**, p. 494]. Therefore, in those cases, the person in question is *entitled* but not justified in her belief that q, even though the source of her belief that q is a proposition.

All this shows that on Burge's account, the sources of epistemic entitlements are varied: visual representations, for instance, are quite different entities than assertions. It is not clear that Burge's suggested classification system for warrants is the most useful one. Rather than distinguishing only two main species of warrants, one might see justification as just one special type of warrant among many different kinds of warrant. Thus we might speak of justificational warrant, interlocutionary warrant, perceptual warrant, preservational warrant (memory), etcetera. In what follows, I will sometimes speak in this way.

When the source of a warrant of someone's belief does not contain reasons, Burge calls the resulting warrant an *entitlement* (rather than a justification), and the resulting belief is then an *entitled* (rather than a justified) belief. In practice, if a belief is warranted, the warrant is likely to involve both reasons and entities that are not reasons. If the source of a warrant contains at least one reason, then Burge calls the warrant a justification [**Bur13b**, p. 3–4]:

> A *justification* is a warrant that consists partly in the operation or possession of a reason. An individual is justified if and only if the reason is operative or relied upon in the individual's psychology. An *entitlement* is a warrant whose force does not consist, even partly, in the individual's using or having a reason.

Burge thus regards justification and entitlement as the two species of the genus warrant [**Bur11a**, p. 489]. Warrants can be of only two distinct kinds: those that involve reasons, and those that don't.

By way of example, I might have a belief that has the form of a conjunction $p \wedge q$, and have arrived at it in the following way:

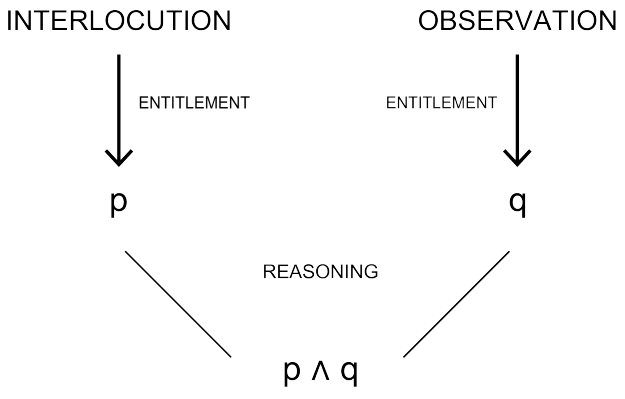

FIGURE 1

So for Burge, in warrants, reasons are *dominant*, and non-reasons are somehow *recessive*.[8]

2.2. Preservative memory and the a priori

Our discussion of Burge's theory of warrant has so far been rather abstract. Let us now see how it does work for us in epistemology. The first application of his theory of warrant that we will consider in some detail, concerns the epistemology of memory.[9]

Most mathematical proofs are too long for people to keep in mind as a whole at any one given time. Therefore mathematicians rely on memory when they prove theorems. Patricia recalls that she proved a mathematical proposition φ last year, but she no longer recalls how her proof went. Presently, she uses φ as a premise in her proof of a new theorem ψ. Since she has proved ψ, she knows ψ. Is her knowledge of ψ a priori, or is it empirical?

On the standard view, Patricia's knowledge of ψ, like most mathematical knowledge, is a priori. However, Chisholm formulated an argument that her knowledge is only a posteriori. His argument goes more or less as follows [**Chi77**].

Among Patricia's reasons for ψ, we find the belief that she would express by the sentence:

(‡) I remember that I proved φ.

Granted, in the scenario under consideration, (‡) is true, and Patricia knows that (‡). But (‡) is not the sort of statement that can be known a priori. If introspection delivers a priori knowledge, then she may know a priori that she *thinks* that she has proved φ. But finding out whether she really did, requires empirical work. For instance, she might consult her diary, find an entry which says "I have proved φ today", and take this as empirical confirmation of (‡). But now, since at least one of her reasons is empirical, by the recessiveness of a priori reasons (see p. 19), Patricia's knowledge of ψ is a posteriori. If this is right, then already for considerations such as these alone, much of Patricia's mathematical knowledge is not a priori.

Against Chisholm, and in agreement with the standard view, Burge argues that in the scenario under consideration, Patricia acquires a priori knowledge that ψ. If that is so, then where does Chisholm's argument go wrong?

The problem is, according to Burge, that (‡) is not a reason that is part of Patricia's justification of ψ; (‡) is not a premise or a step in Patricia's justification. When she first proved φ, she acquired a justified belief that φ. But the occurrent belief that φ that she forms while she is constructing a belief that ψ, is not *inferred*. It is not based on one or more reasons, so it is not justified. Nonetheless, if her memory is functioning normally—and here we are assuming that it is—then Patricia's present occurrent belief is warranted. Hence her warrant is of a different kind than justification: her warrant is an *entitlement*!

Moreover, if Patricia's warrant for φ was a priori, when she proved this statement, then when she later remembers φ, her warrant for that occurrent belief is likewise a priori. In this way, preservative memory is a mechanism for shifting a belief, with its whole epistemological status intact, from one time and context to another.

---

[8] Compare this with Passeau's distinction (see p. 19) between dominant and recessive *reasons*.

[9] See [**Bur93**].

It is a familiar refrain in contemporary epistemology that aprioricity does not entail infallibility. So it is in Patricia's case. Her exercise of her competence of preservative memory malfunctions from time to time. This does not mean that in each case, she has an epistemic duty to check if her competence is functioning properly on the occasion. Patricia need not be aware of the fact that her memory is functioning normally in order to be entitled to her belief. It is sufficient that her competence is a reliable guide to the truth. This is an expression of Burge's externalist outlook.

Only when she has evidence to the contrary that it does that cannot properly be ignored, must she acquire overriding evidence that her competence does function properly on that occasion. Indeed, results from experimental psychology teach us that in many contexts, our memory is *systematically* unreliable. There are therefore many situations in which we are not warranted to rely on our memory without (empirically) checking its proper functioning. And in those situations where we appropriately do so, and rightly conclude that our memory functions properly in that instance, the resulting beliefs constitute a posteriori knowledge if the knowledge of the remembered proposition was acquired in an a priori fashion. But there are also many cases where the warrant delivered by the faculty of memory does not need to be shored up by empirical checks—we are assuming that the above scenario is a case in point. Moreover, even when the empirical checks need to be done, are carried out, and override the antecedently available counterevidence, this does not mean that the default entitlement to believe what memory suggests on this occasion has been forever 'canceled' by the counterevidence and you only have empirical justification for your belief. On the contrary, your default entitlement is rehabilitated and typically reinforced by the empirical evidence.

If the world is *very* uncooperative, then Patricia can gather as much evidence that her competence is functioning well as she likes, while it is in fact malfunctioning and delivers false beliefs. In such a case, Patricia is still entitled to these beliefs, for she is epistemically blameless. But because her beliefs are false, they then don't constitute knowledge. This is only the Gettier phenomenon all over again.

2.3. Interlocution and computer proofs

In the previous Section, we saw how Burge argued from the distinction between justification and entitlement, to the claim that reliance on memory in a proof does not ruin the a priori status of the proved theorem. He thus employed a new epistemological distinction for a rather unsurprising conclusion. Now we will see how Burge used the very same distinction for a surprising conclusion, and how he later shied away from drawing the surprising conclusion.

We again start with a fictional scenario. You find yourself in an unfamiliar town, and ask an arbitrary passer-by for the shortest route to the train station. She says: "Take the second right, and then the first left. That will take you straight to the train station." You take her word for it, and do as she suggests.

Again we ask what the epistemological status is of your belief that p, where p is the proposition that the quickest way to the station is to take the second right, and then the first left.

In this case, it is controversial whether your belief that p constitutes knowledge at all. You have formed your belief that p as an immediate response to her telling you that p. You do not *infer* your belief from anything, so your belief is not justified.

In particular, you do not infer your belief from the belief that people are more likely to tell the truth than not. Even the thought that this passer-by was more likely than not in this situation to tell the truth, did not cross your mind.

Burge claims that in this scenario, you know p. Knowledge requires warrant, and your warrant is not justification. So your warrant is an *entitlement*: an "entitlement of interlocution", we might say. If in a scenario such as this you would not be warranted to believe p, then we would not know much. After all, so many of our beliefs rest in part or in full on what we have been told. It would be unreasonable and unjust to accuse you, in the scenario under consideration, of dereliction of epistemic duty. For the vast majority of what we are told, we simply do not have the time, energy, and resources to gather independent confirmation that our interlocutor (the postman, the newsreader, the woman in the street, the clerk in the town hall,...) is a reliable source of information.

If you have evidence against the reliability of your interlocutor, then this evidence should not be ignored. In that case, further evidence must be gathered that on this occasion, despite this counterevidence, is truthful. If on this occasion, evidence is found that trumps the initial counterevidence, then you are *justified* in believing what your interlocutor tells you. But in our fictional scenario, no initial counterevidence is present, so the need for evidence-gathering does not arise. The default position is that you take your interlocutor as a source of reason.

Even though, in the scenario under consideration, you do not need to do so, you can, if you want to, try to seek empirical evidence for the reliability of your interlocutor's testimony. Suppose you seek such evidence, and obtain it. Then your initial entitlement to believe p does not disappear. You still have it. Your entitlement has even become stronger than before: your empirical evidence reinforces it [**Bur13c**, p. 270]. Suppose you are in a situation where you have evidence that your interlocutor's testimony might not be reliable in a particular instance. Then you must gather further evidence before trusting what she says. Suppose you do gather further evidence, the further evidence that you find is inductive in nature and it assuages your initial fears. Then you can trust your interlocutor: the initial entitlement has been reinstated. In this case, you are both entitled and inductively inductively justified to believe what you are told.

Thus far, Burge has argued for a controversial, but not altogether surprising claim. But in [**Bur93**] and [**Bur97**], he goes further and argues for a truly remarkable contention. He argues that, like in the case of memory, through interlocution it is possible to acquire *a priori* warrant. To see how this goes, consider a slightly different scenario. Your friend Christine tells you that the Tanayama-Weil conjecture has been proved. Again, you simply take her word for it. You do not even go to the library to check out where this is supposed to be proved in order to verify that the argument was given by reputed mathematicians in a renown journal of number theory. Again Burge claims that when you thus form your belief in the Tanayama-Weil conjecture, your belief constitutes knowledge. But Burge goes further. He claims that, as in the case of memory, the epistemic status of the warrant for the Tanayama-Weil theorem, as we may now call it, is transferred from the mathematicians who proved it to you. Since their justification of it is a priori, so is your entitled belief [**Bur93**, p. 251–252]. Thus, through interlocution, you acquire a priori knowledge of the Tanayama-Weil conjecture.

This is a surprising claim. We tend to think that a priori knowledge is knowledge that is arrived at "purely internally", i.e., without relying on what is delivered to us by the external senses. But your auditory experience has played an essential role in your acquisition of the belief that the Tanayama-Weil conjecture is true.

However, matters are not that simple. We have seen earlier[10] that sense experience is sometimes ineliminably involved in the acquisition of a priori knowledge: the question is *how* sense experience is involved. Burge claimed that in interlocution, perception of our interlocuter's assertion play only a *triggering* role, but that it does not contribute to the force of the warrant [**Bur97**, p. 294]:

> Strictly speaking, we do not perceive the assertive mode, or the conceptual content, of utterances. We understand them. [...] We understand assertions by perceiving other aspects of assertions. We understand the concepts in assertions, by perceiving expressions of them. But here perception is part of the condition for exercising the intellectual capacity, not—or not normally—part of the warrant for the individual's relying on his understanding. It is a necessary triggering mechanism, but it is not the understanding itself.

In [**Bur98a**], Burge related his view on the epistemology of interlocution to a philosophical debate about computer-assisted proofs.

We have seen earlier that there are mathematical proofs that essentially rely on the assistance of computers for checking cases.[11] Tymoczko has argued that such proofs can only deliver *a posteriori* knowledge of theorems [**Tym79**]. The irreducibly empirical element in our justification of the four colour theorem, for instance, is the fact that we only have empirical evidence for the claim that the computer has checked the cases correctly: it would take us humans prohibitively long mentally to go through the computer-generated proofs of the cases. This is why the relevant computer programs are run several times over and on multiple computers, why several programs for checking one and the same case are written and executed, etcetera.[12]

Burge took issue with Tymoczko's view of the epistemological status of computer proofs [**Bur98a**]. He argued that in computer-assisted proofs, computers can be treated as interlocutors. Because humans have programmed them, they can be seen, like humans, as sources of reason. It is true that we have to read the computer output ("`Case 431 has been verified...`") in order to prove the theorem. But this empirical element in the proceedings only plays a triggering, or enabling role; it does not contribute to the force of the warrant. Moreover, Burge admits that without the empirical checks on the correctness of the computer programmes, and the correctness of their execution, Appel and Haken cannot be said to have proved the four colour theorem. The empirical checks are needed because of the difficulty of the problem. An extraordinarily difficult problem requires extraordinary rational powers for its solution. The empirical checks are part of what allows us to access this rational source. Again, they do not contribute to our warrant for believing what this rational source delivers to us. In sum, despite the empirical element

[10] See p. 19.

[11] See p. 20.

[12] See for instance [**AH78**].

involved in computer-assisted proofs, Burge concluded[13] that they deliver a priori knowledge just as traditional proofs do.

In later work, Burge recants his earlier claim that interlocution can deliver *a priori* warrant [**Bur13c**, p. 284]:

> Knowledge that relies on warrant for comprehension—including knowledge that relies essentially on the default prima facie warrant to believe what another says on a particular occasion—is always empirical, even if sometimes just barely.[14]

This retraction is significant.[15] Burge is a reliabilist, but also a *rationalist*. His work is concerned with safeguarding and if possible extending the domain of the a priori [**Bur13a**, Preface]. If Burge's startling claims about the epistemology of interlocution and computer proofs were correct, then they would contribute to his rationalist programme. They would support the claim that much of our mathematical knowledge is a priori. Mathematicians rely on interlocution on a daily basis in their work. They believe that a theorem is true on the basis of their colleague saying so. They consult textbooks of subjects in fields that are related to their field in order to find theorems that might help them solve their problem—very often they do not go through the proofs of these auxiliary mathematical propositions. If all this affects the a priori status of their mathematical knowledge, then most mathematicians don't even have a priori knowledge of their own best theorems.

But perhaps the loss for Burge's rationalist project that results from the retraction of his startling claims, is not as dramatic as it may seem. Its impact is mitigated when we shift our attention from the individual mathematician to the mathematical community. Even if knowledge through interlocution is always empirical—if only barely so, as Burge says,— interlocution still secures "a priori paths" to most of our mathematical knowledge. It is just that for most mathematical theorems, no single mathematician walks the whole path: she hitches a ride part of the way. And as far as computer proofs are concerned: as long as the engineers have done their job well, what is wrong with hitching a ride in a driverless car?

However this may be, much of Burge's account survives this retraction. It is still true, according to Burge, that our warrant for knowledge through interlocution is of a fundamentally different kind than justification. In the case of computer proofs, the situation is less clear, because the empirical elements involved seem somehow more fundamental. Burge's considered view seems to be that our warrant for believing computer proofs to be partly inductive in nature [**Bur11b**, p. 501]:

> A computer-assisted proof, like the proof of the Four-Color Theorem, does not give anyone full understanding of the proof. The mathematician knows much of the proof, understands the principles used in it, and has inductive reason to think that the computer has carried out a proof. Understanding is partial. It

---

[13]Somewhat tentatively: see [**Bur98a**, p. 341].

[14]In the same vein, he says elsewhere : "Contrary to what I said in 'Content Preservation', and elsewhere, I think that the comprehension that is needed to bring pieces of communication from others under the Acceptance Principle is inevitably warranted empirically. The force of warrant for one's comprehension depends on perceiving others' linguistic input competently and reliably" [**Bur13b**, p. 31]. (The Acceptance Principle is discussed in Section 2.4.)

[15]As Burge admits: see [**Bur13b**, p. 31].

> is partial understanding of how the proof goes, backed by inductive ground that the proof has been completed. It is partial, idealized, but genuine understanding of the necessity of the conclusion given the premises.

2.4. The acceptance principle

In a way similar to memory, we have an *entitlement* to believe that we exist [**Bur98b**]. The Cartesian *Cogito* thought is not inferred from reasons. Our warrant for our belief in our own existence is a priori, so the situation is more similar to that of memory than to that of interlocution. On the other hand, in a way that is structurally similar to interlocution, we have an *entitlement* to believe in the existence of other minds [**Bur13e**]. Unlike memory, and like interlocution, our warrant for belief in other minds is ("barely") a posteriori: our perceptions (seeing, hearing,... other people) play a role in the force of the warrant.

Burge believes that there is a general principle behind these different kinds of entitlements. He calls this the *Acceptance Principle* [**Bur93**, p. 237]:

> A person is entitled to accept as true something that is true and that is intelligible to him, unless there are stronger reasons not to do so.

For Burge, this is a rock bottom principle of rationality. It is the cornerstone of his theory of epistemic warrant. He believes that it can a priori be seen to be true. Indeed, in his *justification of the Acceptance Principle*, he explains how it flows from the nature of rationality [**Bur93**, p. 238]:

> A person is entitled to accept a proposition that is presented as true and that is acceptable to him, unless there are stronger reasons not to do so, because it is prima facie preserved (received) from a rational source, or resource from reason; reliance on rational sources—or resources for reason—is, other things equal, necessary to the function of reason.

A few pages later, he explains his justification of the Acceptance Principle further as follows [**Bur93**, p. 240–241]:

> We are apriori prima facie entitled to accept something that is prima facie intelligible and presented as true. For prima facie intelligible propositional contents prima facie presented as true bear an apriori prima facie conceptual relation to a rational source of true presentations-as-true: Intelligible propositional expressions presuppose rational abilities and entitlements; so the intelligible presentations-as-true come prima facie backed by a rational source or resource for reason; and both the content of intelligible propositional presentations-as-true and the prima facie rationality of their source indicate a prima facie source of truth.

Fricker remarks that it seems to her that Burge's "justification of the Acceptance Principle" is not intended to be *suasive*: it does not intend to convince someone who somehow initially finds the Acceptance Principle doubtful [**Fri06**, Section 4]. Burge agrees, and clarifies the status of his "justification" as follows [**Bur13c**, p. 266, footnote 18]:

> I believe that reflection on practice and taking care to avoid hyper-intellectualization are the best grounds for a philosopher's coming to accept the principle. My account is intended to articulate an underlying rationale for the principle, granting its truth.

We have seen how Burge withdrew his earlier claims about the possibility of purely a priori knowledge through interlocution. But he held onto the Acceptance Principle and its rationale. In particular, he maintains up to this day that we have epistemological warrants for what people tell us that fundamentally differ from justifications. We have also seen how this is bound up with Burge's fundamentally rationalist outlook. From an empiricist viewpoint, all this does not make much sense. Indeed, from an empiricist starting point, Fricker has challenged the Acceptance Principle [**Fri06**].

Fricker argues first of all that the Acceptance Principle is superfluous. She believes that "the typical position of a mature adult faced with a piece of testimony is that she has in her cognitive background, and brings to bear, a wealth of empirical knowledge relevant to the assessment of that testimony", and therefore "she does not need recourse to a default principle licensing its acceptance in the absence of such relevant empirical information" [**Fri06**, p. 81]. Moreover, she argues that it is irrational for people to apply the Acceptance Principle: it is a "charter for gullibility" [**Fri06**, p. 80]: people very frequently do not tell the truth, and in many circumstances, it is *rational* for people not to tell the truth.

Instead of relying on the Acceptance Principle, human adults do and should draw on background knowledge about human nature (folk psychology) and social roles when assessing the veracity of acts of testimony [**Fri06**, p. 83]. This background knowledge is broadly *inductively* justified. Thus our warrant for believing what someone tells us is ultimately inductive in nature.

Burge responds to Fricker's critique of the Acceptance Principle in [**Bur13c**, Section II]. We saw above how Burge urges us to be wary of hyperintellectualisation. This admonition was specifically directed at Fricker's account of testimony. Consider again our example on p. 48. Is it really credible that in scenarios such as this, I implicitly go through an empirical argument to conclude that there is a high probability that the passer-by on this occasion speaks the truth? Burge adds to this that for very small children it is *in principle* impossible to go through such an argument. They have not been able to build up the required induction base for such an argument. Moreover, they have not acquired the requisite inductive reasoning skills, nor do they even have the conceptual arsenal that its required for acquiring these skills [**Bur13b**, p. 26–27]. And how would they acquire warrants for these reasoning skills: inductively? Fricker could reply that children do not have the same epistemic obligations as adults do.[16] But then, how could children, or the human species as a whole, for that matter, ever acquire the required warrant for the needed canons of inductive reasoning?

Here we find ourselves getting drawn into the age-old debate between rationalism and empiricism. It is needless to say that this debate will not be settled in this book. My aim is far more modest. In subsequent Chapters, I want to explore the potential of Burge's theory of entitlement further, relate it to Crispin

[16]It indeed seems that Fricker mainly argues that it is unreasonable for "mature adult humans" to adhere to the Acceptance Principle.

Wright's theory of entitlement, to extend it a bit, and to apply it to new epistemic phenomena.

2.5. Entitlement and scepticism

The concept of entitlement also plays a key role in an influential article by Crispin Wright [**Wri04b**]. Wright is mainly concerned with our epistemic warrant for certain general propositions that he calls *cornerstone propositions*. Cornerstone propositions are statements that play an organising role in our cognitive representation of reality. An example is the proposition:

There is an external world.

Wright is occupied with sceptical challenges that seek to undermine our epistemic warrant for our cornerstone propositions, and thereby undermine our presumed warrant for believing ordinary propositions (such as "it is snowing outside") in everyday circumstances.

One might think that scepticism about the outside world can in a Moorean fashion ([**Moo39**]) be refuted by what Wright calls a I-II-III argument [**Wri02**]:

I My visual perception suggests to me that I have hands.
II I have hands.
III There is an external world.

In Wright's view, such an argument for an anti-sceptical conclusion is not rationally acceptable. We have a case of *warrant transmission failure*. The problem is that the argument from I to III is *question-begging*: one can only rationally accept the argument as a whole, and in particular the inference from I to II, on the condition that III holds. So it seems that we have to establish III in an independent way, and there seems no way to do this.

If Burge is right, then it suffices to look at my two hands, and say to myself "here are two hands", i.e., to come to believe that I have two hands on the basis of my perception. On Burge's account, we are entitled to believe in II *directly* on the basis of our perception instead of on the basis of an *inference* from a statement about perceptual seemings. Since III can be seen to follow from II, we are *justified* in believing that there is an external world.

Whether Moorean arguments yield justification for belief in an external world is related to deep issues in the philosophy of perception. We have seen how Burge defends a externalist picture of our warrant for perceptual beliefs. Wright, on the other hand, is an internalist in these matters. It is beyond the scope of this book to adjudicate between Burge's and Wright's accounts of perceptual belief. For now, however, let us accept Wright's description of the Moorean argument. Let us also, for now, accept Wright's claim that the Moorean argument, thus understood, does not yield justified belief in the existence of an external world, and see where this leads him.

Wright proposes to resolve the conundrum by denying that III is in need of justification. Instead, we are entitled to *assume* or *presuppose* or *accept* III without justification. Proposition III is then a *presupposition of cognitive project*: doubting III would rationally commit one to doubting the significance or competence of one's cognitive project[17] [**Wri04b**, 193]. The presupposition III then allows us to

[17]Wright uses the notion 'cognitive project' in a somewhat technical sense: see below, Section 2.6. For now, the intuitive notion of cognitive project suffices.

be *justified* in inferring II from I: I am therefore justified in believing that I have hands

The notion of *entitlement of cognitive project* can then be defined along the following lines [**Wri04b**, 191–192]:

> ...an entitlement of cognitive project [...] may be proposed to be any presupposition P of a cognitive project meeting the following additional two conditions:
> (i) We have no sufficient reason to believe that P is untrue
> (ii) The attempt to justify P would involve further presuppositions in turn of no more secure a prior standing...

In the light of this, Wright then holds that we are entitled to *accept* or *assume* or *trust* or *presuppose* or *presume* cornerstone proposition III without having justification for it. This *still* does not give us the epistemic right to *believe* III for the very reasons that we have gone through before: it would be question-begging.

Earlier we saw that an epistemic warrant has a source and an object (p. 44). Now we see that the object of an epistemic warrant need not always be a belief: it can also be a presumption. More needs to be said about what a presumption (or assumption, or presupposition, or...) is. Moreover, it is not clear that to accept, to assume, to trust,...are all one and the the same epistemic attitude. However, let us leave these worries aside for now, and postpone them until Section 2.7.

In the process, a concession has been made to the sceptic: I have as yet no *knowledge* of III, even though I am entitled to presuppose it. In this sense, Wright proposes a *sceptical* solution to the sceptical challenge. But something important has been gained in the process: my trust in III earns me the epistemic right to believe particular beliefs such as II, i.e., my ordinary belief-forming processes are rational.

Not everyone agrees with Wright's thesis of the non-transference of warrant. James Pryor argues that *under certain circumstances*, going through a I-II-III argument can give an epistemic agent, call her Elisabeth, knowledge of the existence of an outside world [**Pry04**]. The following is a highly idealised scenario that nonetheless conveys the idea. Suppose Elisabeth has, to begin with, an open mind concerning the existence of objects outside her own mind. She looks at her hands, has the experience as of having hands, and forms the belief that she has an experience as if she has hands. On the basis of her visual experience she forms the belief that she has hands. From the proposition that she has hands she infers that there is an external world: she comes to believe that there is an external world on the basis of this argument. Then, Pryor, says, Elisabeth has acquired justified belief in the proposition that there is an external world [**Pry04**, p. 369]. So in this situation, her warrant does transfer.

In the situation where Elisabeth has *doubts* or *reservations* about the existence of an external world, the situation is different. Then she cannot transfer her warrant from I to II, and therefore not get via the I-II-III argument to knowing III. Indeed, the I-II-III argument is not *dialectically effective* in a discussion with a sceptic: it is not a line of reasoning that the sceptic can accept. After all, the sceptic would insist, we cannot exclude a Cartesian scenario in which a malicious demon causes me to have an experience as if I have hands, even though no material world exists. Moreover, it is also not even a good *philosophical response* to the sceptic (even though a good philosophical response to the sceptic need not be dialectically

effective against the sceptic), for it fails to diagnose and criticise the argument of the sceptic.

In sum, Pryor insists that someone *can* come to know that there is an outside world by going through the I-II-III argument. At the same time he concedes to Wright that the I-II-III argument does not take away sceptical worries.

Another worry about Wright's account was raised by Jenkins.[18] She suspects that Wright is committed to *epistemic consequentialism*. According to Wright, Victoria (say), going through the I-II-III argument, is entitled to trust in III *because* doing so has good epistemic consequences. Jenkins then goes on to argue that having good epistemic consequences is never in and of itself a good reason for adopting an epistemic attitude.

Jenkins recognises that Wright does not want to appeal to epistemic consequentialism in his account [**Jen07**, p. 28]. Her point is that that it seems difficult to justify Victoria's entitlement to trusting in III (even if she herself does not have to posses this justification) in any other way than by appealing to its epistemic consequences [**Jen07**, p. 27–31].

Jenkins' critique rests on a misunderstanding of Wright's project. Wright is not trying to *justify* Victoria's epistemic entitlement to III. In this sense, Wright's account is Wittgensteinian: all explanations must come to an end, and Wright's assertion that Victoria is entitled to her trust in III is a fundamental claim of his theory. That people are in similar situations entitled to what they rely is not derived from more basic principles of rationality. Instead, the plausibility of the claim that Victoria is entitled to her trust in III (and similar claims) is supposed to derive from the plausibility of the overall picture of rationality that Wright proposes. Ultimately, Wright intends to give a descriptively accurate account of what people are epistemically entitled to presuppose. Like Burge has proposed putative rock bottom principles of rationality, Wright has proposed his basic claims about rationality.

Davies has argued that entitled trust in III is not needed to warrant Victoria in believing II. He argues that, rather than an entitlement to to trusting in III, having a *negative entitlement* not to doubt III suffices. This negative entitlement is an entitlement "not to adopt the attitude of doubt where Wright has an entitlement to adopt the attitude of trust" [**Dav04**, p. 226]. Davies says that this negative entitlement is not a kind of epistemic warrant, for it "is not an entitlement to assume, trust, or believe any proposition" [**Dav04**, p. 243]. This can perhaps be disputed, since it can be seen as an epistemic warrant not to doubt something. Be that as it may: when this backing of a negative entitlement is in place, normal perception yields a *positive entitlement* to believe II. All in all, this takes us, as Davies recognises [**Dav04**, p. 230–231], fairly close to a Burgean account of Victoria's warrant for believing in II.

2.6. Cognitive projects and their presuppositions

We have seen how Wright argues that we have an *entitlement of cognitive project* to assume III. But what, exactly, is a cognitive project?

A cognitive project is "defined by a pair: a question, and a procedure one might competently execute in order to answer it" [**Wri12**, p. 466]. This is a bit abstract.

[18] See [**Jen07**].

Waxman explains in more concrete terms what this amounts to [**Wax17**, Chapter 3]:

> Let a cognitive project be a pair consisting of a question Q and a procedure P which, if successfully executed, will deliver an answer to Q in which you have justification to believe. For example there is a cognitive project associated with the question "what is the time right now?" and the procedure of looking at one's watch; there is another that involves the same question and the different procedure of asking a passer-by. Next, let us say that a presupposition for a given cognitive project is a condition C such that, if you were justified in believing that C failed to obtain, any justification in the output of the procedure would be defeated. The thought is that, in many if not all cases, a cognitive project is bound to rest on the satisfaction of a number of conditions such that, if doubts were to arise about them, would compromise the ability of the procedure to return a justified verdict on the relevant question. So the presuppositions for the project ⟨what is the time?, looking at one's watch⟩ include conditions like: the watch's being set correctly when it was last set; its subsequent normal functioning (without any major mechanical problems); there being nothing interfering with the veridicality of our perception of the clock-face; our being able to tell the time, i.e. to understand what time the watch is representing it as being; and so on.

By extension, a cognitive project may consist of a set of questions accompanied by a smaller or larger battery of procedures. For instance, on a large scale, mathematics is a cognitive project, with mathematical proof as one of its main 'procedures'. Metaphysics is, I think, is also a cognitive project. And larger scale cognitive projects may have smaller scale projects as sub-projects. Thus Wright concept of cognitive project contains echos of Kuhn's notion of paradigm [**Kuh62**], Foucault's notion of disciplinary matrix [**Fou66**], and Lakatos' notion of a scientific research programme [**Lak68**]. The difference is just that the latter are by definition rather large-scale, whereas Wright's cognitive projects can vary in scale from very small to very large.

Presuppositions always have to be made in order to have justification [**Wri04b**, p. 189]:

> To take it that one has acquired a justification for a particular proposition by the appropriate exercise of appropriate cognitive capacities–perception, introspection, memory, or intellection, for instance–always involves various kinds of presupposition. These presuppositions will include the proper functioning of the relevant cognitive capacities, the suitability of the occasion and circumstances for their effective function, and indeed the integrity of the very concepts involved in the formulation of the issue in question.

Wright's point then is that we have an entitlement of cognitive project to accepting all the presuppositions of the cognitive project ⟨do I have hands?, looking⟩, and III is one of them.

I doubt that *presupposition* is a felicitous choice of terminology in this context. Since Wright takes presuppositions to attach to collections of *beliefs*, the notion of presupposition that is used here cannot be the *semantic* notion of presupposition that goes back to Strawson's seminal article [**Str50**]. Thus we are driven to a more *pragmatic* notion of presupposition. But this also does not quite fit. Stalnaker, for example, writes that "presuppositions, on [the pragmatic] account, are something like background beliefs of the speaker" [**Sta74**, p. 198]. But the *entitled* 'presuppositions' that Wright is interested in, are precisely propositions that *shouldn't* be believed by the speaker. So perhaps an attitude that is somewhere in the neighbourhood of, but not quite like, presupposition is what is relevant here. In the following Section, we take a closer look at this question.

2.7. Belief, acceptance, trust

Part of Davies and Burge's critique of Wright's view of perceptual warrant (see Section 2.5) is justified. Wright's *internalist* reconstruction of the Moorean argument seems a result of over-intellecualisation: we do not typically *infer* II from something like I. A more externalist account, along the lines given by Burge, seems more plausible. Whether Victoria's entitlement to her belief II nonetheless requires an entitled some form of acceptance of III, is another matter. In the absence of more information about the content of the attitude of trust that is appealed to, this is hard to adjudicate. So far, the relevant attitude has been described in a number of different ways: relying on, presupposing, trusting (implicitly), assuming, accepting, presuming, taking for granted...[19] It is not at all clear that all the terms have the same content. It will not do simply to list a battery of attitude terms, and hope that one or more of them function as they are supposed to. We must do better.

One hallmark of acceptance, in Wright's sense of the word, is that if a person accepts p, then she *acts in all respects as if she believes* p [**Wri04b**, p. 180]. But this criterion is purely behavioural. Therefore it is unsatisfactory as an ultimate account of the content of the relevant concept. Indeed, answering 'yes' to the question 'do you believe that p' is part of acting in *all* respects as if believing p. But that would mean that acceptance is belief-entailing after all, which would completely undermine Wright's account. Nonetheless, Wright's account of the notion of acceptance is not without merits. It clearly brings out the fact that the attitude that is relevant in this context has *pragmatic aspects* as some of their key components: acceptance is intimately connected with action.

Van Fraassen famously described a notion of acceptance (of a theory) that is not purely 'as if'. It does not entail full belief in the theory, but only the belief that the theory is empirically adequate [**vF80**, Chapter 2, Section 1]. The pragmatic aspect of acceptance is also emphasised by van Fraassen, who argues that acceptance of a scientific theory involves a practical *commitment* to the theory and to the wider research programme to which it belongs [**vF80**, p. 12]. It involves a long-term commitment to follow the theory where it leads, and to take it as a practical guide for action (designing experiments, articulating research questions,...). This long-term aspect is also a property of the attitude that Wright is appealing to. Notions such as supposing, assuming, presuming, do not carry this connotation of being

[19] A number of these supposedly equivalent terms can be found on [**Wri04b**, p. 176].

long-term.[20] For this reason, I regard these terms as unsuitable for conveying what is aimed at. For van Fraassen, even though long-term, a commitment to an empirical theory is never absolute. If it leads to persistent conflicts with experience, for instance, then the commitment may well be given up.

Van Fraassen's view points to an important distinction within the concept of acceptance. Intuitively, acceptance can be guarded, or it can be unrestricted and unconditional, and it can be anywhere in between. One paradigmatic example of guarded acceptance is *instrumental acceptance*. The kind of acceptance of empirical scientific theories that van Fraassen recommends is instrumental. Likewise, the kind of acceptance of higher mathematics that Hilbert recommended is instrumental [**Hil26**]. For van Fraassen, scientific theories are primarily guides to empirical predictions; for Hilbert, higher mathematics is primarily a guide to finitary mathematical statements.[21]

In the context of mathematical knowledge, Torkel Franzén distinguishes in Hilbertian vein between *accepting as consistent* and *accepting as sound* [**Fra04a**]. Like van Fraassen, he takes acceptance to involve belief (belief of consistency, or belief of truth). He takes accepting a theory S as consistent to entail the belief that S is consistent, and he takes accepting S as sound to entail belief that S is true. However, it is important to keep in mind that a person might believe, in a dispositional sense, all theorems of a theory, without believing that the theory is sound. This can be the case, for instance, if she doe not (or not yet) possess the concept of truth. This may sound like nitpicking, but it will prove to be relevant later.

In a prescient remark in his philosophical notebooks, Gödel anticipates the subtleties involved in the relation between belief and acceptance, and in the distinction between different forms of acceptance of a formal theory [**Göd21**, p. 242]:

> Psychology Remark: What is lost in the transformation from the inference rules to the concept of an immediate deduction is the question of "acceptance" of a formal system (only the description remains). To accept is: to make it the system of maxims behind one's assumptions and actions. This is different from "believing it" in the sense that every sentence represents an objective reality and that one can see (perceive) this objective reality. It is also different from belief in consistency. Other possible "attitudes": belief that every controllable conclusion is true, provisionally accepted [...]

However all this may be, we have seen that for Wright's purposes, the notion of acceptance should not entail belief at all. So van Fraassen's and Franzén's concepts of acceptance will not do at all.

Jonathan Cohen describes the relation between belief and acceptance as follows [**Coh89**, p. 368]:

> [I]n my sense to accept that p is to have or adopt a policy of deeming, positing, or postulating that p—that is, of going along with that proposition (either for the long term or for immediate purposes only) as a premiss in some or all contexts for

[20] See [**Coh89**, p. 368].

[21] Van Fraassen classifies the kind of acceptance of our best scientific theories that he thinks is rational as *full acceptance*, but I take that to be a mere terminological matter.

> one's own and others' proofs, argumentations, inferences, deliberations, etc., whether or not one assents and whether or not one feels it to be true that p [...] Belief that p, on the other hand, is a disposition to feel it true that p, whether or not one goes along with the proposition as a premiss.

Observe that, like Franzén, Cohen's definition of belief entails that in order for a person to believe any proposition whatsoever, she must have a concept of truth, which seems wrong. However, let us not dwell on this for now.

Acceptance, in Cohen's sense, does not entail belief [**Coh89**, p. 369]:[22]

> [A] person who does not fully believe that p can nevertheless justifiably accept that p. For example, this may happen when he has a hunch that not-p, though the balance of presently available evidence makes p the only opinion that deserves acceptance within his community. Or he might accept that p out of solidarity with an old friend, even though there is no evidence to produce a belief that p. Or for professional purposes a lawyer might accept that his client is not guilty even though he does not believe it.

Nonetheless, Cohen's notion of acceptance also cannot play the role that it is expected to play in Wright's account. Belief is involuntary,[23] whereas acceptance, on Cohen's understanding of the term, is voluntary [**Coh89**, p. 369–370]:

> [Y]ou cannot decide to believe that it will rain tomorrow, or to believe that it will not. You can decide only to accept that it will, or to accept that it will not: the belief may then ensue, but it may not. Acceptance occurs at will, because at bottom it executes a choice—the accepter's choice of which propositions to take as his premisses. But belief is not normally achieved at will because it is caused in each kind of case by something independent of the believer's immediate choice [...]

But consider Victoria's 'acceptance' that there is an external world once more. This is not the product of a voluntary decision at all. It is just as involuntary as my belief that it will not rain tomorrow.

I conclude that the concept of acceptance is not quite suitable for playing the role that it is expected to play in Wright's account of epistemic entitlement. I will now argue that the notion of *trust* has better prospects for fitting the bill.

The distinction between *propositional belief* and *objectual belief* is familiar enough. Objectual belief *primarily* has persons as its objects. We believe a person (on an occasion) when we believe what she says (on that occasion). In a derived sense, objectual belief sometimes has other entities as its object ("I do not believe that clock."). A form of belief that has some importance for our discussion is the notion of believing *in*. Audi calls this notion *attitudinal belief* [**Aud08**, p. 88]. In the primary sense of that concept, we again believe in people. In a secondary sense of attitudinal belief, we believe also in other things than people: someone might be said to believe in democracy, for instance.

---

[22] Also the converse implication does not hold [**Coh89**, p. 369].

[23] This thesis is defended against Descartes' doxastic voluntarism in [**Wil78**].

Audi denies that attitudinal belief is a form of objectual belief [**Aud08**, p. 88]. I am not sure about this. Attitudinal belief seems closer to objectual belief than to propositional belief. Believing a person seems almost a species of attitudinal belief.[24] I said earlier that when we say that we believe a person (on an occasion), we mean that we believe the assertions that she produces. In the same vein, it is tempting to surmise that when we believe *in* a person, this amounts to having a *positive attitude*[25] towards this person's future actions. But this is not quite right. Our positive attitude is in the first place directed to the person herself as the *source* of future actions;[26] only derivatively do we have a good feeling about what we think these future actions might be.

Anscombe held that the notion of *believing a person* is absolutely primary, and that the notion of propositional belief is very secondary and somehow derivative from it [**Ans79**]. This view is not as bizarre as it might seem. The positive attitude that is operative in believing a person is one of *trust*. It hardly necessary to add that trust is a heavily pragmatically laden concept. On the occasion that we trust a person, we mostly do not trust her with our life, but we always trust her with some of our actions. Believing a person means having a modicum of trust in her (on the occasion, at least). Also in believing *in* a person, trust is a central component. Even in propositional belief, there is trust: we trust the believed statement as a 'producer' of a proposition in a way that is similar to the way in which we can trust a person as a source of propositions. Propositional belief has primitive origins.

If this is along the right lines, then belief always involves trust. But we are interested in the converse direction. *Is there trust without belief?*

As with belief, there is objectual trust (in the first place directed at people), attitudinal trust, and propositional trust. Victoria trusts her friend Evelyn, she has trust *in* Evelyn, and she trusts that there is an external world. An even stronger case than with belief can be made for the thesis that propositional trust is derivative from objectual or attitudinal trust. When we say that Victoria trusts that there is an external world, what do we mean? At least large part of the meaning of that sentence is that Victoria trusts her outer senses, that she willingly relies on her outer senses to tell her how things are.

In *this* sense, even a very young child may be said to trust that there is an external world. All it takes, is that she accepts the offerings of her outer senses at face value. For this to be true of her, she need not even yet possess any of the concepts 'outer sense', 'external world', or 'truth'. We, as mature adults, have by no means outgrown this basic trust in our senses. It is only in extreme cases of psychic pathology, or in extremely unusual circumstances (involving hallucinogens, for instance) that this form of trust in our outer senses is absent. In such situations, our cognitive faculty is debilitated.

Our young child does not *believe* that there is an external world. She has not yet mastered the concepts that are needed to entertain this thought: she trusts without believing. Again, mature adults are no exception. As Audi points out, I can trust that my good friend survives cancer, without either believing or disbelieving this [**Aud08**, p. 92].

---

[24]'Almost' because it concerns past and present events as well as future ones.

[25]This is a term of Audi: See [**Aud08**, p. 90].

[26]This point is stressed in [**Mor05**].

This does not mean that there is no connection between propositional trust and propositional belief. According to Audi, propositional trust entails a *disposition* or openness to the corresponding propositional belief [**Aud08**, p. 96]. In the context of a discussion of religious faith instead of secular trust,[27] he writes [**Aud08**, p. 93]:

> If I have faith that God loves humanity, I have a certain positive disposition toward the proposition that this is so. This disposition is something beyond hope. But the cognitive component of propositional faith, though stronger than the minimal cognitive element required for hoping, does not entail belief. Propositional theistic faith is, to be sure, incompatible with believing that God does not exist; but that is a different point. Because of the positive way in which propositional faith is more than hope, it is also incompatible with a pervasive or dominating doubt that God exists, though it can coexist with some degree of doubt or even with a tendency to have moments of deeply unsettling doubt.

However, even the thesis that propositional trust entails a disposition to believe seems wrong to me, for the following reason. Very often, a person has propositional trust without having considered the proposition in question, even though, as a mature adult, nothing prevents her from considering it. Consider Nathalie, who is out on a day trip without sunscreen or her sun hat. When she leaves her appartment, she sees the tube of sunscreen on the living room table, and her sun hat on the coat rack, but she sees no reason to take them—which is not to say that she sees a reason not to take them! When she looks out the window before she leaves, she says to herself: 'it is going to be a glorious day'. It is apparent from these and other aspects of her behaviour that she trusts that she won't get sunburnt, even though she has not asked herself the question. Now suppose that her friend asks her if she does not think she will get sunburnt. Nathalie chews this over for a few moments, and judges her trust to be careless. She forms the belief that she *will* get sunburnt, and proceeds to act accordingly. This shows that it would be simple-minded to try to reduce belief to unreflective behaviour—which does not mean, of course, that there are no close links between belief and behaviour.

With Audi, I distinguish between *doxastic* (propositional) trust and *fiducial* (propositional) trust [**Aud08**, p. 96].[28] Often, propositional trust is in fact accompanied by the corresponding propositional belief. Because propositional trust is often accompanied by propositional belief, there is a strong temptation to analyse propositional trust in terms of propositional belief. But since doxastic trust does not entail propositional trust, this temptation must be resisted: *trust must be understood in its own terms* [**Aud08**, p. 101]. This is not easy. The temptation is great to reduce propositional trust to some combination of propositional belief and action, or even to behaviour only, However, when we take Audi's recommendation to heart, we see that propositional belief does not accompany propositional trust as often as is commonly thought.

Propositional trust can to some extent be, but need not be *voluntary* [**Aud08**, p. 91–92]. Certainly Victoria's trust that there is an external world cannot be

[27] I believe that there are indeed deep connections between trust in cornerstone propositions and religious propositional faith, but I will not press the point here.

[28] Again, Audi speaks of faith instead of trust.

labelled as voluntary. In a discussion of Hume's scepticism, Strawson writes [**Str85**, p. 11]:

> [Hume] points out that all arguments in support of the skeptical position are totally inefficacious; and, by the same token, all arguments against it are totally idle. His point is really the very simple one that, whatever arguments may be produced on one side or the other of the question, we simply cannot help believing in the existence of body, and cannot help forming beliefs and expectations in general accordance with the basic canons of induction. He might have added, though he did not discuss this question, that the belief in the existence of other people (hence other minds) is equally inescapable.

Here Strawson is speaking of belief in the existence of body. But his remarks apply even more to our *trust* that there is an external world.

Propositional trust is not a luminous state. Someone can trust that p, without being aware that she trusts that p.

Some instances of propositional trust are *rational* or warranted, others are not. Nathalie's initial trust that she would not get sunburned, for instance, was careless and therefore unwarranted. Doxastic propositional trust may be supported by *reasons*. In certain cases, therefore, doxastic trust can be justified. In cases of fiducial trust, there is no belief to be supported by reasons, so they cannot be justified. Nonetheless, fiducial trust can be rational.[29] Rational instances of fiducial trust are cases of *entitled trust*. In sum, I claim that fiducial trust is the attitude that meets the needs of Wright's account: it is the kind of attitude to which we can have Wrightean entitlement.

Moreover, I claim that on this point, Wright's view is in harmony with Burge's account. Burge has argued that we have an entitlement to certain belief-forming competences.[30] So, like Wright, Burge recognises entitlements that have objects other than beliefs. In particular, we have an entitlement to rely on our preception [**Bur03**]. Our fiducial trust in the existence of an external world is warranted by our entitlement to rely on our outer senses. In sum, what appear to be distinctive features of Wright's account of entitlement—such as our entitlement to rely on III—have natural counterparts in Burge's theory.

Note that this not mean that Burge's theory provides answers to *all* questions regarding entitlement for trust. For instance, as far as I can tell, Burge's theory does not speak to the question of the existence and nature of my fiducial entitlement that my friend will survive cancer. Also, the foregoing should not obscure the fact that Burge's epistemological theory is overall very different from that of Wright. In particular, I remind the reader (see p. 58) that I side with Burge, and against Wright, on embedding the theory of entitlement in an *externalist* overall framework.

2.8. Inference and entitlement

Suppose Alice is taught the proof of Pythagoras' theorem. She believes the mathematical basic principles involved in the proof in a justified manner. She follows the proof, and is on the basis of the argument convinced that the theorem

[29] Audi discusses the question of fiducial faith in [**Aud08**, p. 100].

[30] An excellent account of Burge's view of entitlement to rely on competences is found in [**Gra20**].

holds. In such a scenario, we are inclined to say that Alice has come to *know* Pythagoras' theorem.

Alice does not need to to have a justified belief in the validity of the logical inference rules used in the proof in order to come to know Pythagoras' theorem. She is entitled to trust implicitly in the logical steps that she is taking when she is following the proof: she *reasons blindly* [**Bog03**].

Indeed, Burge would say that we have a prima facie entitlement to rely on rules of inference in our reasoning. Our entitlement "resides in [our] actual competence to make the relevant deductive transitions, not in an ability to understand and represent the rule governing the competence" [**Bur11b**, p. 492]. In Wright's terminology, this is another entitlement of cognitive project. We have here another entitlement of cognitive project [**Wri04a**]. This shows that entitlements of cognitive project do not only consist in trust that certain matters of fact obtain; they also comprise trust in the reliability of *actions* (taking logical inference steps).

Suppose that we are entitled to rely on our logical faculties in this way. In particular, supposed that we are entitled to trust the Rule of Modus Ponens (MPR) in this way. Then a next question is whether we are also warranted to, in Wright's terminology, *claim* knowledge of the corresponding logical principle, i.e., the material conditional corresponding to MPR. Let us call this corresponding conditional MPP ("Modus Ponens Principle"). Wright answers this question in the following way [**Wri04a**, Section VIII, p. 173]:

> ... [I]f we are entitled to claim knowledge of a statement which we have recognised to follow from known premises by inference in accordance with entitled rules, then we are surely entitled to claim knowledge of a statement which we have recognised to follow from an *empty set* of premises by inference in accordance with entitled rules. But—assuming an entitlement to [MPR] and conditional proof—that is just what a rule-circular derivation of [MPP] provides for.

Such a derivation could proceed like this:

1	(i)	P	Assumption
2	(ii)	If P, then Q	Assumption
$1, 2$	(iii)	Q	$(i), (ii)$ M. Ponens
1	(iv)	If (if P, then Q), then Q	$(ii), (iii)$ Cond. Proof
	(v)	If P, then if (if P, then Q), then Q	$(i), (iv)$ Cond. Proof

This seems right. So we have a (schematic) *justification* of the logical principle (viz. MPP) that guarantees the reliability of the rule MPR. Remember how Pryor claimed that in a similar way, we obtain a justification of our belief in principle III that underwrites the reliability of our perceptual faculties (see p.56). This raises the puzzle why exactly, in Wright's view, the argument quoted above "transmits warrant", whereas no Moorean argument does. I will not try to resolve this puzzle here.

It is commonly assumed that the kind of epistemic warrant for believing the conclusion that is generated by application of logical rules is always justification. But this is not so. Consider again the familiar propositional logical rule $\rightarrow$I of conditional proof ("Arrow-Introduction"):

$$\begin{array}{ccc} (i) & P & [\text{Assumption}] \\ \vdots & \vdots & \vdots \\ (k) & Q & [\text{Assumption}] \\ (k+1) & P \to Q & (i-k), \to \text{I} \end{array}$$

The input of →I does not consist of premises (reasons), but of a *derivation*. A derivation is not something that has propositional structure, so it does not function as a reason. Therefore application of →I does not yield *justification* for belief in its conclusion. But it can convey epistemic warrant nonetheless: it can produce *entitled* belief! So, in some way, →I is a curious logical rule. The reason why this difference between →I and, for instance, MPR, passes unobserved, is that wherever there is an entitlement to believe $P \to Q$ on the basis of →I, there is a justification of $P \to Q$ nearby. When a reasoner has gone through a logical derivation of the form

$$\begin{array}{cc} P & [\text{Assumption}] \\ \vdots & \vdots \\ Q & [\text{Assumption}], \end{array}$$

she *may* observe

"There is a logical derivation of Q from P",

and use this proposition as her reason for believing $P \to Q$. However, this should not obscure the fact that this proposition does not figure as a *premise* in an application of the logical rule →I.

Proof systems in which conditionalisation fails are characteristic for *partial logic*.[31] Suppose we assume the principles of Peano Arithmetic, not in the context of classical logic, but of partial logic, and where we assume that conditionalisation holds for arithmetical formulas, but not for all formulas of the whole language.[32] Call the system in which we are working S. Then full conditionalisation can be obtained by adding a following weak proof theoretic reflection rule, in the following way. Let $Der_S(x, y)$ be an arithmetical formula express that in S, from assumption x, the formula y can be derived. Clearly $Der_S(x, y)$ can be constructed in such a way that

$$\begin{array}{cc} P & [\text{Assumption}] \\ \vdots & \vdots \\ Q & [\text{Assumption}] \\ Der_S(\ulcorner P \urcorner, \ulcorner Q \urcorner) & \end{array}$$

is an admissible rule of S. (Since it contains a modicum of arithmetic, S has this much 'introspective' power.) Now consider the following weak reflection rule:

$$\frac{\vdash_S Der_S(\ulcorner P \urcorner, \ulcorner Q \urcorner)}{P \to Q}$$

In the context of classical logic, this rule is equivalent to the rule of local reflection.

[31] See [**Bla02**].

[32] See for instance [**Hor11**, Section 9.5].

Putting these two together yields full conditionalisation, and hence full classical logic. The moral of this is that classical logic may be seen as arising from reflection on a weaker logic.

2.9. Entitlement to reflection

Daphne's mother lives in a care home in the UK, far from where Daphne lives and works. Daphne's mother is in an advanced state of dementia. She does not recognise her daughter anymore; Daphne cannot communicate with her anymore. Daphne has no siblings. Aside from Daphne, there are no friends or relatives who visit Daphne's mother in the care home. Daphne herself has not been able to visit her mother for over six months because of travel restrictions connected to the coronavirus pandemic. Shortly after her last visit to the nursing home, Daphne's mother was assigned a new primary caregiver: nurse Myrtle. Since then, Myrtle has become Daphne's sole source of information about how her mother is doing. Daphne has been calling and skyping with Myrtle on a weekly basis, as she did with her mother's previous primary caregiver. Daphne has heard disturbing news reports of care homes where elderly residents are not well taken care of. But she has no evidence that this also holds for the nursing home where her mother is residing. In fact, she has no evidence whatsoever that her mother was not well treated over the past months. Daphne's life has been stressful since the outbreak of the pandemic. She has had to work from home, and take care of her children at the same time, because the schools are closed. For this reason, she has not been thinking about her mother as much as she normally does. Until today, Daphne has not asked herself the question whether Myrtle takes good care of her mother. Nonetheless, over the past months, Daphne has come to trust Myrtle. (As a matter of fact, Myrtle does take good care of Daphne's mother.) Today, after hanging up the phone after a conversation with Myrtle, Daphne realises that she has come to trust, in the fiducial sense of the word, that Myrtle treats her mother well. She reflects on this for a few moments. It would not be easy in the present circumstances, but she might acquire evidence concerning this matter. But she doesn't. She does not change her stance towards Myrtle in any way, and forms the propositional *belief* that Myrtle takes good care of her mother.

Daphne has formed her trust in Myrtle through her interaction with Myrtle, and at least in part in response to Myrtle's actions and behaviour. If Myrtle's behaviour in the skype conversations or her way of speaking in the phone conversations would somehow have aroused suspicion, then Daphne would not have come to trust her. Nonetheless, Daphne has not taken Myrtle's actions and behaviour as *evidence* for her trustworthiness: her reaction has been far more immediate and far less deliberate than that. According to Audi, "[fiducial] faith [that a friend will survive risky surgery] is not mainly a response to evidence (and need not be so at all)" [**Aud08**, p. 100]. I believe that this holds equally true of the closely related concept of fiducial trust. Moreover, Daphne has not formed her propositional belief that Myrtle treats her mother well by some form of *inference* from her trust in Myrtle. In this sense, the situation here is structurally different from the situation at the end of the previous Section, where we *derived* a conditional statement (something we believe) from a rule of inference (something we do). Viewing Daphne's transition from trust to belief as an inference, is a form of over-intellectualisation.

Daphne's initial fiducial trust, and her subsequent propositional belief, are *rational*. Daphne is *entitled* to her fiducial trust that Myrtle is kind to her mother. Moreover, she is also *entitled* to the propositional belief that she has formed through reflection on her trust in Myrtle: her rational fiducial trust is her warrant.

Does not Daphne's prior trust in Myrtle *already* constitute propositional belief? No. At the earlier time, she did not have a *disposition* to assent to the relevant proposition yet. How she would react if she were to realise that she had come to trust Myrtle, still hung in the balance. She *could* instead have reacted like Nathalie (see p. 62), by giving up or at least qualifying her trust.

As soon as Daphne has *realised* that she has come to trust that Myrtle takes good care of her mother, her epistemic position has irretrievably been altered. She has now lost her state of epistemic innocence. It is now not a rational option for her anymore to refrain from believing that Myrtle takes good care of her mother, while continuing, in an unconditional manner, fiducially to trust that Myrtle takes good care of her mother. Through reflecting on her trust, Daphne has incurred an epistemic obligation to harmonise her belief and her trust on the matter. This is a *new* epistemic duty: she did not have it before she reflected.

Daphne had no *epistemic duty* to reflect: not reflecting on her trust would have made her no less rational. Indeed, very small children cannot reflect in this way—they do not have the cognitive machinery yet,—and this does not make them irrational. Moreover, she had no *rational obligation* to react in the way that she did when she did reflect on her trust. She would have been no less rational if she had instead suspended her trust and sought evidence instead. This means that I am embracing a *liberal conception of rationality*. Van Fraassen canvasses this conception of rationality in the following way [**vF89**, p. 171–172]:

> The difference [between Russell's traditional conception of rationality and the 'liberal' conception of rationality] is analogous to that between (or so Justice Oliver Wendell Holmes wrote) the Prussian and the English conception of law. In the former, everything is forbidden which is not explicitly permitted, and in the latter, everything permitted that is not explicitly forbidden. When Russell is still preoccupied with reasons and justification, he heeds the call of what we might analogously call the Prussian concept of rationality: what is rational to believe is exactly what one is rationally compelled to believe. I would opt instead for the dual: what is rational to believe includes everything that one is not rationally compelled to disbelieve. *Rationality is only bridled irrationality.*

In other words, "rationality is a concept of permission rather than of compulsion" [**vF89**, p. 180].[33]

It would not be correct to describe Daphne's propositional belief as a voluntary act. This is so merely because, as we have seen earlier (see p. 60), forming a belief is *in general* not a voluntary act.

The warrant for the propositional belief that Daphne has formed is a posteriori rather than a priori. This is because it is an essential ingredient in her process of

[33]This does not mean that "anything goes": see [**vF89**, Chapter 7].

reflection that she remembers her worldly actions and her behaviour vis-à-vis Myrtle. But to remember this, she must have *observed* her own actions and behaviour first.

Suppose we grant that Daphne is entitled to the propositional belief that she has acquired through reflection on her trust. Does her belief constitute *knowledge*? This depends, I think, on the strength of her non-justificatory warrant for her prior *trust* that Myrtle treats her mother well. Non-justificatory warrant for trust, like justification (see p. 16), is a matter of degree.[34] If Daphne's entitlement to her trust is strong enough, then her propositional belief obtained by reflection can amount to knowledge. Whether that is the case or can be so in the imaginary scenario that I described, I am not sure.

Daphne has exercised her *rational faculty of reflecting on fiducial trust*. She has a prima facie entitlement to rely on this faculty in ways that are similar to her entitlement to rely on other rational competences, such as memory or perception. But unlike memory or interlocution (and *like* perception), this faculty is ampliative. Like other forms of entitlement, Daphne's entitlement to reflection is *prima facie*. If she had had counter-evidence to the proposition that Myrtle takes good care of her mother, then she would not have been entitled to form the propositional belief that she did form. Instead, she would have had an epistemic duty to look into the matter, i.e., to seek further evidence. Moreover, she would then not have been entitled to maintain her unqualified trust in Myrtle.

Our entitlement to reflection on fiducial trust flows, in the context of a liberal conception of rationality, from the *principle of rationality* of which I made use earlier:

> *Through reflection on our fiducial trust that p, we incur an epistemic obligation to align our fiducial trust that p with our belief that p.*

I see no way to motivate the above principle of rationality from any more fundamental principles. It is merely an appropriately qualified way of saying that our beliefs must cohere with our actions. As far as I can see, this is rock bottom.

I hasten to add once more, however, that reflection is not rationally bound to end in propositional belief. As we have seen earlier, it may just as rationally end in restricting or abandoning fiducial trust. Or one might decide to seek further information.

[34]This is a point that Burge does not seem to dwell on in his writings.

CHAPTER 3

Reflection

We will see that the concept of reflection has a long philosophical history, going back to the ancient Greeks. It is by no means the aim of this Chapter to give an exhaustive treatment of this philosophical history. Rather, we will only consider some episodes in it, and we will not consider any of these episodes in great depth. Philosophers from the early modern period—both rationalists and empiricists—will play an important role.

The concept of reflection has been put to work in various ways in the history of philosophy. But few philosophers have reflected on their own philosophical use of the concept of reflection, and almost no philosopher has worked out a systematic philosophical theory of reflection.

A brief comparison with the concept of *abstraction* may be instructive here. The concept of abstraction also has been put to philosophical use since the period of the ancient Greeks. In philosophy, the word 'abstraction' can refer to a process and to the product of a process. Abstraction plays a role not only in philosophy, but also in a technical way in mathematics (since the nineteenth century), when equivalence classes are taken and considered as mathematical objects in their own right. The use of abstraction in mathematics was famously described and philosophically exploited by Frege in his *Grundlagen der Arithmetik* [**Fre84**]. Frege noticed that principles describing how products of abstraction are related to what they are abstracted from—for instance, *Hume's Principle*—have proof theoretic strength. In this sense, abstraction in Frege's sense is a knowledge-producing mechanism.

Like the word 'abstraction', the word 'reflection' also refers not only to one or more *processes*,[1] but also to the products of such processes. Like abstraction, reflection is also a concept that plays a role not only in philosophy, but also in mathematics.[2] One important use of the word 'reflection' in philosophy has to do with "the mind bending back on itself" and with *introspection*, as we will see below. In mathematics, reflection has to do, among other things, with the process of *axiomatisation*, where mathematicians "bend back" on their own proof practice, in a given discipline. Gödel observed that principles that certain principles concerning such reflections (called *proof theoretic reflection principles*) have proof theoretic strength.

In recent decades, great efforts in the philosophy of mathematics have been made to understand abstraction. Despite the work of Gödel and others, it is fair to say that we do not understand philosophical or mathematical *processes* of reflection well at all. There has been no "Frege for reflection". This is in part because, as we

[1] We will see later that in one of its philosophical meanings, the word 'reflection' refers not to a process but simply to a relation.

[2] In fact, reflection also plays a role in physics, although I will not have much to say about physical notions of reflection in this book.

will see, 'reflection' is a word that can be used to express several different concepts, whereas 'abstraction' is not.

3.1. The many faces of reflection

It is generally bad form to appeal to a dictionary in philosophical discussions. Nonetheless, in the present case, I think that it is instructive to start our investigation with a close inspection of a dictionary entry.

The *Oxford English Dictionary* (OED) describes the various meanings of the word 'reflection' thus [**LCF$^+$73**, p. 1777]:

> **Reflection**, **reflexion** [...]
> **1.** The action, on the part of surfaces, of throwing back light or heat (rays, beams, etc.) falling upon them. The phenomenon of light and heat being thrown back in this way. **b.** Reflected light or heat [...] **2.** The action of a mirror or other polished surface in exhibiting or reproducing the image of an object; the fact or phenomenon of an image being produced in this way [...] **b.** An image or counterpart thus produced [...] **3.** The act of bending, turning, or folding back [...] **4.** The act of throwing back, or fact of being thrown or driven back, after impact [...] **b.** *Phys.* Reflex action [...] **5.** Animadversion, blame, censure, reproof [...] **b.** A remark or statement reflecting on a person [...] **c**. An imputation; a fact or procedure casting an imputation or discredit *on* one [...] **6.** The act of turning (back) or fixing the thoughts on some subject; meditation, deep or serious consideration [...] †**b**. Recollection or remembrance *of* a thing [...] **c.** *Philos.* The mode, operation, or faculty by means of which the mind has knowledge of itself and its operations, or by which it deals with the ideas received from sensation and perception [...] **7.** A thought or idea occurring to, or occupying, the mind [...] **b.** A thought expressed in words; a remark made after reflection on a subject [...]

The dictionary entry shows that the meaning of the word 'reflection' has a core. It is often used to refer to a *process* that consists of a *source*, a transformation procedure, and a *product*. But the dictionary entry also shows that these basic components of the meaning of the word can be filled out in very different ways. Indeed, it is clear that the word 'reflection' has a rather diffuse meaning. Although its multiple meanings are clearly related, the word is certainly not uniformly applied with the same meaning. Some meanings of the word 'reflection' will not play a significant role in what follows. For instance, we will not be much concerned with meanings **4.** or **5.**. Other meanings of the term will be of significance to us. Meanings **2b**, and **6c.**, for instance, will play an important role in this Chapter and in some later Chapters. Let us call these meanings *type* **2** *reflection* and *type* **6** *reflection*, respectively. Meaning **3.**, which we call type **3** reflection, will also be seen to play a role in the beginning of the philosophical history of reflection.

The concept of reflection has a long history in Western culture: in literature, physics, poetry, theology, psychology, the arts... We will in this Chapter be concerned with the history of reflection *in philosophy*, and in later Chapters also in *the*

mathematical sciences (and especially in logic) in the twentieth century. The history of reflection goes back at least to Plato. Philosophers have made determinate attempts to get reflection to do metaphysical and epistemological work for close to 2500 years. Many avenues have been pursued doggedly over many centuries. One factor that makes this history difficult to trace is that, as I intimated above, reflection has not been a real *theme* in philosophy in the way that, for instance, the concept of moral value has been since Antiquity. Attempts were made to make the concept of reflection do philosophical work, i.e., to *use* it, but in a somewhat oblique, mostly not fully conscious manner.[3] What has been lacking in the history of philosophy, one might say, is systematic and explicit *philosophical reflection on the philosophical concept(s) of reflection.* This I propose to do, from a historical perspective, in the present Chapter. So the following sections will to a large extent be an exercise in *metaphilosophy.*

We will discuss some elements of the history of the concept of reflection in philosophy. This history is intricate and complicated: far be it from me to claim that the historical overview that is given here even touches on all of the main strands of this history. My interest in the concept of reflection is not primarily as a chapter in the history of ideas. Instead, my leading question will be thematic in nature: *which philosophically fruitful uses can be made of concepts of reflection?* Later in the book,[4] one such theoretically fruitful use of reflection (in logic and the foundations of mathematics), namely type **6** reflection, will be singled out for special attention.

Philosophical uses of reflection are often metaphorical or even allegorical in nature. Think for instance of the famous allegory of the cave in Plato's *Republic*, where sensible objects are somehow likened to shadows (read: reflections of outlines of) of puppets that are projected on a wall. The arduous task of philosophers—especially of *analytic* philosophers—is then to extract *clear* and *literal* metaphysical or epistemological content from such metaphors and allegories, and to probe the philosophical merits of this content. Sometimes there seems reason for optimism; at other times the philosopher is driven to despair.

The history of reflection in philosophy is indeed a story of hit and miss. Descartes' *Cogito* is widely regarded as a success story, whereas the idea that there is a deep parallelism between the cosmos ('Macrocosmos') and Man ('Microcosmos') is commonly seen as a mistake of monumental proportions.[5] Some roads are regarded with much suspicion today, without being quite dead yet. Neoplatonist 'emanation' metaphysics, for instance, is felt by many contemporary philosophers to be dangerously close to discredited Microcosmos/Macrocosmos hypotheses. Yet it is not always easy correctly to gauge the prospects of philosophical theories of reflection. We will see that in the late Middle Ages, Philo of Alexandria's theory of reflection came to be regarded as obsolete. I consider this judgement to be premature. I will argue, against this judgement, that Philo's philosophical theory of reflection was almost two millennia ahead of his time, and that it is one of the most powerful theoretical ideas that have been developed in the history of Western thought.

---

[3]The use of the word 'reflection' or 'reflectio' as a term for the referring to the soul's knowledge of itself or of its own acts goes back only to the late middle ages, it seems: see [**Men12**, p. 65].

[4]See Chapter 8.

[5]For an eloquent expression of this latter viewpoint, see [**Boa80**].

3.2. Echo and the pool

Our historical exploration of the concept of reflection starts with the Greek myth of Narcissus. The most extensive version of (a version of) the myth that we have is found in Ovid's *Metamorphoses* [**Gra92**, p. 286–288]:

> NARCISSUS was a Thespian, the son of the blue Nymph Leiriope, whom the River-god Cephisus had once encircled with the windings of his streams, and ravished. The seer Tereisias told Leirope, the first person ever to consult him: 'Narcissus will live to a ripe old age, provided that he never knows himself.' Anyone might excusably have fallen in love with Narcissus, even as a child, and when he reached the age of sixteen, his path was strewn with heartlessly rejected lovers of both sexes; for he had a stubborn pride in his own beauty.
>
> Among these lovers was the nymph Echo, who could no longer use her voice, except in foolish repetition of another's shout: a punishment for having kept Hera entertained with long stories while Zeus's concubines, the mountain nymphs, ever evaded her jealous eye and made good their escape. One day when Narcissus went out to net stags, Echo stealthily followed him through the pathless forest, longing to address him, but unable to speak first. At last Narcissus, finding that he had strayed from his companions, shouted: 'Is anyone here?'
>
> 'Here!' Echo answered, which surprised Narcissus, since no one was in sight.
>
> 'Come!'
>
> 'Come!'
>
> 'Why do you avoid me?'
>
> 'Why do you avoid me?'
>
> 'Let us come together here!'
>
> 'Let us come together here!' repeated Echo, and joyfully rushed from her hiding place to embrace Narcissus. Yet he shook her off roughly, and ran away. 'I will die before you ever lie with me!' he cried.
>
> 'Lie with me!' Echo pleaded.
>
> But Narcissus had gone, and she spent the rest of her life in lonely glens, pining away for love and mortification, until only her voice remained.
>
> (*)
>
> One day, Narcissus sent a sword to Ameinius, his most insistent suitor, after whom the river Ameinius is named; it is a tributary of the river Helisson, which flows into the Alpheius. Ameinius killed himself on Narcissus's threshold, calling on the gods to avenge his death.
>
> Artemis heard the plea, and made Narcissus fall in love, though denying him love's consummation. At Donacon in Thespia he came upon a spring, clear as silver, and never yet disturbed by cattle, birds, wild beasts, or even by branches dropping off the trees that

> shaded it; and as he cast himself down, exhausted, on the grassy verge to slake his thirst, he fell in love with his reflection. At first he tried to embrace and kiss the beautiful boy who confronted him, but presently recognised himself, and lay gazing enraptured into the pool, hour after hour. How could he endure both to possess and yet not to possess? Grief was destroying him, yet he rejoiced in his torments; knowing at least that his other self would remain true to him, whatever happened.
>
> Echo, although she had not forgiven Narcissus, grieved with him; she sympathetically echoed 'Alas! Alas!' as he plunged a dagger in his breast, and also the final 'Ah, youth, beloved in vain, farewell!' as he expired. His blood soaked the earth, and up sprang the white narcissus flower with its red corollary, from which an unguent balm is now distilled at Chaeronea. This is recommended for affections of the ears (though apt to give headaches), and as a vulnerary, and for the cure of frost-bite.

This text consists of two very distinct parts: the part up to (*), and the part from (*) onwards. We will see how these two parts foreshadow two very different ways in which reflection has played a role in the history of philosophy.

Both parts of the myth of Narcissus are clearly related to meanings **2** and **2b**—"mirroring"—of the term 'reflection'. The visual reflection of Narcissus in the pool is of course an essential component of the the part from (*) onwards. But an echo is merely the auditive counterpart of a visual reflection, so meaning **2** is of central importance in the part up to (*) also. Meaning **3** in the dictionary entry—"bending back"—also plays an important role. It is essential to the story that echo of Narcissus' voice is "thrown back" at him, and that Narcissus sees his own reflection in the pool. However, meaning **6c** of the word reflection does not seem to be present in the myth.

The imperfection of the image as compared to the original plays no role. On the contrary, the situation is so set up that the reflection appears to be *perfect* ("a spring, clear as silver,…"). An essential *motif*, however, is the unbridgeable distance between the reflection of Narcissus and Narcissus himself. Indeed, because the reflection appears as perfect as Narcissus himself, and Narcissus is irresistibly attracted to his reflection and not conversely, the *direction* of the reflection relation becomes somehow unclear.

The part up to (*) is equally rich and fascinating, if not more so. One theme here is the concept of *repetition*: Echo can only repeat what others say. One observation in the text is that repetition seems by its very nature unproductive and unoriginal. The text explores, with typical ancient Greek playfulness, whether by repetition it might nonetheless be possible for fundamentally *new* content to be expressed. The suggestion is that this can perhaps be achieved using indexicals ("here", "now"). By means of indexicals, Echo attempts to make an assertion that is about herself and the place and time where she finds herself. In this sense, meaning **3** in the dictionary entry ("bending back") plays a key role. It is not clear whether Echo's attempt is quite successful. The apparently new content—"come (to me)", as uttered by Echo, for instance—can be seen as the result of a *misunderstanding* on the part of Narcissus, namely that Echo makes an *assertion* rather than merely repeat without assertive force. But another interpretation is also possible. Perhaps

Hera was not *completely successful* disabling Echo from speaking in her own name. Perhaps she only succeeded in severely restricting the manner in which Echo was able to do so. Deception is a common theme in Greek mythology, and it is a theme in the myth of Narcissus. Just as Zeus found ways to deceive Hera, perhaps Echo succeeded, at least to some extent, in eluding Hera's curse?

The myth of Narcissus articulates pre-theoretical understanding of reflection. We have seen how meaning **3** ("bending back") plays a role in both forms of reflection that are described in the myth of Narcissus, but that apart from that, they are quite different from each other. In the remainder of this Chapter, we will look into the way in which these two forms of reflection have migrated into and evolved through the course of the history of philosophy. In later Chapters, we will be concerned with the way in which the first of these forms of reflection has come to play a role in logic in the twentieth century. Of this, let me now give a short preview.

Fast forward more than two millennia, and transport the context from intimate relationships to mathematical theorising. Then assertion becomes proof, and fundamentally new content becomes proof-theoretically independent statements. A standard arithmetical provability predicate Bew for a mathematical theory S (extending a weak theory of arithmetic) is a *repetition machine*:

$$S \vdash \mathrm{Bew}_S(A) \Leftrightarrow \mathrm{Bew}_S(A) \Leftrightarrow S \vdash A.$$

In this context "genuinely new statements" become statements that are independent of S, and assertion becomes provability in S. Gödel's 1931 insight consists in the fact that, using indexicals and the repetition machine, a statement that "bends back" upon itself and that is genuinely new (compared to what has been asserted, namely, S) can be produced. This is of course the Gödel sentence for S. It would be quite a stretch to claim that the first incompleteness theorem is foreshadowed in the myth of Narcissus. Not even all key ingredients of Gödelian incompleteness arguments are present in the part of the myth up to (*). *Negation*, for instance plays a crucial role in the construction of the Gödel sentence, but it does not even figure in Echo's statements (or repetitions). Nonetheless, the first part of the myth of Narcissus was a first step on the road to exploring how repetition and indexicals interact with each other.

3.3. Philo's angel

Let us now turn to the question how the two philosophical senses of reflection that are prefigured in the myth of Narcissus entered Western philosophical thought. In particular, we will describe how the dictionary meaning **2** of the word 'reflection' found its way to philosophy through Philo of Alexandria, and how dictionary meaning **6** found its way to philosophy through the later Neo-platonist philosophers.

Philo of Alexandria was a Jewish bible commentator, who lived in Alexandria in the first century AD. His aim was to reconcile Greek philosophy with the truth of Scripture (the Pentateuch). We will not go into the details of the complicated metaphysical system that Philo constructed, but limit our discussion to elements of it that are directly relevant to our present concerns.

As a philosopher, Philo took a Platonist stance; indeed, he can and has been seen as a precursor of Neo-platonism. The platonistic concept of *idea* therefore plays a key role in his metaphysical thought. He frequently uses the concept 'idea' in the *structural* sense in which Plato and Aristotle use it, namely in the sense of

pattern. However, he is also the first philosopher to this concept in a radically new sense, namely in the sense of *image*.[6] (This is probably due or at least related to the fact that Scripture describes Man as being created after the image of God.)

Philo's starting point is the thesis, which is also an original thought of Philo, that God is radically transcendent. The gulf between God and Man is so vast that no knowledge of God and His intentions with the world is possible for us without mediation. For Philo, this mediation is provided by the *externalised Logos*, which can be seen as an abstract image of the mind of God.[7] This externalised Logos is a created (abstract) entity that we humans can to some extent understand. It exists outside the essence of God—which infinitely exceeds our powers of comprehension,—but we humans cannot distinguish it from the essence of God. Philo explains this thought in the following passage in his, *On Dreams*, which is quoted in [**Seg77**, p. 163]:

> Thus in another place, when he had inquired whether He that is has a proper name, he came to know full well that He has no proper name, [the reference is to Exodus 6:3] and that whatever name anyone may use for Him he will use by licence of language; for it is not in the nature of Him that is to be spoken of, but simply to be. Testimony to this is also afforded by the divine response made to Moses' question whether He has a name, even "I am He that is (Exodus 3:14)". It is given in order that, since there are not in God things that man can comprehend, man may recognise His substance. To the souls indeed which are incorporeal and occupied in His worship it is likely that He should reveal himself as He is, conversing with him as a friend with friends; but to souls which are still in the body, giving Himself the likeness of angels, not altering His own nature, for He is unchangeable, but conveying to those who receive the impression of His presence a semblance in a different form, such that they take the image not to be a copy, but that original form itself.

In other words, an 'angel' *reflects* the essence of God in the form of an image. In this way, dictionary meaning **2** of the word 'reflection' is at work here. But this angel-image is such a perfect copy that we cannot distinguish it from God in any way, so we humans tend to take such 'angels' to be God himself. Philo also observes—he was clever indeed!—that this theory leads to a semantic problem. Since we cannot distinguish God from certain 'angels', there is nothing we can do to ensure that the word 'God' refers to God rather than to one of the angels. So, literally speaking, on Philo's view, God is unnameable.

Philo thus postulates a *reflection principle*: God is reflected, in the sense of mirroring (type **2** reflection), in an entity in the world (an angel). We will see later[8] that reflection principles of this sort are theoretically very powerful: complexity of the reflected object can be deduced from such principles.

But neither Philo nor his philosophical or theological readers in Antiquity seem to have been aware of the theoretical strength of such kinds of reflection thoughts.

---

[6] See [**Wol47a**, p. 238]. Plato uses the term 'image' only to refer to the visible world.

[7] *Noûs* or *Logos* is one of the most complicated concepts in Greek philosophy. For the role of the concept of Logos in Philo's metaphysics, see [**Wol47a**, Chapter 4]

[8] See Section 6.5.

Why was it so hard? It has to do with Philo's conception of God. Like many ancient and medieval philosophers, Philo takes *absolute simplicity* to be one of the cardinal properties of God [**Wol47b**, chapter XI, section I], and this muddles the theoretical picture. If God is absolutely simple, then the reflecting object, if it reflects perfectly, will also be simple. Philo's mirroring principle comes into its own when God is taken to be *infinite*, where infinity is understood in something like the contemporary sense of the word, and where the reflecting object is in some sense part of the reflected object. But the concept of infinity was not theoretically understood in Philo's days. Indeed, only the beginnings of *any* conception of God as infinite can be discerned in Philo's work.[9]

Nonetheless, most elements required for theoretically exploiting his reflection principle were available to Philo. According to Philo's metaphysical system, God created an abstract blueprint of the world, the "externalised" Logos of God [**Wol47a**, chapter IV, section IV]. This externalised Logos is complex indeed, and it contains abstract counterparts of the angels, which are the reflecting objects. If this externalised Logos is reflected in an angel that is indistinguishable from God, then the world *must* be infinite in the modern sense of the word. It is only because he did not possess the modern concept of infinity that this line of reasoning was not available to Philo.

At any rate, Philo postulates a form of reflection that has a *direction*: from God, to the world. A few centuries later, Augustin postulates a form of reflection with an inverse directionality: from the world (which includes 'our' natural numbers) to the mind of God. In Augustine's writings, an explicit connection with the concept of infinity is made.

Augustine's views of infinity did not remain stable throughout his theological career.[10] But his writings contain a thought concerning quantitative infinity that has proved to be remarkably prescient. Augustin's thought concerns the multiplicity of the natural numbers. This multiplicity forms a potential infinity in the world. But in God's knowledge this multiplicity is limited in the sense that He can somehow assign a number to it, so that something that is infinite for us, is finite for God [**Aug**, 12.18]:

> The infinity of number[s], although there is no number for infinities of numbers, is yet not incomprehensible by Him of whose understanding there is no number. And thus, if what is comprehended in knowledge is made finite by the comprehension of this knowledge, then all infinity is in some ineffable way finite to God, for it is not incomprehensible to His knowledge.

Thus the infinity of natural numbers is somehow *reflected* in a *bounded* entity in God's thought. In *Mitteilungen zur Lehre des Transfiniten I*, this passage is lauded by Cantor as a prefiguration of his theory of transfinite numbers [**Can32**, p. 402]:[11]

> Energischer als dies hier von S. Augustin geschieht, kann das Transfinitum nicht verlangt, vollkommener nicht begründet und verteidigt werden. [...] Indem nun h. Augustinus die totale, intuitive Perzeption der Menge der natürlichen Zahlen [durch das

[9] See [**Gel05**, Section 2].

[10] See [**Dro19**].

[11] For more on the influence on Augustine's philosophy of mathematics on Cantor, see [**vdVH13**].

> Wissen Gottes] behauptet, erkennt er zugleich diese Menge *formaliter* als ein aktual-unendliches Ganzes, als ein *Transfinitum* an, und wir sind gezwungen, ihm darin zu folgen.

We will return to Cantor's views in infinity and reflection in a later Chapter.[12]

The idea that there is a need for mediation between a transcendent God and the world already plays a role in certain microcosmos / macrocosmos-theories that were advocated from late Antiquity onwards, particularly in forms of Neo-platonism and in early medieval philosophy.[13] In many of these theories, Man was taken to play this mediating role. This is because Man belongs in part to the heavenly world (because she has a soul), and in part to the material world (because she has a body); moreover, the two are intimately intertwined [**All44**, p. 355]. In this sense, Man is a microcosmos: she is a small-scale copy of the structure of the whole cosmos. Over the centuries, this idea was developed in myriads of ways. For instance, theories were formulated that argued that the state is a macrocosmos compared to Man.

Microcosmos / macrocosmos ideas obtained a new lease on life in Renaissance philosophy, and the concept of type **2** reflection plays an important role in Leibniz' *Monadology*.[14] Yet, from the end of the middle ages onwards, the influence in Western metaphysics of this concept of reflection gradually decreases. An awareness of this evolution is expressed as early as the late 13th century AD by Odo Reginaldus (as quoted in [**Côt02**, p. 78]):

> How can the finite attain [knowledge of] the Infinite? On this question some say that God will show Himself to us in a mediated way, and that he will show Himself to us not in his essence, but in created things. This view is receding from the aula...

Beside the anthropomorphism at the core of many of the microcosmos / macrocosmos theories, a fundamental problem with this 'research programme' was a poor understanding and theoretically unconstrained use of the correspondence relation of reflection. The correspondence relation between microcosmos and macrocosmos was variously interpreted as some form of similarity, analogy, mirroring, metaphysical causality, symbolising, or a combination of those. This made the epistemic strength of the inferences from properties of the microcosmos to properties of the macrocosmos and *vice versa* difficult to evaluate.

Philo, as we saw, already worked with a clearer notion of similarity: *absolute indiscernibility*. But really precise notions of similarity that can do useful work in this context are of a mathematical nature, and would not become available until much later. One cluster of relevant notions consists of mathematical concepts of *structural* similarity, such as bijection, isomorphism, homomorphism, and the like. Another cluster of relevant notions, as we will see in a later Chapter,[15] comprises relations of *elementary equivalence*, i.e., the relation of making the same sentences true.

---

[12] See Section 6.5.

[13] See [**All44**].

[14] Cfr infra, Section 3.7.

[15] See Section 6.5.

3.4. See the flying man

The historical path to type **6** reflection is difficult to reconstruct, and is at present very imperfectly understood. In our discussion of it, we to a considerable extent follow [**Men12**].

According to Menn, the origin of this concept of reflection can be traced to a dissatisfaction, on the part of Aristotle, with Plato's cosmological theory. Plato postulated the existence of a *World Soul*. In order to account for the movement of the heavens, Plato postulated a spiritual revolving ball coextensive with the material cosmos. This spiritual revolving ball is held responsible for the movements of the stars, and individual revolving spiritual spheres responsible for the movements of the planets. Already Aristotle in *De Anima* complained that by conceiving of the World Soul as physically revolving, Plato commits a 'materialistic fallacy' ([**Ari**, 406b24–407a2], as quoted in [**Men12**, p. 47]:

> The soul does not seem to move the animal in this [sc. merely mechanical] way, but through some choice and thinking [sc. and therefore teleologically]. And [the?] Timaeus too physicizes that the soul moves the body in the same way [as Democritus holds], [that is] that through being moved itself it moves the body too, since it is interwoven with it. For after it has been constituted out of the elements [that is, being, sameness, otherness] and divided according to harmonic numbers, so that it might have a connate sensation of harmony and so that the universe might be locally moved with concordant locomotions, he *bent back the straight line* [my emphasis] into a circle; and having divided one circle into two attached at two [opposite points], he then divided one of them into seven circles, as if the motions of the soul were the locomotions of the heaven.

In this passage, we see also that type **3** reflection, namely reflection in the sense of "bending, turning, or folding back" is at playing a key role.

The Neo-platonists shared Aristotle's misgivings, and argued that a *Vergeistlichung* of Plato's theory is urgently called for. They arrived at the view that the World Soul does not physically move, but continually thinks and *thereby* moves the heavens. This is of course a somewhat anthropomorphic way of thinking. According to a very natural pre-theoretic way of thinking, mental events (beliefs and desires) *cause* physical events such as limbs being moved and eyebrows being raised. In a structurally somewhat similar way, the Neo-platonists believed the World Soul to move the heavens.

In this way, physical motion is transformed in mental directedness. Moreover, the Neo-platonists interpreted Plato's 'bending back of the straight line' as the divine souls thinking *themselves* and thereby understanding the world (Proclus, *In Timaeum* 2.248, 11–23, as quoted in [**Men12**, p.64–65]):

> [. . .] since the vital aspect of the soul is intellectual and directed toward returning and unwinds the intelligible multiplicity, it returns again to the same [starting-point]; and since the soul moves other-moved things [while? by?] being turned toward itself and moving itself; for all these reasons the circular [shape or motion] is appropriate to it.

Thus the idea of *autonomous self-reflection* is born [**Men12**, p. 59–60]:

> The late Neoplatonists are thus led [...] to develop and defend the *Phaedrus* argument by saying that the only genuine primitive self-motion (that is, a self-motion that is not decomposable into one part of a thing moving another part) is self-*thinking* (more precisely, not just any act in which something thinks itself under some description, but what we might call *reflexive* thinking—thinking whose content is essentially *de se*) [...]

Already in late Antiquity, especially in Porphyry and in Augustine, one finds the thought that not only the *World Soul* and other divine beings self-reflect in the way described by Proclus.[16] Human souls, too, have privileged intellectual access to themselves as souls. They, too, can be *present to themselves* in a way that gives them knowledge of their essence and existence as a thinking being.

Porphyry moreover argued that the knowledge thus obtained is knowledge of the purest and highest kind. When the subject and the object of knowledge differ from each other, perfect knowledge is not possible, because the knowledge will then necessarily be tainted by subjective elements that are extraneous to the object of knowledge. But self-reflective knowledge does not suffer from this defect, because in this form of knowledge the subject and the object of knowledge coincide:[17]

> Those people are present to themselves who are able to go intellectually to their own essence and know their essence and in that knowledge and in the recognition of that knowledge *to grasp themselves in the unity of knower and known* [my emphasis]...

This argument does not convince contemporary philosophers. If the mystic oneness with oneself that seems to be described here is no more than pure self-identity, then it is not clear how it qualifies as a form of knowledge at all.

Avicenna later likewise defended the thesis that humans can come to know their own essence and existence by self-reflection. He argued for this by means of a striking *thought experiment argument* known as the argument of the Flying Man:[18]

> The inquiry leads us to concern ourselves with grasping the quiddity of this thing that is called 'soul'. We must here indicate a way to affirm the existence of our soul, with an affirmation that may serve as an admonition and reminder. This will be a pertinent indication for one who has the ability to observe the truth by himself, without needing to be instructed or rebuked, or averted from errors. We say that one of us must imagine himself as if he were created all at once and as a whole, but with his sight covered so that he cannot see anything external, and created falling through the air or a vacuum, but falling in such a way that he encounters no air resistance nor anything else that would allow him to have any sensations, and with his limbs separated from one another so that they do not meet or touch. Then consider whether he will affirm the existence of his

[16] See [**Sor07**, p. 61–64].

[17] From Porphyry, Sententiae 41, as quoted in [**Sor07**, p. 62].

[18] Avicenna, The healing: Soul 1.1 (Rahman 15.18–16.17) (= De Anima 1.1, from Latin of Avicenna Latinus, ed. S. Van Riet), quoted in [**Sor07**, p. 65–66].

> essence. For he will not have any doubt in affirming existence for his essence, yet he will not along with this affirm [the existence of] the extremities of his limbs, nor his innards, his heart, his brain, or anything external to him. Instead, he will affirm [the existence of] his essence, without affirming that it has length, breadth or depth. Nor, if in that state he were able to imagine there to be a hand or other body part, would that it was a part of himself. You know that what is affirmed is different from what is not affirmed, and that what is grasped immediately [literally: 'what is near at hand'] is different from what is not so grasped. Therefore the essence whose existence is affirmed [by the Flying Man] is proper to him, insofar as it is his self, not his body or his limbs, which he does not affirm. Thus he is admonished and has a way of being awake to the existence of his soul as something distinct from the body and immaterial, and he knows and is aware of it [sc. his soul]. But if he is oblivious of it, he will need to be rebuked.

Avicenna expresses the conclusion of his thought experiment as follows:[19]

> What [the Flying Man] will then have grasped is his essence, which he will then perceive. Indeed there is nothing which grasps a thing without grasping its [own] essence as grasping.

In these late ancient and medieval descriptions of the mind that goes out of itself and comes back to itself in a reflexive movement, we may descry the origin of the Cartesian thought that some form of rational introspection lies at the basis of an understanding of the world.

3.5. Cartesian thoughts

Every philosopher is familiar with the overall structure of the argumentation in the *Meditations* [**Des41**]. There Descartes starts by methodically suspending all his beliefs on the strength of the possibility that he is deceived by an evil demon. He then realises that there is at least one belief that survives this process of methodological doubt. This is the famous *Cogito*-belief, i.e., the content of the statement:

I am thinking.

From this, he obtains the insight that he is a Thinking Subject. Moreover, he finds in his mind an idea of God. He uses this idea to construct a proof of the existence of God: from the existence in his mind of an idea of God, he "deduces" that a benevolent God exists. From the existence of a benevolent God, he then concludes that most of his former beliefs about the external world must be correct after all.

There is a wide consensus that Descartes' proof of the existence of God is deeply problematic and probably unsalvageable: this is seen as one main reason why his foundational programme is ultimately doomed to failure. Nonetheless, at the same time Descartes is seen as one of the most important philosophers that ever lived. This is in part because the Cartesian conception of the Subject, as articulated in the *Meditations*, has proved to be extremely influential in philosophy

[19] Avicenna, Reply to Bahmanyâr and al-Kirmânî, paras 58–59, translated by J. Michot, as quoted in [**Sor07**, p. 66].

and far beyond. But it is also because there is a strong feeling that Descartes was onto something in his reflections on the epistemic significance of the *Cogito*.[20] Let us to some extent pursue this latter thought with some more rigour and precision than Descartes himself was able to.

Let the operator A_i stand for "it can be a priori known by i that", and let the operator $\Box$ stand for "it is metaphysically necessary that". Let c stand for the proposition expressed by the *Cogito* sentence "I am thinking", and let e stand for the proposition "I exist". Furthemore, let c_i stand for the content of c relative to subject i (so with 'I' in the *Cogito* sentence interpreted as subject i), and similarly for e_i.

PROPOSITION 3.1. $A_i(c_i)$.

PROOF. The following is a procedure for coming to know c:
(1) *Think* proposition c.
(2) By introspection, come to believe c.
(3) By introspection again, form the second-order belief that you believe c.
(4) Claim: This second-order belief is knowledge.
The reason for (4) is the following. By (2), the second-order belief is true. Since introspection yields justification, by (3) the second-order belief is also justified. Moreover, no Gettier-condition is present.
(5) This knowledge is a priori, since introspection is a source of a priori belief. □

From Proposition 3.1, we can go on in Cartesian fashion to infer:[21]

PROPOSITION 3.2. $A_i(c_i \wedge e_i)$.

PROOF.

(1) $A_i(c_i)$	Proposition 3.1
(2) $A_i(c_i \rightarrow e_i)$	existential generalisation on c_i
(3) $A_i(e_i)$	from (1) and (2)
(4) $A_i(c_i \wedge e_i)$	from (1) and (3)

□

Descartes then very quickly infers from this to the conclusion "I am a thinking substance"; but this goes beyond the content of proposition 3.2. However, Kripke observed that we are now in a position to conclude that the concepts of necessity and of a prioricity are not co-extensional [**Kri80**]:

PROPOSITION 3.3. *There are contingent a priori propositions.*

PROOF. This claim is witnessed by c_i and e_i, in a situation where i is thinking (and therefore exists). These propositions are a priori knowable (proposition 3.2). But since i could have failed to exist, and would in such a situation not have thought, these propositions are contingent. □

Bernard Williams also recognises that the propositions c and e are of special epistemological significance. He argues that they have a property that he calls *incorrigibility* [**Wil78**, p. 59]:

[20] See for instance [**Wil78**, Chapter 3].
[21] *Pace* the reservations formulated by Bernard Williams at the end of Chapter 3 in [**Wil78**].

Definition 3.4 (B. Williams).
A proposition x is *incorrigible* if and only if for all i:

$$\Box[(i \text{ believes that } x) \to x].$$

Proposition 3.5 (B. Williams).
Propositions c and e are incorrigible.

Proof. Consider any situation w in which i believes c. Since believing is a form of thinking, i is thinking in w. So, in w, proposition c is true. And since one cannot believe without existing, also proposition e is true in w. □

Next, suppose that in the *Cogito* sentence we replace "subject i thinks" by "formal system S proves", where system S contains a minimal amount of arithmetic. Call the resulting proposition c^*. In proof theory, such a proposition is called a *Henkin sentence*.[22] Then c^* has a property analogous to the incorrigibility of c:

Necessarily, if c^* is provable in S, then it is true.

Moreover, Löb has shown that it is consequence of Gödel's second incompleteness theorem that c^* is indeed provable in S (and therefore also true) [**BBPJ02**, Corollary 18.5].

Now consider the cartesian proposition $\overline{c}$, which is the content of the sentence:

I am *not* thinking this thought.

This proposition also belongs to a noteworthy class of propositions, namely the *elusive* ones:

Definition 3.6.
A proposition x is *elusive* if and only if for all i:

$$\Box[((i \text{ thinks } x) \to \neg x) \wedge ((i \text{ does not think } x) \to x)].$$

Proposition 3.7. *$\overline{c}$ is elusive.*

Proof. This follows from the self-referent properties of $\overline{c}$. □

In this sense, the sentence $\overline{c}$ is analogous to the Gödel sentence g_S for a system S, for which we have, by Gödel's first incompleteness theorem, that if S proves g_S, then g_S is false, and if S does not prove g_S, then g_S is true.

Proposition 3.7 thus shows that it is in principle impossible, for a mind, to entertain all and only the true propositions in thought. This it can be seen as a weak incompleteness theorem: as a *cartesian incompleteness theorem*. Indeed, Descartes had all the conceptual and theoretical tools that are required for arriving at this insight. As with the myth of Narcissus (up to (*)), Descartes merely failed to consider *Cogito*-like self-referential propositions that essentially involve the concept of *negation*.

[22] See [**BBPJ02**, p. 236]. If we substitute "subject i thinks" instead by "it is true that", then we obtain the proposition that is known in the truth theory literature as the *truth teller*.

3.6. Lockean reflection

In the foregoing Sections, we have seen how concepts of reflection have been taken to task for doing some heavy philosophical lifting in the history of philosophy. But to the best of my knowledge it is fair to say that no philosopher prior to Locke developed a worked-out theory about what reflection *is*: Locke was the first philosopher to do so. Moreover, even though his theory of reflection, too, is rudimentary in certain respects, it is developed in more detail than even his successors' theories of reflection.

Locke's theory of reflection went on deeply to influence later philosophical thought about reflection. It is not an exaggeration to say that, despite the fact that his successors took issue with elements of his theory of reflection, Locke's theory quickly became dominant in philosophy, and continues in analytic philosophy to be dominant up to this day.

Locke's theory of reflection is a part of his philosophy of mind and cognition. At a first approximation Locke can be taken to liken the human mind to a kind of vessel that contains *ideas*. He does not have very much to say about the nature of ideas, except that they are *representations*. According to Locke, there are only two sources of ideas in the human Mind: *sensation* and *reflection*.

Locke describes the process of reflection in general terms as follows [**Loc89**, II, 1, 4]:

> By REFLECTION then, in the following part of this Discourse, I would be understood to mean, that notice which the Mind takes of its own Operations, and the manner of them, by reason whereof, there come to be *Ideas* of these Operations in the Understanding.

Thus sensation may be called a faculty of external sense, and reflection may be called a faculty of inner sense [**Loc89**, II, 1, 4]:

> The other Fountain, from which Experience furnisheth the Understanding with Ideas, is the *Perception of the Operations of our own Minds* within us, as it is employ'd about the *Ideas* it has got [...] This source of *Ideas*, every Man has wholly in himself: And though it be not Sense, as having nothing to do with external Objects; yet it is very like it, and might properly enough be call'd internal Sense.

Whereas perceptual knowledge, for instance, is first-order, reflective knowledge has a *second-order* character for Locke.

Unlike external sense, reflection functions as a source of *a priori* knowledge. As in the case of Descartes' *Cogito*, type **6** reflection is at work here. From the early modern period onwards, *something like* this meaning comes first to philosophers' minds when they think about reflection as a technical philosophical concept. To repeat, the Lockean conception of reflection dominates in philosophy. In particular, it has shaped and continues to exercise a deep influence on philosophical discussions about *introspection*, which is taken to be a special "means of learning about one's own currently ongoing, or perhaps very recently past, mental states or processes" [**Sch19**, p. 1].

Even though reflection plays a central role in his theory of mind, Locke did not go into great detail about the nature of the mental process of reflection and

its products. Some interpret Locke as claiming that, necessarily, when an idea is formed in the mind, then also, by reflection, an idea of that idea is formed. From this, the conclusion is drawn that the human mind carries out an infinite number of acts of reflection, and contains infinitely many ideas.[23] This, then, is taken by many to be in conflict with the finiteness of the human mind.[24] Others interpret Locke in such a way that a *voluntary act* is involved in reflection, which need not always occur when an idea is formed. Scharp, for instance, argues that Locke believes that when an idea is formed by the mind, the formation act of the mind leaves an *impression* in the mind which is transformed into a reflective idea only if the mind (voluntarily) turns its attention to this impression.[25]

This Lockean theme is not of mere historical interest. The idea that when an idea is formed in the mind, a higher-order idea is necessarily also formed, is echoed in views by influential contemporary analytic philosophers such as Brandom, Davidson, and McDowell, who hold that one cannot have beliefs at all without having higher-order beliefs. I do not propose to carry out Locke-scholarship here, nor am I qualified to do so. So I will not attempt to adjudicate between the above-mentioned interpretations of Locke's theory of reflection. Likewise, we shall not here further enquire into the merits and demerits of the view that first-order belief necessitates higher-order belief.[26]

We have seen in Sections 3.4 and 3.5 how in the works of authors such as Porphyrius, Avicenna, and Descartes, dictionary meaning **6** of the word reflection is connected to *self-consciousness*: to the way in which the mind acquires knowledge of itself as an entity and of its own essence. Lockean reflection, however, is a faculty that is restricted to knowledge by the human mind of its own *operations*. The question how the human mind obtains knowledge of itself as an entity and of its nature is answered by Locke as follows [**Loc89**, IV, 4, 3]:

> *we have an intuitive Knowledge of our own Existence*, and an internal infallible Perception that we are. In every Act of Sensation, Reasoning, or Thinking, we are conscious to ourselves of our own Being.

Since Locke believes the knowledge that the Mind has of itself (rather than of its individual operations) to be intuitive, it is *first-order* knowledge, rather than knowledge obtained through a process of reflection.[27] The question then arises—particularly for an empiricist such as Locke—how this intuitive knowledge that the Mind has of itself comes about.[28]

3.7. Leibniz and apperception

Leibniz has expressed and advocated core elements of a theory of reflection. Actually, *two* concepts of reflection play a role in Leibniz's work. We will discuss them in turn. We will see that Leibniz's conception of one of the concepts of reflection that is at work in his philosophy is closely related to that of Locke, but

[23]For a discussion of this objection, see [**Sch08**, Sections 1.5 and 1.6].

[24]That the human mind is finite, is claimed, for instance, by Descartes in [**Des41**, Meditation 3].

[25]See [**Sch08**, Section 2].

[26]For a critical discussion of this view, see [**Kor12**, Chapter 2].

[27]For a discussion of this question, see [**Thi94**, Section II].

[28]We do not go into this further question here.

also differs from Locke's conception of reflection in key respects. Moreover, it will become clear that it is not easy to understand *exactly* what Leibniz's views of this concept of reflection are, and that Leibniz does not have a clearly worked out theory of reflection [**Thi94**, p. 199].

One concept of reflection plays a central role in the *Monadology* [**Lei91**]. The concept of reflection that is at work there, is dictionary meaning **2** of the word 'reflection'.

A fundamental thesis in the *Monadology* is that the world consists of monads that are causally isolated from each other, where every monad reflects each other monad completely, albeit mostly not very clearly. For any two monads a and b, then, a is structurally very similar to a part of b, and b is structurally very similar to a part of a. This implies that every monad is structurally very similar to a proper part of itself. And it also implies that every monad is structurally very similar to the world as a whole. As with Philo's conception of reflection, this makes one suspect that there is a connection with the concept of infinity here. Also as with Philo,[29] one main reason why this connection was not explored, even though Leibniz was a great mathematician, is the fact that the concept of actual infinity was not well understood and even by many believed to be intrinsically paradoxical.[30]

In Leibniz' picture, God is "the central monad", which is likewise reflected in all other monads. Thus we distinguish between reflection of the Absolute in the world (of monads) on the one hand, and reflection between 'ordinary monads' on the other hand.

Leibniz has also contributed to the theory of the concept of reflection that is expressed by dictionary meaning **6** of the word 'reflection', which he mostly calls *apperception*. This will actually be our main concern in this Section.

Leibniz draws a distinction between perception and *apperception* that is reminiscent of Locke's distinction between sensation and reflection. Apperception relates to the mind's past only, and can therefore be seen as a form of immediate memory [**Thi94**, p. 199]. In a famous passage from *Principes de la Nature et de la Grâce, fondés en Raison* (1714) Leibniz describes the concepts of perception and apperception as follows:[31]

> [...] il est bon de faire distinction entre la PERCEPTION qui est l'état intérieure de la Monade representant les choses externes; et L'APPERCEPTION qui est la CONSCIENCE ou la connaissance réflexive de cet état interieur, laquelle n'ést point donnée à toutes les Ames, ni toujours à la même Ame.

Through apperception of itself, the Mind is then able to produce an idea of itself in itself:[32]

> Nostre Esprit même nous donne quelque image de cela, et pour rendre ces notions plus aisées par quelque chose d'approchant, je ne trouve dans les creatures rien de plus propre à éclaircir ce sujet, que la Reflexion des Esprits, lors qu'un même Esprit est

[29] See Section 3.3.

[30] For a discussion of the relation between Leibniz's use of reflection in the *Monadology* on the one hand, and reflection principles in set theory on the other hand, see [**vA09**].

[31] As quoted in [**Thi94**, p. 197].

[32] Leibniz, *Remarques sur le Livre d'un Antitrinitaire Anglois* (1693?), as quoted in [**Thi94**, p. 199, footnote 8]. See also [**Lei91**, par 30].

> son propre objet immediat, et agit sur soy même en pensant à soy et à ce qu'il fait. Car ce *redoublement* [emphasis Udo Thiel] donne une image ou ombre de deux substances respectives dans une même substance absolue; sçavoir de celle qui entend, et de celle qui est entendue.

So through apperception the human mind can contain an idea of *itself*, rather than just of its actions.[33]

Thus Leibniz's view of reflective knowledge differs from that of Locke. On the former view, "reflection achieves more than Locke says it does. According to Locke, reflection is a source of ideas of the operation of the mind; for Leibniz, reflection reaches not only the operations of the mind, but the mind itself" [**Thi94**, p. 207]. In Leibniz's own words:[34]

> [...] mais cette *reflexion* ne se borne pas aux seules operations de l'esprit, [...] elle va jusqu'à l'esprit luy même, et c'est en *s'appercevant de* luy, que nous *nous appercevons* de la substance.

In particular, then, for Leibniz, in contradistinction to Locke, the knowledge that the Mind has of itself as an entity is of a *second-order* nature.

Kulstad, in an important monograph on Leibniz's theory of reflection ([**Kul91**]), argues that Leibniz distinguishes between two kinds of reflection. In perception, as well as in apperception, an *idea* in the mind is created by means of a *mental act*. Two kinds of reflection on perception (as well as on apperception) are then possible. In the reflective act, one can focus one's attention on the product of perception (apperception), i.e., on the idea in the mind that is thereby produced. Kulstad calls this kind of reflection *simple reflection*. But one can also, in the reflective act, focus one's attention on the token mental act of perceiving (apperceiving). This kind of reflection is called *focused reflection*.

Whether this distinction between two kinds of reflection is clearly discernible in Leibniz's writings is subject to debate.[35] Also, it is not clear whether in a given act of reflection, one aspect (simple, or focused) can be present without the other also being present [**Thi94**, p. 206]. Nonetheless, it is a philosophically significant distinction.[36]

In any case, Leibniz's views presuppose that every human mind must have an *innate idea* of itself from the outset. According to him, many of our (innate and acquired) ideas are unconsciously present in our mind: these are the so-called *petites perceptions*. This may often apply to our idea of ourselves as well. But when, in an apperceptive act, we turn our attention to this innate idea of ourselves, we become conscious of it.

Through reflection, we can become conscious not only of our innate ideas, but also of relations between them. The principle of non-contradiction may be one such insight. Further reflection then reveals, according to Leibniz, that all of our mathematical insights can ultimately be derived from this principle of non-contradiction. (From a contemporary vantage point, this claim seems far-fetched.)

---

[33]The thesis that the mind can contain an idea of *itself* goes back at least to Philo of Alexandria: see [**Wol47a**, p. 213–214].

[34]As quoted in [**Thi94**, p. 207].

[35]Thiel, for instance, denies this: see [**Thi94**, Section III].

[36]Even critics of Kulstad's interpretation, such as Thiel, concede this [**Thi94**, p. 209].

So reflection is not merely a matter of "pure inner perception". Nor is this the case for our apperception of our own minds: as in other acts of apperception, reasoning and abstraction play a crucial role in it. The following passage is relevant here:[37]

> L'apperception de ce qui est en nous depend d'une attention *et d'un ordre* [my emphasis].

The term 'order' in this passage is taken by commentators to refer to the role of discursive elements and *rationality* in apperception.

It is the role of rational elements, and the background assumption of the existence of innate ideas, that allows apperceptive acts to go beyond the goings on in the mind and to produce substantial knowledge about the world [**Lei91**, par 30]:[38]

> C'est aussi par la connoissance des vérités nécessaires et par leur abstractions que nous sommes élevés aux Actes reflexifs, qui nous font penser à ce qui s'appelle Moy, et a considérer que ceci ou cela est en Nous: et c'est ainsi qu'en pensant à nous, nous pensons à l'Être, à la Substance, au simple et au composé, à l'immateriel et à Dieu même; en concevant que ce qui est borné en nous, est en lui sans bornes.

Leibniz was rather vague about the details of the role that type **6** reflection plays in the formation of knowledge exceeding the contents of our experience. Other rationalists tried to be more precise. In their *Port Royal Logic*, Arnauld and Nicole argued that there is an intimate relation between type **6** reflection on the one hand, and the process of *abstraction* on the other hand [**AN62**, 38, emphasis added]:[39]

> Suppose, for example, I reflect that I am thinking, and, in consequence, that I am the I who thinks. In my idea of the I who thinks, *I can consider a thinking thing without noticing that it is I*, although in me the I and the one who thinks are one and the same thing. *The idea I thereby conceive* of a person who thinks can represent not only me but all other thinking persons. By the same token, if I draw an equilateral triangle on a piece of paper, and if I concentrate on examining it on this paper along with all the accidental circumstances determining it, I shall have an idea of only a single triangle. But if I *ignore all the particular circumstances* and focus on the thought that the triangle is a figure bounded by three equal lines, the *idea I form* will, on the one hand, represent more clearly the equality of lines and, on the other, be able to represent all equilateral triangles [...] Now in these abstractions it is clear that the lower degree includes the higher degree along with some particular determination, just as the I includes that which thinks, the equilateral triangle includes the triangle, and the triangle the straight-lined figure. But *since the higher degree is less determinate, it can represent more things.*

---

[37] *Nouveaux Essais* I.i.25, as quoted in [**Thi94**, p. 206].

[38] For a discussion of the relation between rationality and reflection in Leibniz, see [**Ben16**].

[39] In the beginning of this Chapter I argued that there is at least an interesting analogy between the process of abstraction and the process of reflection: cfr supra, p. 69.

This shows that for Arnauld and Nicole, as, according to Kulstad, for Leibniz, the role of *attention* in type **6** reflection is of great importance.[40]

That abstraction is a process for forming new ideas, and that these are ideas of a special kind, is even clearer in a later work by Arnauld, where he tells the following story [**Arn83**, 74]:

> The philosopher Thales, having to pay twenty workers one drachma each, counted twenty drachmas and paid each worker. He would not have been able to do this unless there were at least two perceptions in his mind: one of twenty men and one of twenty drachmas [...] Having some spare time he began to reflect, and thinking about what the two perceptions or ideas have in common, namely that there is 20 in both, he abstracts from what is particular in them the abstract idea of the number 20, which can subsequently be applied to twenty horses, twenty houses, twenty stadiums. This is a third idea or perception.

The picture, then, seems roughly the following. In an act of type **6** reflection, multiple ideas are compared (idea of 20 drachma, idea of 20 workers). Attention is given to the way in which these ideas are similar to each other. On the basis of this, a more abstract idea is formed (the idea of the number 20), which contains only the features that the multiple ideas have in common.

3.8. Introspection

From the foregoing, it is clear that Locke's theory of type **6** reflection is the birthplace of *introspection* as an important theme in contemporary epistemology.

Schwitzgebel describes the main properties of introspection as follows [**Sch19**, section 1.1]:

(1) Introspection is a process that generates, or is aimed at generating, knowledge, judgments, or beliefs about *mental* events, states, or processes, and not about affairs outside one's mind, at least not directly.
(2) Introspection is a process that generates, or is aimed at generating, knowledge, judgments, or beliefs about *one's own mind* only and no one else's, at least not directly.
(3) Introspection is a process that generates knowledge, beliefs, or judgments about one's *currently ongoing* mental life only; or, alternatively (or perhaps in addition) *immediately past* (or even future) mental life, within a certain narrow temporal window.
(4) Introspection yields judgments or knowledge about one's own current mental processes relatively *directly* or *immediately*.
(5) Introspection involves some sort of *attunement* to or *detection* of a pre-existing mental state or event, where the introspective judgment or knowledge is (when all goes well) *causally* but not *ontologically* dependent on the target mental state.
(6) Introspection is not *constant, effortless, and automatic*.

[40] It is argued in [**Pea19**] that attention plays an important role in Locke's theory of reflection, too.

Introspection is taken to be a source of *reason*; introspection is a faculty of the rational mind. Properties 3,4, and 5 show that introspection is indeed a form of **6** reflection.

The *directness* condition (property 4) shows that discursive elements play at best a secondary role in introspection. Moreover, property 2 indicates that introspection does not by itself deliver knowledge of states of affairs outside the mind. These two elements are in line with Locke's notion of reflection, and go against Leibniz's conception of **6** reflection. Indeed, as we have said before, rationalist claims that reflection yields knowledge of the world outside the mind are regarded with much suspicion nowadays.

In contemporary philosophy, the cognitive contents of the rational mind are not taken to be ideas, but *beliefs*. But many questions of early modern philosophy concerning **6** reflection translate into questions concerning introspection.

One such question is: which are the basic logical principles concerning belief (of a rational subject) that are connected to introspection? Putative such principles are called *introspection principles*. If B is a predicate that expresses, then some of these basic schematic principles are the following:[41]

PI	$B\phi \rightarrow BB\phi$ (Positive Introspection)
CPI	$BB\phi \rightarrow B\phi$ (Converse Positive Introspection)
NI	$\neg B\phi \rightarrow B\neg B\phi$ (Negative Introspection)
CNI	$B\neg B\phi \rightarrow \neg B\phi$ (Converse Negative Introspection)

If these introspection principles were to hold, then belief would be a *luminous* condition in the sense of [**Wil00**].

If introspection were an automatic process, then Positive Introspection would be a true principle. However, we have seen earlier that Leibniz observed that this would mean that the rational mind contains infinitely many beliefs,[42] which many take to be an objectionable conclusion. Moreover, we have seen at the beginning of this Section that introspection is not typically taken to be an automatic or constant process (property 6). So principle PI is highly suspect. For similar reasons, NI is suspect. We will later see that simple diagonal arguments show that almost all general introspection principles for logical reasons cannot hold in full generality for a consistent rational mind.[43] In other words, belief is not a luminous state of mind.

The early history of proof theory has shown that all sound mathematical theories S that extend elementary arithmetic have certain positive introspective capacities. Any such theory S, formulated in a language $\mathcal{L}$, contains an 'idea' of itself in the form of a *standard provability predicate* Bew_S, such that for each $\varphi \in \mathcal{L}$:

$$\mathrm{S} \vdash \mathrm{Bew}_S(\varphi) \Leftrightarrow \varphi \text{ is a theorem of S.}$$

This can be seen as a positive introspection property.

Moreover, this provability predicate will satisfy what are called the *Hilbert-Bernays derivability conditions* for every $\varphi \in \mathcal{L}$:

(1) If $\mathrm{S} \vdash \varphi$, then $\mathrm{S} \vdash \mathrm{Bew}_S(\varphi)$;
(2) $\mathrm{S} \vdash \mathrm{Bew}_S(\varphi) \rightarrow [\mathrm{Bew}_S(\varphi \rightarrow \psi) \rightarrow \mathrm{Bew}_S(\psi)]$;

[41] We are sloppy with coding here.

[42] Unless the mind's belief that p would somehow be *identical* to its belief that it believes that p.

[43] Cfr Section 6.7. See [**SH22**] for an extended discussion of introspection principles for rational belief.

(3) $\mathrm{S} \vdash \mathrm{Bew}_S(\varphi) \to \mathrm{Bew}_S(\mathrm{Bew}_S(\varphi))$.

The Hilbert-Bernays derivability conditions are regarded as minimal conditions that a predicate Bew_S must satisfy if it is to be taken to express the concept *provability in* S.

An analysis of the proof of Gödel's second incompleteness proof shows:

THEOREM 3.8. *If a provability predicate* Bew_S *for a theory S satisfy the Hilbert-Bernays conditions, then:*

$$\mathrm{S} \vdash \mathrm{Bew}_S(\mathrm{Bew}_S(\varphi) \to \varphi) \to \mathrm{Bew}_S(\varphi).$$

This is known as *Löb's theorem*. From this theorem, it immediately follows that the negative introspection property

$$\mathrm{S} \vdash \neg\mathrm{Bew}_S(\varphi) \Leftrightarrow \varphi \text{ is not a theorem of S}$$

fails dramatically for sufficiently strong sound mathematical theories. This is the moral of Gödel's second incompleteness theorem.

3.9. Dedekind's perfectly reflective minds

We now turn to the question whether *perfect* reflective knowledge of oneself is theoretically possible, and, if it is, what such possibilities might look like. With the word 'reflection', here, reflection in the sense of meaning **6** is intended. Moreover, a *structural* notion of perfection is aimed at: self-knowledge is taken to be *perfect* if the structure of the mind is mirrored in an idea in the mind. Thus, in perfect self-knowledge in this sense, type **2** reflection is also involved. To conclude, throughout this investigation, I will use categories (mind, idea,...) that were part of the shared philosophical vocabulary in the early modern period rather than those of contemporary epistemology.

Perfect reflective self-knowledge is possible if certain *assumptions* are made, some of which have been discussed previous sections. The assumptions that are needed are the following. Firstly, the human mind is not *itself* an idea. Secondly, the human mind contains *nothing but* ideas. Thirdly, the human mind can contain not only ideas of external entities, but also contain *ideas of ideas*. This, of course, is just the concept of reflection that was defended by Locke and by Leibniz. Fourthly, by reflection, the human mind can form an idea of itself. As we have seen, this assumption was defended by Leibniz, but denied by Locke. Fifthly, ideas can contain other ideas as *part of their content*. For instance, an idea of a yellow rubber duck may somehow contain an idea of yellowness, an idea of rubber, and idea of duck as parts of its content. In particular, when b is an idea of a, then a is part of the content of b.

Let us call the human mind M. We have seen that it is populated with (zero or more) ideas $(m_0, m_1, m_2, \ldots)$. Moreover, we have seen that through a process of reflection, an idea *of* another idea can be created. If an idea m_i is in mind M, then m_i is a part of the content of M. Likewise, if an idea m_i is about an idea m_j, then m_j is a part of the content of m_i. Let us abbreviate the transitive closure of the 'being an idea of'-relation as $\prec$. Then of course $\prec$ can be seen as a transitive directed graph.

We do not want to require of the relation $\prec$ that it be anti-symmetric. It seems possible, for example, for two paintings p_1 and p_2 to be such that p_1 represents p_2 and p_2 represents p_1; likewise, it is hard to exclude that an idea a is a representation

of the idea b and *vice versa*. Similarly, we make no assumptions about the extent to which $\prec$ is or is not a *reflexive* relation.

Now we ask the following question:

QUESTION 3.9. What do minds that have structurally perfect self-knowledge look like?

This only becomes a precise question once it is explained in precise terms what is meant by structurally perfect self-knowledge. This is done in the following natural way:

DEFINITION 3.10. A mind M has structurally perfect self-knowledge iff

$$\exists m_i \neq M : m_i \prec M \wedge m_i \cong M,$$

where $\cong$ stands for the relation of isomorphism. In other words, a mind has structurally perfect self-knowledge if it contains an *idea* of the Mind that is isomorphic to the structure of the Mind.[44]

The model of the completely empty Zen-mind does not represent a mind that has perfect self-knowledge: according to our assumption one above, the mind is not an idea, so we cannot have $M \prec M$, whereby the Zen-mind is not perfectly self-knowing in our sense of the word. Moreover, since in addition $\prec$ is transitive, also the slightly more complicated model of the Mind M that contains only one idea m, which is an idea of M does not represent a mind that has perfect self-knowledge. All this does not mean that, according to the assumptions that we have made, the Mind cannot belong to the content of ideas—the point is just that the Mind is not an idea of itself.

PROPOSITION 3.11. *The mind M with ideas $m_0, m_1, m_2, \ldots, m_i, \ldots$ (for all $i \in \mathbb{N}$), such that*

$$\ldots \prec m_i \prec \ldots \prec m_0 \prec M,$$

and which contains no other ideas, is a mind that has structurally perfect self-knowledge.

PROOF. Obviously $m_0 \cong M$. □

Intuitively, this Mind can be conceived of as follows: m_0 is an idea of the mind M formed by reflection,[45] m_2 is a reflexive idea of m_1, and so on, *ad infinitum*.

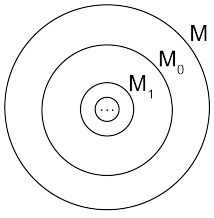

FIGURE 1

[44]Using the notion of homomorphism instead, ideas of less-than-perfect self-knowledge can be expressed.

[45]As mentioned earlier in this Section, this is not something Locke himself would agree with.

This is a variant of Dedekind's notorious "proof" of the existence of an infinite collection [**Ded88**, 66]:

> My own realm of thoughts, i.e., the totality S of all things which can be objects of my thought, is infinite. For if s signifies an element of S, then the thought s' that s can be an object of my thought, is itself an element of S. [...] then S is infinite, which was to be proved.

With the word 'infinite' in this quote, Dedekind means *Dedekind-infinite*, which is defined as follows [**Ded88**, 64]:

DEFINITION 3.12. A set S is Dedekind-infinite if there is a $S' \subsetneq S$ such that there is a bijective function from S to S'.

Dedekind-infinity, then, is a *reflection property*, with 'reflection' taken in sense **2** of the word.

Then we have the following simple observation:

PROPOSITION 3.13. *Every perfectly self-knowing mind contains an infinite reflection chain*

$$\ldots \prec m_i \prec \ldots \prec m_0 \prec M.$$

PROOF. This follows by a simple cardinality argument, using the assumption that there is an idea m_0 of the Mind. □

In other words, every perfectly self-knowing mind is infinite, and the mind M of Proposition 3.11 is a *minimal model* of a perfectly self-knowing mind. Moreover, we see that the minimal model of the proof of Proposition 3.11 in fact contains *infinitely many* ideas that perfectly reflect the mind: every m_n does so.

Observe that our minimal model of a perfectly self-knowing mind is not well-founded—it is conversely well-founded. Indeed, no well-founded perfectly self-knowing minds exist:

PROPOSITION 3.14. *No perfectly self-knowing mind has a well-founded $\prec$-relation.*

PROOF. This follows from Proposition 3.13. □

Of course we do not want to make too much of the non-well-foundedness involved here: we might as well have focussed on the converse of $\prec$.

Now contain the following axiom:

AXIOM 3.15. There is a perfectly self-knowing mind.

Proposition 3.13 shows that this is an *axiom of infinity*. Given minimal assumptions on minds, ideas, and reflection, this axiom postulates objects (minds) that are infinite. Dedekind's "proof" of the existence of an infinite collection can be seen as an argument for this Axiom.

The Mind described in Proposition 3.11 can in a straightforward sense be seen as a model

$$\mathfrak{M} = \langle \{m_0, m_1, \ldots\}, \prec \rangle$$

for a first- or second-order language $\mathcal{L}_\prec$ that has $\prec$ as its sole non-logical symbol. If $\mathcal{L}_\prec$ is second-order, then Axiom 3.15 can be expressed in it, and be seen to hold in $\mathfrak{M}$.

Axiom 3.15 is a reflection principle—with 'reflection' taken in sense **2** of the word. It follows from work of Frege[46] that Axiom 3.15 has significant consistency strength, i.e., mathematical strength: in a second-order context, it allows one to interpret full second-order Peano Arithmetic. Moreover, Axiom 3.15 entails a version of the schematic second-order Bernays reflection principle:[47]

$$\exists m \forall Y \forall y \prec m : \varphi(Y, y) \leftrightarrow \varphi^{m, \mathcal{P}(m)}(Y \cap m, y),^{48}$$

and that any witness in $\mathfrak{M}$ of Axiom 3.15, seen as a model in its own right, is elementary equivalent with $\mathfrak{M}$. However, in the context of pure second-order logic, this Bernays reflection principle does not entail the existence of an ω-sequence, as can be seen as follows.[49] Consider a non-standard model $\mathfrak{N}$ of second-order arithmetic containing $2^{\aleph_0}$ elements, among which there are two (non-standard) numbers m, n such that according to the model, $m < n$, and such that any formula $\phi(x)$ holds (in $\mathfrak{N}$) of m if and only if it holds (in $\mathfrak{N}$) of n.[50] Then the restriction of $\mathfrak{N}$ to all elements $\leq n$ (call this model $\mathfrak{N}^{\leq n}$), Bernays reflection holds. But $\mathfrak{N}^{\leq n}$ thinks of itself that it is finite.

It may be instructive briefly to compare Axiom 3.15 to the basic *first-order* set theoretic reflection principle of Montague and Levy, ([**Lev60b**], [**Mon61**]),[51] which says that every (first-order) truth in the set theoretic universe is such that it also holds when the domain of discourse is restricted to some set in the universe. In a fairly weak fragment of set theory (consisting of the Empty Set Axiom, Extensionality, and the Singleton Axiom), infinitely many sets can be proved to exist in the universe, i.e., a potential infinity can be proved to exist. The Montague-Levy Axiom can then be used to derive from this potential infinity that also an actually infinite set must exist. There is a sense in which Axiom 3.15 is stronger than the Montague-Levy principle. We have seen that even in the absence of the Axiom of Foundation, and without a potential infinity being assumed at the outset, Axiom 3.15 immediately posits an actual infinity.

Iteration is often taken to be of central foundational importance both in arithmetic and in set theory. The natural numbers are often regarded as "formed" by iteration of the successor operation; the sets are often regarded as "formed" by iteration of the power set operation (the iterative conception of set).[52] Reflection, on the other hand, is often seen as an *addition* to the more basic idea of iteration, or even to some extent derived from it (since Montague-Levy reflection is provable in ZFC).

But the foregoing considerations indicate that reflection can likewise be taken to be *basic*. Dedekind observed that, in the context of a countable choice axiom, the postulation of a Dedekind-infinite collection is *equivalent* to the postulation of an ω-sequence, or *simply infinite system*, in his terminology. Dedekind-infinity is

---

[46] See [**Fre84**].

[47] Bernays reflection is discussed on p.168.

[48] Here the superscripts m and $\mathcal{P}(m)$ indicate relativisation of the first- and second-order quantifiers, respectively, where $\mathcal{P}$ is the power set operation.

[49] Thanks to Sam Roberts for this observation and for the following argument.

[50] Indeed, a famous theorem of Harvey Friedman says that every countable non-standard model of Peano Arithmetic is isomorphic to an initial segment of itself [**Kay91**, Section 12.1]. This is a strong ontological reflection theorem for non-standard models of arithmetic.

[51] Montague-Levy reflection is discussed on p. 167.

[52] See for instance [**Boo71**].

motivated by a type **2** reflection thought, whereas ω-sequences are motivated by type **6** reflection considerations. From a formal point of view, at least, and *modulo* the Axiom of Choice, both have an equal claim to fundamentality.

In this way, in the work of Dedekind, a meaningful connection with mathematics is made, and basic insights into the connection between reflection and infinity are finally obtained.

3.10. From Hume to Kant and beyond

Let us return to our whistle tour review of the history of reflection in philosophy. We pick up the thread after Locke and Leibniz.

Hume is sceptical about the possibility of obtaining knowledge of the self through reflective mental acts [**Hum40**, Book 1, Chapter 4, par 6]:

> [T]here are some philosophers, who imagine we are every moment intimately conscious of what we call our self [...] For my part when I enter most intimately into what I call myself, I always stumble on some particular perception or other, of heat or cold, light or shade, love or hatred, pain or pleasure. I never can catch myself at any time without a perception, and never can observe anything but the perception.

In this way, he denies claims made by rationalist philosophers such as Descartes and Leibniz. Over time, this Humean claim has become almost a commonplace in philosophy. For instance, an echo of this view is found in Wittgenstein's *Tractatus*, where he writes [**Wit22**, par 5.633]:

> Wo in der Welt ist ein metaphysisches Subjekt zu merken? Du sagst, es verhält sich hier ganz wie mit Auge und Gesichtsfeld. Aber das Auge siehst du wirklich nicht.

This does not mean that Hume agrees with Locke that our knowledge of ourselves as a subject is first-order and intuitive in nature. Concerning self-knowledge, he is a more thorough-going empiricist than Locke. Hume argues that there is no underlying carrier of our ideas and impressions. Rather, the mind is no more than a plurality or *bundle* of impressions [**Hum40**, Book 1, Part 4, par 6]:[53]

> [...] they are the successive perceptions only, that constitute the mind [...] [T]here is properly [...] no identity [of the mind] [...] [at] different [times].

However, we ordinarily assume that we do have a mind that is the bearer of our experiences. To explain our pre-theoretic thought patterns, Hume needs an error theory. He argues that *similarities* between different perceptions are responsible for our error: "our propension to ascribe identity where empirical evidence suggests diversity 'is so great' that our imagination creates the notion that there is something that underlies the succession of related objects and binds them together into a unitary and identical entity" [**Thi11**, p. 391].

Kant agrees with Hume that there is no intuition of the self "through which it is given as object" [**Kan87**, B408]. Against Hume, Kant believes that the unity of consciousness over time requires the assumption of a mind that *has* the conscious experiences. But since, for Kant, our knowledge of ourselves is not experiential, the

[53]For a detailed discussion of Hume's bundle theory of the mind in its historical context, see [**Thi11**, Part VI].

mind takes the form of a *theoretical posit*: it is needed to explain how we can have knowledge at all.

Despite Hume, Kant takes a notion of reflection to be of crucial importance: he calls this notion *transcendental reflection*.[54] Transcendental reflection "determines the origins of key cognitive concepts in sensibility, understanding, or reason, and the a priori roles and relations of these concepts in cognitive judgment, and thus their contributions to the possibility and validity of knowledge, especially of synthetic knowledge *a priori*" [**Wes03**, p. 141].

A transcendental reflection is an argument that starts from a *thought experiment* that brings out certain cognitive incapacities. But it also essentially contains *discursive* components. Transcendental reflection allows us to obtain knowledge *with modal strength* about the structure of our cognition. Thus Kant agrees with the rationalists (and against the empiricists), but for completely different reasons, that reflective reasoning can yield synthetic knowledge that exceeds the content of our mind.

A few examples may be helpful. Kant states that "[o]ne cannot make oneself a representation that there be no space" [**Kan87**, A24 / B38]. Trying to imagine a non-spatial world reveals that non-spatial representations are not possible for us. From this, Kant infers that *our* world is *necessarily* spatial. In this way, knowledge that goes beyond the content of our mind is obtained. In a similar vein, Kant asserts that we must be able to identify our representations as our own, for "otherwise I would have as multicolored, diverse a *self* as I have representations of which I am conscious" [**Kan87**, B 134]. A Humean plurality of experiences is not what our mind can be; for human beings like us, it is impossible to exist without having a mind that is sensibly affected.

Such arguments can be labelled *thought experiment arguments*. In such arguments, thinking about *concepts* is not typically the *main* concern. In the first example that we discussed, Kant is in the first place concerned with the nature of representations. In Kant's transcendental reflection on the unitary cognitive mind, what seems primary at stake, is not the concept of knowledge, but its nature. In this sense, transcendental reflections are arguments that are not obviously or primarily of a higher-order nature.

There has been a vast discussion about the persuasiveness of such Kantian arguments, and about what kind of modal strength is involved here: I do not propose to adjudicate here in any of the relevant debates. What is important for us, however, is to observe that *if* we are dealing with reflection at all here, then a different notion of reflection is at work here than the concept of reflection that was at stake in the debate between the rationalists and the empiricists. This is then not type **6** reflection, but with reflection in the sense of a *thought experiment argument*: type **8** reflection, if you will.

Thought experiment arguments have played and continue to play a central role in analytical metaphysics at least since the 1970s. The role of *examples* has become very important here, but it is not completely new [**Bur13a**, p. 539]:[55]

[54] For an account of the nature of transcendental reflection and its role in Kant's philosophy, see [**Wes03**].

[55] Indeed, Avicenna's argument of the flying man in Section 3.4 is a thought experiment argument.

> In the twentieth century a number of philosophers have emphasized the role of examples in attaining illumination through reflection. This emphasis is not new. Some classical rationalists, particularly Socrates and Descartes, used examples prominently in reflection. The aim is to use examples to arrive at principles. The classical idea is that in making judgements about examples, we are guided by principles. The examples help make the principles more explicit.

Kripke and Putnam's "twin earth" example constitutes an illustration of how examples guide thought experiment arguments in contemporary philosophy. If a linguistic community on a far away planet called a substance 'water' that behaves exactly like water, feels like water, looks like water..., but has a different chemical composition ("XYZ") from the substance that we call 'water' here on earth, then this substance is not water. Hence, *necessarily*, water is H_2O.[56] More generally, natural kinds necessarily have some of the micro-structural properties that they actually have.

In such philosophical thought experiment arguments, *counterfactual reasoning* plays an essential role: they are *what if?*-arguments. Also, such arguments typically involve thought about concepts or propositions. This does not make these arguments higher-order in the same sense as Locke and Leibniz's reflections are. For Locke and Leibniz, reflection involves attending to what is in your mind. But, as Frege and Putnam have argued: concepts and propositions (meanings) *ain't in the head*.

3.11. The value of reflection

It is not much of an exaggeration to say that in contemporary analytic philosophy, only type **6** reflection is an active research theme. One research question that has received quite a bit of attention in recent years is: *what is the value of type **6** reflection? Wherein consists its epistemic importance?*

We already know that the history of this question goes back at least to Descartes.[57] According to Descartes, our whole knowledge about ourselves, the external world, and God is obtained largely by type **6** reflection. To most contemporary philosophers, it seems that type **6** reflection cannot bear this weight. Nonetheless, I will argue in Chapter 8 that type **6** reflection plays an essential role in certain processes in which *new* knowledge about the external world is obtained.

The Lockean thought that type **6** reflection provides us with a unique way of coming to know the operations of our mind is, in some forms, still defended today. There is a strong feeling that type **6** reflection gives us, in the form of beliefs by acquaintance (introspection), privileged first person access of the contents of our mind.[58]

Nonetheless, this view has come under attack from several corners in the second half of the twentieth century. In the post-war period, Wittgensteinian ordinary language philosophers argued that the mind-body duality that is presupposed in the

---

[56] A similar point could be made about Burge's equally famous "arthritis" thought experiment argument, which is formulated in [**Bur79**].

[57] See Sections 3.5, 3.6, and 3.7.

[58] For a classical discussion of privileged access, see [**Hei88**].

Lockean thought is highly problematic.[59] On their view, mental properties attach to actions rather than to alleged 'inner events'. This smacks forms of reductionism from the mental to the physical (behaviorism) that is no longer fashionable. Nonetheless, in more recent times, naturalistic philosophers have argued forcefully for the less extreme claim that the cognitive sciences, and experimental cognitive psychology in particular, provide us with much more detailed and reliable information about what we believe than introspection can ever do.[60]

Lastly, there is the thought that type **6** reflection improves the epistemic quality of our first-order beliefs. The idea is that this type of reflection improves the average *reliability* and / or our average *justification* of the beliefs that we hold. Again, this is a natural thought. Suppose you have constructed a complicated machine that produces a specific kind of foodstuffs. Then, motivated by a a desire to ensure reliability of quality the finished product, you add a variety of control devices to selected parts of the machinery (thermometers, pressure guages, The worker who operates the machine then makes adjustments to the machine, depending on her readings of the monitoring systems. So it is, one might think, with one's belief system.[61]

Here, too, however, naturalist epistemologists voice a sceptical note.[62] Studies from experimental psychology indicate that when a person is asked critically to reflect on the way in which she has arrived at a given belief (the belief that p), she tends to engage in *confabulation* about the sources of her belief: confabulation by which she herself is taken in. Only rarely does she, through reflection, come to revise her belief that p, even if this belief has in fact been unreliably formed. The process of reflection tends in practice to be a *self-congratulatory* affair. Type **6** reflection on the sources of her belief gives a cognitive agent the *illusion* of being in control of her belief-forming mechanisms, but tends in practice to be a very unreliable monitoring mechanism. This is the more so because so many aspects of our belief-forming processes take place unconsciously, and are therefore not readily available to be monitored.

3.12. Burge on reflection

The naturalistic considerations of the previous section are intended to cast doubt on the hypothesis that reflection actually plays a valuable epistemic role in our daily cognitive lives. They do not, however, show that type **6** reflection *cannot* be a powerful epistemic instrument.[63] Tyler Burge has argued that type **6** reflection has epistemic qualities that make it an essential epistemic tool especially *in philosophy*.

As a first approximation, Burge characterises the philosophical process of reflection in the following terms [**Bur13a**, p. 534]:

---

[59] See for instance [**Ryl49**].

[60] See [**Kor12**, Section 1.3].

[61] Some even go so far as to argue that justified belief is not possible without reflecting on one's reasons for one's belief—see for instance[**Bon85**, Chapter 3] Kornblith observes that such views would lead to an infinite regress [**Kor12**, Section 1.1], so we will leave them aside here here.

[62] [**Kor12**, Section 1.3].

[63] The main critic of type **6** reflection, Hilary Kornblith, recognises this: see [**Kor12**, p. 136].

> As a first approximation, reflection is a type of rational cognition with four significant features. It goes beyond what is immediately obvious. It is higher-order, in the sense that it involves thought about psychological states or representational contents, although its conclusions need not be about psychological states or representational contents. It aims at constitutive understanding. And it develops such understanding by drawing conclusions, without acquiring new premises, empirical or otherwise, beyond what is already understood or known. Usually reflection aims at improving pre-reflective understanding or knowledge. Much reflection makes use of empirical background knowledge. Where the force of its warrant is independent of sense experience, reflection is apriori.

Burge's point of departure is what he calls the rationalist conception of reflection, of which he sees Leibniz as the main exponent. In other words, he starts from what we have called type **6** of reflection—leaving aside the sense in which Leibniz's monads reflect each other,—but is not concerned with early modern *empiricist* theories of type **6** reflection at all.

I have tried to argue that there are different *concepts* of reflection at work in the history of philosophy, and that we conflate them at our peril. In particular, it seems to me that what Kant calls 'transcendental reflection' refers to a philosophical cognitive process that is fundamentally different from type **6** reflection. Burge, in contrast, takes there to be a uniform target—*one* type of philosophical cognitive process—of the rationalist, Kantian, and twentieth century analytical *conceptions* of reflection. In particular, Burge subsumes certain forms of conceptual analysis and philosophical thought experiment arguments under the umbrella of philosophical reflection.

Burge argues that classical rationalist philosophers attribute three cardinal properties to reflection [**Bur13a**, p. 535–537]:

(1) In reflection an individual brings to articulated consciousness steps or conclusions that are implicitly present, subliminally or unconsciously, in the individual's mind before reflection.
(2) Reflection can yield a priori knowledge of objective subject matters, beyond thoughts that the reflector is engaging in.
(3) Successful reflection requires skilful reasoning and is difficult: it is not a matter of one-off introspection or intuition.

Burge endorses Theses 2. and 3., but rejects Thesis 1.

We have seen that according to the rationalist conception of reflection, as well as according to Kant's conception of reflection, Thesis 3 is correct. Also if we take the best known thought experiments of contemporary metaphysics and philosophy of mind as paradigmatic examples of reflection, Thesis 3 applies. So we may (and will) take Thesis 3 as given.

Thesis 2, on the other hand, sounds somehow suspect to the contemporary analytical philosopher. We have already seen that thought experiment arguments yield "knowledge of objective subject matters, beyond thoughts that the reflector is engaging in".[64] To be sure, not all such knowledge is a priori. Our knowledge that,

[64]See Section 3.10.

necessarily, water is H_2O, for instance, is a posteriori—it depends, among other things, on an empirical scientific discovery. But sometimes the knowledge obtained by reflection is, according to Burge, a priori [**Bur13a**, p. 555]:

> Reflection can serve as an adjunct to any enterprise. But I believe it also offers substantive insights that are not parasitic on or merely supplemental to the natural or mathematical sciences. The insights are limited and fallible. Their nature remains to be better understood. But it seems to me that apriori reflection can yield limited autonomous insight, and even knowledge, in certain parts of semantics, philosophy of mind, and epistemology. This is because we are reasoning about reasoning itself. Apriori understanding in these areas is constitutive of understanding fundamental aspects of critical reason, and of us as critical reasoners.

The worked-out example that Burge adduces of a priori knowledge obtained through reflection concerns our knowledge of the concepts of logical consequence and logical validity, obtained as the result of reflection on the practice of good deductive consequence [**Bur13a**, section III]. This is a successful case of *informal rigour* in the sense of [**Kre67**].

Burge's main reason for disagreeing with Thesis 1 is that reflection is more often than not applied in a situation where we do not have even an implicit, unclear, or confused idea or conception of a concept, i.e., in situations where we have not yet developed a concept at all. We may, for instance, merely have a small number of examples that we are inclined to see as similar in a way that we cannot describe. Or we may be disposed to classify a fairly well circumscribable number of examples as being similar in a significant way without this disposition being in any way conceptualised by us: think of this disposition as being hard-wired without an accompanying cognitive representation even at the sub-personal level.

When we consider Kant's examples of transcendental reflection, Kripke and Putnam's thought experiments, or even the analysis of validity and logical consequence, Thesis 1 indeed sounds implausible. As Burge says [**Bur13a**, p. 539]:

> Psychology has brought out that accessible higher-order, person-level cognitive control plays a very small role in much of our propositional activity. Principles that best explicate a mathematical or natural-scientific concept are often discoverable only by developing new knowledge, knowledge that it is not psychologically plausible to impute to the unconscious of reasoners before the new knowledge is discovered.

In the light of the findings of empirical science it is simply unlikely that we subliminally all along had the concept of natural kind that Kripke and Putnam argued to be the correct one; it is simply implausible that we already unconsciously had the Tarskian concept of logical consequence in Antiquity.

3.13. What is reflection?

Now that we have come to the end of our historical journey, let us return to our main question: *What is reflection, in the philosophical sense of the word*?

We have isolated what we have called type **2** reflection. This is a relation between a reflected object and a reflecting object, which holds when there exists a salient *similarity* between them. Not all similarity relations will do. Yet there are at least two ways of making the notion of similarity precise. It can be done using a relation of structural ontological similarity (isomorphism, or a related notion), or using a relation of ideological similarity (elementary equivalence).

We have also isolated what we have called type **6** reflection. This is a process in which the human mind cognitively relates to its own processes and products.

Both type **2** reflection and type **6** reflection are *iterable*. Moreover, as we have seen,[65] there exist deep relations between these two types of reflection. And both of them play a role not only in philosophy, but also in parts of mathematics. Type **2** reflection has in the twentieth century proved to be fruitful in mathematics—in set theory, for instance; type **6** reflection plays a central role in the motivation of proof theoretic reflection principles.[66]

The Kantian notion of transcendental reflection and the modern notion of thought experiment argument are not clearly related to any of the dictionary meanings of the word 'reflection'. Burge, however, believes that these should rather be seen as better conceptions of type **6** reflection: he wants his article on reflection to be seen as a constructive criticism of the "traditional conception of reflection, and highlights Kant's anticipation of a more adequate conception" [**Bur13a**, p. 534]. A main reason for this suggestion is that, in his view, *higher-order elements* such as thinking about concepts play a role in these philosophical cognitive processes.

I am not convinced. As we have seen, in Kantian transcendental reflection, it is not always, or perhaps not even typically, *concepts* that are at the center of the argument. Many of the paradigmatic modern philosophical thought experiment arguments are primarily about the the nature of certain concepts. But being essentially about or essentially involving thought about concepts is not sufficient for being an instance of type **6** reflection. Otherwise the concept of **6** reflection becomes so far as to cover much if not most of philosophical argumentative activity.[67] It would then cover all varieties of *conceptual analysis*, such as concept explication in the sense of Carnap, ordinary language philosophy in the sense of the later Wittgenstein, informal rigour in the sense of Kreisel,... The resulting concept of reflection in the philosophical sense seems to me too general for us to expect it to be theoretically very fruitful. What is more, large tracts of mathematical activity would then also count as reflection in this sense. As an example, recall Lakatos' case study of the repeated revision of the concept of polyhedron in the course of the history of proving general forms of Euler's theorem [**Lak76**].

The boundaries of type **6** reflection are not completely clear. Type **6** reflection has a *de se* component: it is about a cognitive first person relation that one has to one's own mind and what is going on in it. Concepts are not in the mind, so conceptual analysis does not by itself qualify as type **6** reflection, nor do informal rigour, and axiomatisation of parts of mathematical practice. But what is in our mind is *related* to concepts, and to formal theories. Concepts and propositions belong to the *content* of what is going on in our minds; axiomatic theories capture the

[65] See Section 3.9.

[66] See Chapter 8.

[67] An example of a (now largely forgotten) view that the distinctive feature of philosophical thinking in general is that it is reflective in a technical philosophical sense, is [**Hod78**].

content of what we believe. So axiomatically expressing the maximal mathematical theory that one currently believes, for instance, does count as type **6** reflection.

At any rate, I claim that Kant's transcendental reflections, the conceptual analysis of logical consequence, and especially contemporary thought experiments should not be classified as type **6** philosophical reflections. They *may* be still labelled philosophical reflections of some other type—*what's in a name?*—but this would then be some loose sense of the word.

Nonetheless, as we will see later,[68] *what if*-arguments can play a crucial role **in** type **6** reflection arguments. As mentioned before, philosophical thought experiments play a major role in contemporary philosophy. Thought experiment in philosophy is a subject that cries out for meta-philosophical inquiry into its nature and evidential force. For an interesting theory of philosophical thought experiment arguments, see [**Wil07**, chapter 6].

With all this in mind, let us return to Burge's evaluation of the three cardinal properties that rationalist philosophers attribute to type **6** reflection. Thesis 3 is unaffected by the preceding considerations, so it still stands. It also seems to me that, largely for the reasons that Burge adduces, that Thesis 1 is false for type **6** reflection. We have seen that discursive components form part and parcel of this kind of reflection. It is simply implausible that the conclusions of such reflective argumentations are somehow already 'unconsciously' present in the mind before it has gone through the reflective arguments.

Thesis 2 is a different matter. We have seen that Burge's support for Thesis 2 crucially depends on examples of philosophical argumentation that, in my view, should not be seen as cases of type **6** reflection: the metalogical conceptual analysis of the concept of logical consequence, and certain examples of thought experiment arguments from contemporary philosophy. So Burge's argumentation leaves Thesis 2 completely open. I will argue later,[69], however, that for reasons unrelated to Burge's arguments, Thesis 2 also holds for type **6** reflection.

3.14. Taking stock

This Chapter was to a significant extent an attempt at a brief overview of the history of reflection. We wanted to find out how concepts and conceptions of reflection have evolved in philosophy over time, and we inquired into the theoretical power and fruitfulness of philosophical concepts of reflection. Two main philosophical concepts of reflection emerged: type **2** reflection and type **6** reflection, where type **2** reflection is an ontological form of reflection, whereas type **6** reflection is an epistemic form of reflection. In the remainder of this book, we will *predominantly* be occupied with the latter type of reflection, but the former will also be given a considerable amount of attention.

The concept of type **2** reflection, reflection as mirroring, appears to be the oldest in the history of philosophy. At least from Philo onwards, for a long period, it led a largely independent life. The concept went through many incarnations (versions of Microcosmos / Macrocosmos-theories), up to Leibniz's *Monadology*. We have been mainly concerned with the thought that a transcendent whole is reflected in some of its immanent small parts. But we have learned from the Microcosmos /

[68] See Section 8.3.2.

[69] See Chapter 8.

Macrocosmos-theories and from Leibniz's writings that there may also be reflection relations *between* small parts of a transcendent whole.

In Plato's cosmology, a form of type **3** reflection was at work. This form of reflection lies at the root of a long historical process leading to the isolation of type **6** reflection: mental attention to the contents and operations of one's mind. A gradual "Vergeistlichung" of Plato's *World Soul* played a crucial role here. By the time of the Neoplatonists (such as Porphyry), a conception of type **6** reflection was established. It played an important role in the formation of the Cartesian subject.

In Leibniz's views on *apperception*, a sort of synthesis of type **2** and type **6** reflection is attained. On the one hand, the Leibnizian "redoublement", in which the Mind forms an idea of itself, is clearly a form of type **6** reflection. On the other hand, this "redoublement" results in a mirroring (type **2**) of the Mind in itself in the form of a reflexive idea. This furthermore led to a deep connection with the modern conception of infinity.

Concerning the period starting with Kant and onwards, our review of the history of reflection in philosophy has been somewhat sketchy. I am aware that already concerning Kant's view of reflection, we have barely scratched the surface. Moreover, nothing at all was said about the important role that ideas of reflection play in nineteenth century Idealism. There is no particular reason for this over and above the bare fact that I feel utterly incompetent to do so. For all I know, there may be one or more important concepts of reflection waiting to be discovered in the work of nineteenth century philosophy.

Already our very limited historical exploration has revealed that type **2** reflection and type **6** reflection are theoretically powerful and fruitful ideas. In Section 3.5, we saw that type **6** reflection is related to results about the scope and limitations of knowledge; in Section 3.9 we saw how type **2** reflection is related to infinity. These connections will be deepened and explored further in what follows. We have already seen that type **6** reflection remained a theme in analytic philosophy from the post-world-war two years until the present. Especially Burge's theory of type **6** reflection will turn out to be of importance in what follows.[70]

[70]See especially Chapter 8.

Part II

CHAPTER 4

Some foundationally significant theories

After our discussion of the philosophical concepts of epistemic warrant and reflection, in this part we slowly turn our attention to formal reflection principles. As a preamble to this, in this Chapter we review some formally significant mathematical background theories. In the Chapter thereafter, we review some formal truth theories, which play a significant role in the discussion of reflection principles of the proof-theoretic kind. The material that is discussed in the present Chapter and in the the next Chapter does not contain much that is new. But in later Chapters we will refer back to it and use the results that are discussed here. Readers who are familiar with the material may therefore want to skip most of the following two Chapters, and leaf back later when the need arises.

In the present Chapter, we consider not only theories from pure mathematics (arithmetic and set theory), but also theories of probability. We will state, or in some cases sketch, axiomatic presentations of the theories in question, and point to some important metamathematical results concerning them. We will consider both 'full' theories and important fragments of full theories; we will consider both first-order and second-order theories.

The foundational significance of the presented theories will also be discussed. We will see that from a naturalist perspective, the full theories are typically of most philosophical importance: they are natural formalisations of theories that mainstream mathematicians reason in. From various foundationalist perspectives, in contrast, certain fragments of full systems are judged to be of fundamental importance.

In what follows, we assume that the reader is familiar with classical first-order logic and the most central metamathematical results.[1] Moreover, it will be assumed that she has some basic background knowledge of second-order logic.[2]

4.1. Arithmetic

Our theory of the natural numbers is an absolutely basic part of mathematics. Gödel's incompleteness theorems show that the collection of all arithmetical truths cannot be captured by any single axiomatic system. In this sense, our best theory of the natural numbers will always be *incomplete*. Nonetheless, we can attempt axiomatically to capture our best theory of the natural numbers, or fragments of it. This is an activity that started in the second half of the nineteenth century, and is still ongoing.

Two logical frameworks for formalising arithmetic need to be distinguished. On the one hand, we have a framework in which only quantification over natural

[1] The material in [**BBPJ02**] suffuces.

[2] Knowledge of the material in Part II of [**Sha91**] suffices.

numbers is allowed. This is called first-order arithmetic.[3] The background theory for this is classical first-order logic. On the other hand, there is a framework in which also quantification over *collections* of numbers is allowed. This is called second-order arithmetic. The background theory for second-order arithmetic is some classical theory of second-order logic, which can either be the full standard theory of second-order logic, or a fragment of it.

In general, theories of the natural numbers are of foundational importance because over the past two centuries, many have entertained and pursued the hypothesis that much of mainstream mathematics can *in some sense* be reduced to first- or second-order arithmetic.

4.1.1. First-order Peano Arithmetic and some of its fragments. The best known axiomatisation of first-order number theory is known as *first-order Peano Arithmetic* (PA). The language of PA ($\mathcal{L}_{PA}$) contains 0, s, $+$, and $\cdot$ as its sole non-logical symbols, and contains the following axioms:

PA1 $\neg\exists x : s(x) = 0$;
PA2 $\forall x \exists y : s(x) = y$;
PA3 $\forall x, y : s(x) = s(y) \to x = y$;
PA4 $\forall x : x + 0 = x$;
PA5 $\forall x, y : s(x + y) = x + s(y)$;
PA6 $\forall x : x \cdot 0 = 0$;
PA7 $\forall x, y : x \cdot s(y) = (x \cdot y) + x$;
PA8 $[\phi(0) \wedge \forall x(\phi(x) \to \phi(s(x))] \to \forall x \phi(x)$ for all fomulas $\phi(x) \in \mathcal{L}_{PA}$.

Here PA8 is not a single axiom but an *axiom schema*: it is called the *mathematical induction schema*. It says that if a definable property (ϕ) propagates ('progresses upward') in the natural numbers, then it holds for all natural numbers.

Restricting the mathematical induction schema results in important fragments of PA. For any natural number n, the result of restricting the mathematical induction schema to Σ_n-formulas is the system called $I\Sigma_n$.

In a similar way, the systems $I\Delta_n$ (for $n \in \mathbb{N}$) are defined. The following lemma shows that there is no need to distinguish fragments of PA restricted to *universal* formulas of a given complexity, for the universal hierarchy coincides with the existential hierarchy [**Bek05**, Lemma 4.1]:

LEMMA 4.1. *For each* $n \in \mathbb{N} : I\Sigma_n = I\Pi_n$.

On the other hand, by increasing n we obtain ever stronger fragments of PA:

LEMMA 4.2. *For each* $n \in \mathbb{N} : I\Sigma_n \subsetneq I\Sigma_{n+1}$.

The system $I\Sigma_0$ is very weak. Like the even weaker arithmetical system Q, which results from removing *all* induction axioms from PA, it does not have a good theory of its own syntax. This is related to the fact that the natural way of coding finite sequences of signs (terms, formulas, proof) uses the *exponentiation function*, and $I\Sigma_0$ cannot prove exponentiation to be a total function. To remedy this, one can extend the language with a basic symbol for exponentiation, and add to Q the defining axioms for exponentiation, so that the totality of exponentiation is assumed at the outset.

[3]The standard reference for the metamathematical investigation of first-order theories of arithmetic is [**HP93**].

Especially $I\Sigma_1$ is considered an important fragment of PA. It is a 'basic' system of arithmetic that is nonetheless strong enough for not only Gödel's first incompleteness theorem, but also *Gödel's second incompleteness theorem* to hold for it:

THEOREM 4.3 (Gödel's second incompleteness theorem). *The Gödel sentence for $I\Sigma_1$ is independent from $I\Sigma_1$, and $I\Sigma_1$ cannot prove its own consistency.*

This is a very robust result: it holds for all sound arithmetical systems that are at least as strong as $I\Sigma_1$, and in particular for PA. Moreover, a surprisingly large portion of mainstream arithmetical theorems can be proved in $I\Sigma_1$.

The collection of true *atomic* arithmetical formulas is definable by an arithmetical formula val^+, and the set of false atomic arithmetical formulas is definable by an arithmetical formula val^-. Moreover, *truth* for formulas of restricted complexity is also arithmetically definable:

LEMMA 4.4. *For every $n \in \mathbb{N}$, the collection of the Σ_n arithmetical truths is definable by means of a Σ_n-formula.*

As a counterpoint to this, Tarski has famously shown:

THEOREM 4.5 (Tarski). *The collection of first-order arithmetical truths is not definable in the language of first-order arithmetic ($\mathcal{L}_{PA}$).*

This is known as *Tarski's theorem on the undefinability of truth*. It again is a very robust phenomenon: Tarski's theorem holds not just for $\mathcal{L}_{PA}$, but also for all extensions of $\mathcal{L}_{PA}$.[4]

PA is perhaps the most natural system for formalising 'elementary' number theoretic proofs. Axioms PA1-PA7 express the most basic properties of the natural numbers. The principle of mathematical induction is the work horse for carrying out elementary arithmetical proofs. PA8 is an extremely natural way of expressing the principle of mathematical induction. For this reason, combined with the aforementioned suspicion of many mathematicians that there is a sense in which much of mathematics can be reduced to the theory of the natural numbers, PA is a foundationally significant theory.

By theorem 4.3, PA cannot prove *all* first-order arithmetical truths. Nonetheless, in two important articles ([**Isa87**] and [**Isa92**]), Isaacson argues that *in some sense* the collection of theorems of PA can nonetheless be taken to coincide with the collection of arithmetical truths. This is known as *Isaacson's Thesis*. If it holds, then despites the incompleteness theorems there is a sense in which PA is *complete*.

For Isaacson, the notion of arithmetical truth is in part an epistemological notion. For a statement to be an arithmetical truth, it is not in general sufficient that it belongs to the formal or informal language of arithmetic and is true in the structure of the natural numbers. In addition, its truth must be "directly perceivable on the basis of our [...] articulation of our grasp of the structure of the natural numbers or directly perceivable from truths in the language of arithmetic which are themselves arithmetical" [**Isa87**, p. 217]. Isaacson argues that the statements the truth of which can be perceived in this way are precisely the theorems of PA. In this sense, then, PA is complete for arithmetical truth [**Isa87**, p. 222]. For seeing that any particular statement that is unprovable in PA is nevertheless true in the

[4]Proofs of these and other standard metamathematical results concerning Peano Arithmetic can be found in [**BBPJ02**].

natural number structure, insight is required into concepts that are not strictly speaking arithmetical. Such notions are called *higher-order concepts* by Isaacson. Examples of higher-order concepts are the notion of well-ordering, consistency of a formal system, provability in a formal system, and truth [**Isa92**, p. 96], and also the notion of second-order quantification [**Isa87**, p. 210]. For example, the principle of induction up to the small transfinite ordinal ε_0 is a truth which can be 'expressed' in the language of first order arithmetic (via coding).[5] But to see that this principle is true, insight is required into the notion of well-ordering, which is a set-theoretical and not a purely arithmetical concept. Therefore the principle of induction up to the ordinal ε_0 is *not*, in Isaacson's view, an arithmetical truth.

Most philosophers of mathematics do not follow Isaacson in seeing truth as an epistemological concept. Nonetheless, Isaacson's thesis can be formulated without involving a strong claim about the nature of truth. It can be taken simply to say that the class of arithmetical theorems that have *arithmetical proofs* coincides with PA. Then it is still very much a non-trivial claim. If it holds, then this is another reason why PA is of foundational importance.

4.1.2. Primitive Recursive Arithmetic and Hilbert's program. The weakest natural system of arithmetic that is studied in proof theory is called *Robinson Arithmetic* (after the mathematician Abraham Robinson, who first isolated it). It is the system Q that was briefly mentioned above: it is obtained by removing *all* induction axioms from PA.

An important extension of Q is obtained by adding, for every primitive recursive function f, the defining equations of f to the axioms of Q.[6] The resulting theory is called *Primitive Recursive Arithmetic*, and it is abbreviated as PRA. There is also a so-called *logic-free* presentation of PRA. This theory is formulated in a language *without quantifiers* but with (free) variables. The corresponding theory then of course also has no quantifier axioms or rules, but otherwise its axioms and rules are like those of the 'logical' presentation of PRA. The quantifier-free consequences of the 'logical' version of PRA are exactly those of the logic-free version.

Within mathematics, Hilbert distinguished between *finitary or concrete mathematics* on the one hand, and *infinitary or ideal mathematics* on the other hand [**Hil26**]. Finitary mathematics is the part of mathematics that can be given a concrete interpretation, whereas infinitary mathematics cannot. Roughly, finitary mathematics was thought of as a theory arithmetic that can be interpreted as a theory of concatenation of physical signs (sequences of strokes, for instance).[7] Infinitary mathematics is then just the rest of mathematics. Finitary mathematics was taken by Hilbert to be *true* when interpreted in the intended way. Ideal mathematics, in contrast, remains forever uninterpreted, and is therefore not true. Ideal mathematics has a purely *instrumental* role. Its aim is to prove finitary propositions that do not have feasible proofs in finitary mathematics. Hilbert thought that infinitary mathematics is *conservative* over finitary mathematics for finitary statements. That is, he thought that every finitary proposition that has a proof in infinitary mathematics, also has a proof in finitary mathematics. Moreover, he suspected that this alleged conservativeness could somehow be proved in finitary mathematics. To prove these two conjectures was the aim of *Hilbert's program*.

[5]The ordinal ε_0 is defined on p. 113.

[6]For this, $\mathcal{L}_{PA}$ is first extended by function constants for all primitive recursive functions.

[7]For a discussion of Hilbertian finitism, see [**Par07**, Chapter 7].

In order to make these two conjectures precise, one first has to clarify what *exactly* is meant by finitary mathematics, infinitary mathematics, and finitary proposition. Concerning infinitary mathematics, the details do not matter much: let us take standard Zermelo-Fraenkel set theory[8] as an explication of what is meant by infinitary mathematics. Several ways of making finitary mathematics precise have been proposed. Most philosophers of mathematics accept Tait's arguments[9] for the thesis that Hilbert's finitary mathematics should be identified with PRA, and that the class of finitary propositions is the class of Π_1 arithmetical sentences, which can be taken to be quantifier-free arithmetical formulas with free variables.

Gödel's incompleteness theorems showed that these two conjectures of Hilbert are false. Moreover, the Gödelian incompleteness phenomena are very robust: they are not very sensitive to *exactly* how the notions of finitary statement, finitary proof, infinitary proof are nailed down. Nonetheless, the concepts of finitary mathematics and of finitary statement are foundationally important. And this implies that PRA is a foundationally important theory.

4.1.3. Weak fragments of Peano Arithmetic. The system Q and the system $I\Sigma_0$ are *very* weak fragments of PA. There are some fairly robust fragments of PA that are stronger than these, but much weaker than $I\Sigma_1$: we briefly discuss two of them.

Edward Nelson believed that Hilbert was not cautious enough in his description of the 'safe' part of mathematics. He believed that we are not justified in believing that all primitive recursive functions are everywhere defined. In particular, he doubted that the exponentiation function is total. Not only that, but he believed that PA might well be outright inconsistent [**Nel11**].[10]

Nelson then sought to circumscribe what he took to be the *safe* part of mathematics [**Nel86**]. This led him to the first-order arithmetical theory that is called S^2_1. The list of axioms of S^2_1 is rather long, and not immediately intuitive at first side, so we do not give the list here.[11]

The natural way of coding syntax involves the exponentiation function. One of the noteworthy features of S^2_1 is that despite the fact that it does not prove the totality of exponentiation, it can still serve as a theory of syntax.

A slightly stronger system of arithmetic is the system of *Elementary Arithmetic* (EA). It is formulated in an extension of $\mathcal{L}_{PA}$ by a two-place function symbol e, which expresses the exponentiation function, So it is built into the language of EA, as it were, that exponentiation is a total function.

EA contains all the axioms of PA except its induction scheme, which is severely curtailed. Moreover, it contains natural recursive axioms that govern the behaviour of the exponentiation function:

(1) $\forall x : e(x, 0) = 1$;
(2) $\forall x \forall y : e(x, y + 1) = e(x, y) \cdot x$.

To conclude, EA contains a very restricted amount of mathematical induction. Say that *bounded quantification* is defined by taking $\forall x < t : \varphi$ to stand for $\forall x : x < t \to \varphi$, and $\exists x < t : \varphi$ to stand for $\exists x : x < t \wedge \varphi$, where x is not allowed to

[8]See Section 4.2.

[9]See [**Tai81**]. For a different proposal, see [**Kre60**].

[10]It would take us too far to go into Nelson's reasons for his concerns here. For a brief discussion of Nelson's worries, see [**Par07**, p. 303–305].

[11]The list, and a discussion of the axioms, can be found in [**HP93**, Section V.4].

occur in t. The *principle of bounded induction* is like Axiom PA8, except that all quantifiers in ϕ are required to be bounded. EA contains the principle of bounded induction.

One sense in which EA is weak is given by the fact that already PRA proves the consistency of EA ([**Avi03**, Theorem 2.1]). Another sense in which it is weak is that it cannot prove the totality of *iterated* exponentiation functions. In another sense, EA is remarkably strong. The vast majority of theorems of finitary number theory can be derived in EA. It has even been conjectured (by Harvey Friedman) that Fermat's Last Theorem can be proved in EA.[12]

4.1.4. Second-order Peano Arithmetic and some of its fragments. Second-order arithmetical theories are formulated in an extension of the language $\mathcal{L}_{PA}$ that is called $\mathcal{L}_{PA^2}$. This language is obtained from the language $\mathcal{L}_{PA}$ by adding to it second-order variables $(X, Y, \ldots)$ and second-order quantifiers. These second-order quantifiers are taken to range not over individual natural numbers, but over *properties* or *sets* of numbers.

Second-order arithmetical theories can also be seen as theories of *mathematical analysis*. Real numbers can be seen as certain infinite sets of natural numbers, e.g. as Dedekind cuts or as Cauchy sequences. Thus second-order systems of arithmetic allow the development of a theory of real numbers.

Full second-order arithmetic (PA^2) has full second-order logic as its background, including the unrestricted *second-order comprehension scheme*:

$$\exists X \forall y(X(y) \leftrightarrow \phi(y)) \qquad \text{for all } \phi \in \mathcal{L}_{PA^2} \text{ in which } X \text{ does not occur free.}$$

Moreover, PA^2 contains axioms PA1–PA7, and the scheme PA8 is replaced by the *second-order induction axiom*:

$$\forall X : [X(0) \wedge \forall y(X(y) \rightarrow X(s(y))] \rightarrow \forall y X(y).$$

One important *fragment* of PA^2 results from replacing the second-order induction axiom by the axiom scheme that is just like PA8 except that ϕ now ranges over all the formulas of $\mathcal{L}_{PA^2}$, and restricting the comprehension scheme to formulas that may contain free second-order variables but do not contain any second-order quantifiers. This restricted comprehension scheme is called *arithmetical comprehension*. The resulting sub-theory of PA^2 is called ACA.[13]

The theory ACA is nonetheless proof-theoretically stronger than PA. Formally, this is expressed using the notion of *proof theoretic conservativeness*, which is defined as follows:

Definition 4.6. Let S, S' be axiomatic theories such that $S \subseteq S'$, and $\mathcal{L}, \mathcal{L}'$ be languages such that $\mathcal{L} \subseteq \mathcal{L}'$. Then we say that S' is proof theoretically conservative over S for the language $\mathcal{L}$ if and only if for all $\varphi \in \mathcal{L}$:

$$S' \vdash \varphi \Leftrightarrow S \vdash \varphi.$$

We then have:

Theorem 4.7. *ACA is not proof-theoretically conservative over PA for the class of formulas of* $\mathcal{L}_{PA}$.

[12] These matters are discussed in [**Avi03**].

[13] See [**Hal14**, Definition 8.41, p. 107].

In particular, ACA proves the consistency of PA.

In ACA, the only sets of natural numbers that are recognised—i.e., postulated to exist, i.e., allowed in the range of the second-order quantifiers—are the ones that can be *defined* in first-order Peano Arithmetic. The general doctrine that the only collections of natural numbers that are recognised as existing are collections that are definable in a non-circular manner, is called *predicativism*. (The individual natural numbers themselves are taken as given by this doctrine.)

An example of an *impredicative definition* is the definition of a set of natural numbers A as the intersection of all sets of natural numbers that satisfy a given condition Φ. The reason is that this definition indeed is in a sense *circular*: the entities in terms of which A is defined—the domain of discourse over which the second-order quantifiers in this definition are ranging—contains the set A itself!

ACA is a basic system of predicative analysis, since first-order arithmetical definitions are clearly non-circular. Stronger theories of *predicative analysis* can be obtained by moving to higher-order languages (third-order, fourth-order,...) and iterating the procedure of recognising properties that can be defined by predicative means. For instance, we can add third-order variables $(\mathcal{X}, \mathcal{Y}, \ldots)$ and third-order quantifiers to $\mathcal{L}_{PA^2}$, and thus obtain the language $\mathcal{L}_{PA^3}$. Beside adding the obvious logical quantifier rules governing these new quantifiers, we can then extend the mathematical induction scheme of PA^2 to $\mathcal{L}_{PA^3}$, and add a predicative third-order comprehension axiom:

$$\exists \mathcal{X} \forall Y(\mathcal{X}(Y) \leftrightarrow \phi(Y)) \quad \text{for all } \phi \in \mathcal{L}_{PA^3} \text{ containing no third-order quantifiers.}$$

The resulting theory is called RA_2 (here 'RA' stands for *Ramified Analysis*), where ACA $= RA_1$.[14] Moreover, it can be shown that RA_2 is proof-theoretically stronger than ACA. Continuing in this way, the systems RA_n, for $n \in \mathbb{N}$ are obtained. Then RA_ω, the union of all the systems RA_n, for $n \in \mathbb{N}$, is a natural system of second-order arithmetic. Continuing in this way, taking unions at limit stages, this procedure can be iterated into the transfinite, yielding progressively stronger systems RA_α, with α ranging over ordinals.

Predicativism as a foundational stance goes back to the mathematician Hermann Weyl, who thought that the use of *impredicative definitions* is mathematically inadmissible. He took ACA to be predicatively acceptable, and believed that likewise the finite levely RA_n for $n \in \mathbb{N}$, are predicatively unobjectionable [**Wey18**]. Feferman argued that even for certain transfinite ordinals α, the systems RA_α, are predicatively acceptable theories of analysis. Indeed, if any first-order arithmetical definition is non-circular, then a definition of a set of numbers that quantifies only over sets that are unobjectionably definable in a first-order way, should likewise be regarded as non-circular. And so on. A question that will occupy us later is *how far* predicative definability can thus be extended in a predicatively acceptable manner.[15]

Shapiro has argued that the full second-order system PA^2 is a better framework for formalising our theory of the natural numbers than PA.[16] One of his main considerations is that in applications of mathematical induction in ordinary mathematical practice, we are usually not concerned with excluding quantification

[14] The hierarchy of systems to which this gives rise was first defined in [**Fef64**].

[15] See Section 6.2.3.

[16] See [**Sha91**].

over properties of natural numbers in the induction formula. Similarly, in ordinary mathematical practice we are usually not at all concerned with the question whether a property of natural numbers is *predicatively* definable at all. Hence, he also finds PA^2 a more suitable for the formalisation of our theory of the natural numbers than ACA.

Beside the theories RA_α, There are many systems of second-order arithmetic that are situated "between" ACA and full PA^2. In fact, today ACA is no longer considered to be one of the most important subsystems of PA^2. The statement that least *fixed points* of all positive inductive operators exist is clearly an impredicative statement, and hence unprovable in the systems RA_α. The fragment of second-order arithmetic that has this statement as its core second-order axiom is called ID_1. This system is impredicative (at least in the sense of Weyl and Feferman), but is somehow *close* to being predicative.

4.1.5. Ordinal notation systems and transfinite induction. In what follows, we will often be in a situation where the background setting is some first-order arithmetical theory. In such situations we will want to describe, within the object language, aspects of extending a starting formal theory by new principles transfinitely many times.

In order to refer to the stages of such revision processes, we have to talk about (countable) *transfinite ordinals* that mark the stages of these processes. *Direct* reference to these ordinals is not possible within the context of arithmetic, which after all only directly talks about the finite natural numbers. But by use of coding, it is possible to *simulate* reference to, and use of, transfinite ordinals within an arithmetical setting. Such coded systems for discussing transfinite ordinals are called *ordinal notation systems*. There are many such. But only some of them will play a role in the remainder of this book: powers of ω, *epsilon-numbers*, and the *Veblen hierarchy*. We describe these notation systems informally, leaving aside the details of how they are coded in arithmetic.

4.1.5.1. *Epsilon numbers.*

We take it as read that powers of ω can easily defined arithmetically. Using powers of ω in the base clause, the epsilon numbers ε_α are then inductively defined as follows:

Definition 4.8.

(1) $\varepsilon_0 = \sup\{\omega, \omega^\omega, \omega^{\omega^\omega}, \omega^{\omega^{\omega^\omega}}, \ldots\}$;
(2) $\varepsilon_{\alpha+1} = \sup\{\varepsilon_\alpha, \varepsilon_\alpha{}^{\varepsilon_\alpha}, \varepsilon_\alpha{}^{\varepsilon_\alpha{}^{\varepsilon_\alpha}}, \varepsilon_\alpha{}^{\varepsilon_\alpha{}^{\varepsilon_\alpha{}^{\varepsilon_\alpha}}}, \ldots\}$;
(3) $\varepsilon_\lambda = \sup\{\varepsilon_\beta \mid \beta < \lambda\}$.

So the epsilon function ε is a fast growing function: ε_{n+1} is in some sense *much* larger than ε_n. Moreover, the epsilon numbers are *fixed points*: it can easily be seen that $\varepsilon_0 = \omega^{\varepsilon_0}$, and that in general we have

$$\varepsilon_{\varepsilon_{\alpha+1}} = (\varepsilon_\alpha)^{\varepsilon_{\alpha+1}}.$$

4.1.5.2. *The Veblen hierarchy.*

We now turn to the (binary) *Veblen functions*, which form hierarchy of even faster growing arithmetically definable *normal functions*[17] on the ordinals. We start by defining what it means to be a normal function:

[17]'Normal' means continuous and strictly increasing.

Definition 4.9. A function f on the ordinals is *normal* if and only if f is strictly increasing and *continuous at limit stages*,

where being continuous at limit stages means that for every limit ordinal λ, we have

$$f(\lambda) = \sup\{f(\alpha) \mid \alpha < \lambda\}.$$

If φ_0 is any normal function on a segment of the ordinals, then for any ordinal $\alpha > 0$, φ_α is the function enumerating the common fixed points of φ_β for $\beta < \alpha$. These functions φ_α are then all normal.

We are only concerned with the special case where we take $\varphi_0(\alpha) = \omega^\alpha$. Then the resulting family of functions φ_α is called the *Veblen hierarchy*. The function φ_1 in this hierarchy is the same as the ε function: $\varphi_1(\alpha) = \varepsilon_\alpha$. The functions $\varphi_2, \varphi_3, \ldots$ are then increasingly fast growing functions on the ordinals.

The first ε number ε_0 will be of importance later: it is the least fixed point of φ_0, i.e., the least α such that $\omega^\alpha = \alpha$. In a similar way, $\varphi_2(0)$ is the least ordinal α, such that $\varepsilon_\alpha = \alpha$.

The function Γ enumerates the fixed points of the Veblen hierarchy, i.e., the ordinals α such that $\varphi_\alpha = \alpha$. Also the ordinal Γ_0, the *smallest* ordinal α such that $\varphi_\alpha = \alpha$ will be of importance later. This ordinal Γ_0 is called the *Feferman-Schütte ordinal*.

4.1.5.3. *Kleene's O.*

To conclude this section, we discuss an arithmetical notation system that is due to Stephen Kleene, and which can deal with even larger countable ordinals.[18] Moreover, since it is slightly less straightforward, we will be somewhat more explicit about the arithmetical coding.

This ordinal notation system is called *Kleene's $\mathcal{O}$*. We call $|a|$ the ordinal denoted by an ordinal notation a in Kleene's notation system $\mathcal{O}$ of arithmetical notations of ordinals. $\mathcal{O}$ is partially ordered by a relation $<_\mathcal{O}$. The set $\mathcal{O}$ and the relation $<_\mathcal{O}$ are simultaneously defined as the smallest set and relation for which the following holds:

Definition 4.10.

(1) $0 \in \mathcal{O}$, and $|0|$ is the ordinal number 0;
(2) If $a \in \mathcal{O} \wedge |a| = \alpha$, then $2^a \in \mathcal{O} \wedge |2^a| = \alpha + 1$ and $a <_\mathcal{O} 2^a$;
(3) Let $\{e\}$ is the e-th partial recursive function. If $\{e\}$ is total, the range of $\{e\}$ is a subset of $\mathcal{O}$, and for all n, $\{e\}(n) <_\mathcal{O} \{e\}(n+1)$, then $3 \cdot 5^e \in \mathcal{O}$ and for all n, $\{e\}(n) <_\mathcal{O} 3 \cdot 5^e$ and $|3 \cdot 5^e| = \lim_k |\{e\}(k)|$.
(4) $a <_\mathcal{O} b <_\mathcal{O} c \to a <_\mathcal{O} c$.

We have $a <_\mathcal{O} b$, for two ordinal notations a and b, if and only if $|a| < |b|$. The relation $<_\mathcal{O}$ induces a tree structure on $\mathcal{O}$ whereby $\mathcal{O}$ is well-founded. $\mathcal{O}$ branches only at limit ordinals, and at limit ordinals it branches countably infinitely.

The supremum of the ordinal numbers that are named in Kleene's $\mathcal{O}$, i.e., $\sup\{|a| : a \in \mathcal{O}\}$, is called the first non-constructible ordinal ω_1^{CK} ('omega-1-Church-Kleene'). It is the first ordinal number of which the order type is undefinable in the language of first-order arithmetic.

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 $P =$

[18] See [**Kle38**].

$\{\, b \,|\, b \ <_{\mathcal{O}}\ a\}$ 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 $|P| = \omega_1^{CK}$. 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 $\{\, b \,|\, b \ <_{\mathcal{O}}\ a\}$ is recursively enumerable.

4.1.5.4. *Transfinite induction.*

The *principle of transfinite induction* for α, where α is an ordinal, says that if you have any property Φ that is *progressive*, which means that, for every ordinal β if it holds for all ordinals $\gamma < \beta$, then Φ holds also for β, then Φ holds for α. It is not hard to see that for any α, the principle of transfinite induction up to α is true, for this follows immediately from the principle of transfinite induction that is provable in standard set theory. We have seen how we can simulate talk of small transfinite ordinals in an arithmetical setting, using ordinal notation systems. This allows us to formulate principles of transfinite induction *in the language of arithmetic*.

Let us now define principles of transfinite induction formally. We fix upon a natural notation system for ordinals and call it $\mathbf{O}$. Then both $\mathbf{O}$ and the ordering relation $\prec$ on ordinals defined by elements of $\mathbf{O}$ are definable in first-order arithmetic.

Definition 4.11 (Transfinite induction). Let A be a formula.

(1) Transfinite induction for A for $\alpha \in \mathbf{O}$, denoted as $TI(A, \alpha)$, is the formula
$$Prog(\lambda x A) \to A(t),$$
where t is a notation in $\mathbf{O}$ for α, and $Prog(\lambda x A)$ states that A is progressive along $\prec$, i.e.,
$$\forall x \in \mathbf{O}[\forall y \prec x A(y/x) \to A(x)].$$

(2) For a language $\mathcal{L}$ and ordinal $\alpha \in \mathbf{O}$, the schema of transfinite induction for α, $TI_{\mathcal{L}}(\alpha)$, is the collection of formulas
$$\{TI(A, \alpha) \mid A \in \mathcal{L}\},$$
and the schema of transfinite induction up to (but not including) α, $TI_{\mathcal{L}}(< \alpha)$, is the collection of formulas
$$\{TI(A, \beta) \mid A \in \mathcal{L} \wedge \beta < \alpha\}.$$

In general, the principle $TI_{\mathcal{L}}(< \alpha)$ is strictly weaker than the principle $TI(\alpha)$.

We will see later that there is a strong correlation between proof theoretic strength of a mathematical theory S on the one hand, and the amount of transfinite induction that S can prove on the other hand. The stronger S is, the larger the greatest ordinal α such that it can prove transfinite induction up to α; in this case, α is called the *proof theoretic ordinal* of S.

In this sense, the amount of transfinite induction that a system S can prove is often taken as a measurement of the mathematical strength of S. A typical example of this is Gerhard Gentzen's celebrated theorem:

Theorem 4.12.

(1) $PA \vdash TI_{\mathcal{L}}(< \varepsilon_0)$;
(2) $PA \nvdash TI_{\mathcal{L}}(\varepsilon_0)$.

This means that ε_0 is the proof theoretic ordinal of PA.

4.2. Set theory

Set theory is a universal framework for mathematics, in the sense that all proofs in all branches of mathematics can *in principle* be carried out in set theory. Of course this is not to be understood as a practical recommendation for carrying out, for instance, proofs in algebraic number theory, in the framework of set theory.

4.2.1. First-order. The language of first-order set theory ($\mathcal{L}_{ZFC}$) contains only one non-logical symbol: the relation symbol $\in$ (elementhood). The axioms of standard first-order set theory (ZFC) are the following:[19]

ZFC0 Existence
$\exists x\,(x = x)$
ZFC1 Extensionality
$\forall x\,\forall y\,(\forall z\,(z \in x \leftrightarrow z \in y) \rightarrow x = y)$
ZFC2 Foundation
$\forall x\,[\exists y\,(y \in x) \rightarrow \exists y\,(y \in x \wedge \neg\exists z\,(z \in x \wedge z \in y)]$
ZFC3 Set comprehension scheme / Separation
$\forall z\,\forall w_1, \ldots, w_n\,\exists y\,\forall x\,(x \in y \leftrightarrow x \in z \wedge \Phi)$ for each predicate Φ
ZFC4 Pairing
$\forall x\,\forall y\,\exists z\,(x \in z \wedge y \in z)$
ZFC5 Union
$\forall F\,\exists A\,\forall Y\,\forall x\,(x \in Y \wedge Y \in F \rightarrow x \in A)$
ZFC6 Replacement Scheme
$\forall A\,\forall w_1, \ldots, w_n\,[\forall x \in A\,\exists! y\,\Phi \rightarrow \exists Y\,\forall x \in A\,\exists y \in Y\,\Phi]$ for each predicate Φ

Using Axioms ZFC1–ZFC6, the concepts of subset ($\subseteq$), empty set ($\emptyset$), and ordinal successor (S) can be defined. In terms of these defined concepts, the remaining axioms of ZFC can be introduced:

ZFC7 Infinity
$\exists x\,(\emptyset \in x \wedge \forall y \in x\,(S(y) \in x))$
ZFC8 Power set
$\forall x\,\exists y\,\forall z\,(z \subseteq x \rightarrow z \in y)$
ZFC9 Choice
$\forall F\,[(\forall A \in F\,(A \neq \emptyset) \wedge \forall A, B \in F(A \cap B = \emptyset)) \rightarrow$
$\exists K\,\forall A \in F\,\exists! y \in A\,(y \in K)]$

Naively, one can take PA to describe the natural number structure. Equally naively—or perhaps even more so,—one can take ZFC to describe the set theoretic universe V: the collection of all sets. According to the *iterative conception of sets*,[20] the set theoretic universe V is somehow *generated* in ordinal stages by iterating the power set operation into the transfinite, starting from the empty set.[21] This philosophical view is inspired by the theorem of ZFC that says that V is stratified in layers, called *ranks* V_α, indexed by ordinals. For every set x, there is then a smallest ordinal α such that $x \in V_\alpha$.

---

[19] In the presentation of some of the axioms below, we here (exceptionally) make use of variables in capital letters as means of referring to *sets*. This is just done to increase readability of the axioms; they are not intended to be variables of higher order.

[20] See [**Boo71**].

[21] At limit stages, unions are taken of what is generated in all previous stages.

ZFC is a natural encapsulation of the general proof principles that are generally accepted in mainstream mathematics. In fact, in *most* of mainstream mathematics, not all proof principles in ZFC are used in their full strength. This is related to the fact that most mathematical objects that are of real interest to most mathematicians 'live' in rank $V_{\omega+5}$ or thereabouts. Despite this, few working mathematicians would object to any of the axioms of ZFC in their full generality. Thus, from a naturalist perspective, ZFC is of great foundational importance.

4.2.2. Second-order. The theory ZFC quantifies over sets. But, by Russell's famous diagonal argument, V cannot be a set:

THEOREM 4.13 (Russell). $ZFC \vdash \neg\exists y : \forall x(x \in y)$.

So V does not belong to ZFC's domain of discourse; ZFC does not formally recognise V as a mathematical entity. For similar reasons, ZFC does nor recognise the collection of all ordinal numbers Ord as a mathematical entity.

This does not necessarily mean that V does not exist as a mathematical entity. Collections of sets that cannot themselves be sets, such as the collection of all sets or the collection of all ordinals, are called *proper classes*. (Sets are then "improper" classes.) We can officially recognise proper classes of the mathematical world by adding second-order variables and second-order quantifiers to $\mathcal{L}_{ZFC}$, thus obtaining the language $\mathcal{L}^2_{ZFC}$ of second-order set theory. Then, in analogy with second-order theories of arithmetic, mathematical theories of sets and proper classes can be formulated: such theories are called *class theories*.

The class theory that is the natural analogue of full second-order number theory (PA^2) is called MK (for *Morse-Kelly*), or also ZFC^2. We obtain it by formulating the axioms of ZFC in the extended language $\mathcal{L}^2_{ZFC}$, replacing the replacement scheme ZFC6 and the set comprehension scheme ZFC3 by its second-order universal closures, and adding the full class comprehension scheme:

$$\exists X\forall y : y \in X \leftrightarrow \Phi(y) \qquad \text{for every } \Phi \text{ not containing } X \text{ free .}$$

The theory ZFC^2 is much stronger than ZFC:

THEOREM 4.14. ZFC^2 *is proof-theoretically non-conservative over ZFC for* $\mathcal{L}_{ZFC}$.

Nonetheless, there are elementary natural set theoretic questions that even ZFC^2 cannot decide, such as Cantor's famous *Continuum Hypothesis* (CH):

DEFINITION 4.15 (CH). There are no collections of sets of natural numbers C such that the cardinality of C is strictly greater than the cardinality of $\mathbb{N}$ but strictly smaller than the cardinality of $\mathcal{R}$.

THEOREM 4.16 (Gödel, Cohen). *CH is independent of* ZFC^2.

Because its comprehension scheme postulates the existence of classes that are defined by quantifying over the collection of all classes, the theory MK is an *impredicative* theory of classes. Even if one accepts impredicatively defined sets, one might feel uneasy about impredicatively defined proper classes. In this case, one may opt for a predicative theory of proper classes. The standard predicative theory of proper classes is called NBG (*Von Neumann-Bernays-Gödel*). It is obtained by restricting the full class comprehension scheme to formulas $\Phi \in \mathcal{L}^2_{ZFC}$ that contain no bound variables. This restriction to predicative comprehension makes NBG proof-theoretically much weaker than ZFC^2:

THEOREM 4.17. *NBG is proof-theoretically conservative over ZFC for* $\mathcal{L}_{ZFC}$.

One theory of classes that is intermediate between NBG and ZFC^2 will play some role later. This theory is called ECA.[22] It consists of NBG except that instead of the second-order quantified *axioms* of Separation and Replacement, it has the full second-order *schemes* of Separation and Replacement:

$$\forall z\, \forall w_1, \ldots, w_n\, \exists y\, \forall x\, (x \in y \leftrightarrow x \in z \wedge \Phi) \text{ for } \Phi \text{ any formula of } \mathcal{L}^2_{ZFC},$$

$$\forall a\, \forall w_1, \ldots, w_n\, [\forall x \in a\, \exists! y\, \Phi \rightarrow \exists Y\, \forall x \in a\, \exists y \in Y\, \Phi] \text{ for } \Phi \text{ any formula of } \mathcal{L}^2_{ZFC}.$$

Where ZFC^2 stands to ZFC as PA^2 stands to PA, ECA stands to ZFC as ACA stands to PA.

The foundational significance of class theory derives mostly from the use that is made of it in set theory itself. Not much use is made of proper classes in mainstream mathematics. But in set theory, we do find uses of proper classes. For instance, we find them in the investigation of large cardinals, which are sets that are too large for ZFC prove their existence.[23] For some such uses NBG suffices. For others, stronger class theories come in useful.

4.2.3. Large cardinals. Large cardinal axioms, also known as *strong principles of infinity*, are axioms hat posit the existence of sets that are very large, and that cannot be proved to exist in ZFC. Such principles have been intensively investigated since the 1940s. A distinction is made between *small large cardinal* axioms, *large large cardinal* axioms, and axioms that posit the existence of sets that, were they to exist, would be very large, but whose existence is incompatible with ZFC. The standard textbook on this subject is [**Kan94**], on which we rely heavily in what follows. This subsection is a whistle tour of some of the main large cardinal principles.

The large cardinal axioms that are known are linearly ordered by strength. For (almost) any large cardinal axioms LC and LC' that are known, we know either that LC implies LC', or *vice versa*, or both.

4.2.3.1. *Small large cardinals.* The small large cardinal Axioms are those that are compatible with every set being *constructible*. Here the constructible universe L is defined by transfinite recursion as follows:[24]

DEFINITION 4.18.

(1) $L_0 = \emptyset$;
(2) $L_{\alpha+1}$ contains all and only the sets y that consist of the sets z such that $L_\alpha \models \varphi(a_1, \ldots, a_n, z)$ for some $\varphi(x_1, \ldots, x_n, y) \in \mathcal{L}_{ZFC}$ and $a_1, \ldots, a_n \in L_\alpha$;
(3) $L_\gamma = \bigcup_{\beta<\gamma}(L_\beta)$;
(4) $L = \bigcup_{\alpha \in Ord}(L_\alpha)$, where Ord is the class of all ordinals.

Thus in the recursive definition of L we take the ordinals as given, at successor stages we take definable subsets, and at limit stages we take unions.

L is then a proper class size structure of sets. It can be shown that L is the *minimal* structure containing all ordinals that makes the axioms of ZFC true. The

[22] See [**Fuj23**, p. 151].

[23] We will look at some of them in the next subsection.

[24] The constructible sets were first defined by Gödel. For a good discussion of the properties of L, see [**Dev84**].

statement that the universe V consists of all and only the constructible sets can then be expressed as a first-order set theoretic statement, which is often abbreviated as $V = L$. The statement $V = L$ is *independent* of the axioms of ZFC.

The weakest important modest large cardinal that we encounter is the *Axiom of (strongly) Inaccessible Cardinals* (IC), which can be formulated as follows [**Kan94**, p. 19]:

AXIOM 4.19 (IC). There is a cardinal κ such that $(V_\kappa, \epsilon, V_{\kappa+1}) \models ZFC^2$.

We know that $(V, \epsilon, \mathcal{C}) \models ZFC^2$. So the structure $(V_\kappa, \epsilon, V_{\kappa+1})$ "reflects" the property of making all of making all of ZFC^2 true.

A stronger important small large cardinal principle is the *Axiom of Weakly Compact Cardinals* (WCC):[25]

AXIOM 4.20 (WCC). There is a cardinal κ such that for every Π^1_1 formula (with a second-order parameter) φ and for every $A \subseteq V_\kappa$: if $(V_\kappa, \epsilon, V_{\kappa+1}, A) \models \varphi$, then there is an $\alpha < \kappa$ such that $(V_\alpha, \epsilon, V_{\alpha+1}, A \cap V_\alpha) \models \varphi$,

where in the latter structure V_α serves as the interpretation of the first-order variables, $V_{\alpha+1}$ serves as the interpretation of the second-order variables, and $A \cap V_\alpha$ serves as the interpretation of the second-order parameter. (The former structure then specifies interpretations of variables and parameters in an analogous manner.)

Thus the level V_κ "reflects" all Π^1_1 sentences with second-order free variables. We will later see that this is indeed a paradigmatic example of a set theoretic reflection principle.

All weakly compact cardinals are (strongly) inaccessible, but the converse is not the case. Thus Axiom 4.20 is properly stronger than Axiom 4.19.

4.2.3.2. *Large large cardinals.* Large large cardinals are those that are incompatible with the assumption that every set is constructible. Few doubt the truth of any of the well-known small large cardinal Axioms. The truth of large large cardinals is more controversial, and doubt increases as we move up in the hierarchy of strength of large large cardinal Axioms.

The first important large large cardinal Axiom that we encounter has a measure-theoretic origin:

DEFINITION 4.21. A *measure* over a set S is a function $m : \mathcal{P}(S) \to [0, 1]$ such that:

(1) $m(S) = 1$
(2) $m(\{x\}) = 0$ for all $x \in S$
(3) for pairwise disjoint $\{X_n \subseteq S \mid n < \omega\}$, we have

$$m(\bigcup_n X_n) = \Sigma_n m(X_n).$$

It is easy to see that ω is a *measurable set*. But the existence of *uncountable* measurable sets cannot be proved in ZFC: it is a large cardinal property. This is a motivation for the *Axiom of Measurable Cardinals* (MC):[26]

[25]There are a number of conceptually different but mathematically equivalent definitions of weakly compact cardinals. The definition that is given here is chosen because of its direct connection with the subject matter of reflection.

[26]For an introduction to the concept of measurable cardinal, see [**Kan94**, Chapter 1, Section 2].

Axiom 4.22 (MC). There are uncountable measurable cardinal numbers.

It can be proved that all measurable cardinals are weakly compact. Moreover, measurable cardinals are indeed large large cardinals:

Theorem 4.23. $ZFC + (MC) \vdash V \neq L$.

Almost all large large cardinal axioms, and some small large cardinal axioms,[27] can be expressed as *elemenary embedding principles*, which are defined as follows.[28]

Definition 4.24. An *inner model* of V of a theory S is a transitive substructure of V that makes S true and that contains all ordinals.

Typically, S will be a fairly strong theory of sets, such as ZFC or ZFC^2. A standard example of an inner model for ZFC is Gödel's constructible universe L.

Definition 4.25.

(1) A first-order *elementary embedding* from V into an inner model M is a proper class size bijective function j such that for all formulas $\Phi(x_1, \ldots, x_n) \in \mathcal{L}_{ZFC}$, we have:
$$V \models \Phi(x_1, \ldots, x_n) \Leftrightarrow M \models \Phi(j(x_1), \ldots, j(x_n)).$$
This is denoted as $j : V \mapsto_1 M$.

(2) A *second-order* elementary embedding from V into M is defined as a class function such that for all $\Phi \in \mathcal{L}^2_{ZFC}$:
$$V \models \Phi(x_1, \ldots, x_n, Y_1, \ldots, Y_m) \Leftrightarrow M \models \Phi(j(x_1), \ldots, j(x_n)), j(Y_1), \ldots, j(Y_m).$$
This is denoted as $j : V \mapsto_2 M$.

An embedding function j is said to be *non-trivial* if it is not the identity function on the ordinals. For every embedding function j and for every ordinal α, $j(\alpha) \geq \alpha$. If j is non-trivial, then there must be a smallest ordinal κ that is moved by j. This ordinal κ is then said to be the *critical point* of j; this ordinal typically has large cardinal properties.

Embedding principles thus are axioms that postulate the existence of non-trivial embedding functions, often with certain specific properties. The strength of an embedding principle tends to be positively correlated with the extent to which it forces the inner model M to resemble the set theoretic universe V.

The function j in embedding principles is then a proper class. So elementary embedding principles are class theoretic statements. Nonetheless, the strength of embedding principles does not crucially depend on this second-order feature, since embedding principles relating instead (set-sized) *initial fragments* V_α of V to "inner models" M_α that contain all ordinals of V_α still have the intended large cardinal strength. The latter are of course first-order principles.

The following Theorem shows that the Axiom of measurable cardinal can be regarded as an embedding principle:

Theorem 4.26. *An uncountable cardinal κ is measurable if and only if there is an inner model M and an embedding j such that $j : V \mapsto_1 M$ with critical point κ.*

[27] Axiom 4.20 is a case in point.

[28] For an introduction to the relation between large cardinal axioms and elementary embeddings, see [**Kan94**, Chapter 1, Section 5].

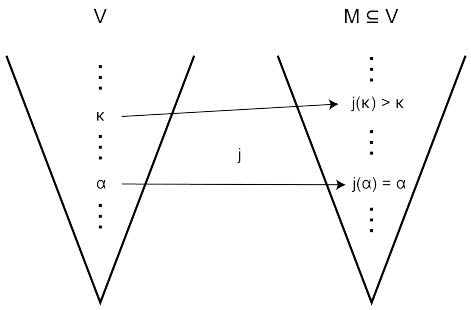

FIGURE 1

By imposing additional conditions on j and on M, stronger embedding principles are obtained. In particular, in stronger embedding principles, M looks more and more like V itself.

Continuing our discussion of the large cardinal hierarchy, we next arrive at the *Axiom of 1-Extendible Cardinals* (EC), which can be seen as a second-order strengthening of Axiom 4.22 ([**Kan94**, p. 311]):[29]

DEFINITION 4.27 (EC). A cardinal κ is *1-extendible* if and only if there is an inner model M and an embedding j such that $j : V \mapsto_2 M$ with critical point κ.

Again all 1-extendible cardinals are measurable cardinals, but not *vice versa*.

In the same way that one distinguishes between sets and classes, one can make a type distinction between classes and *hyperclasses*. Then one can consider V not only with its classes, but also with its hyperclasses, and consider embedding functions j that are not only elementary for sentences of class theory but even for sentences of hyperclass theory (i.e., embedding functions that are 3-elementary), one can in those terms formulate an analogue of Axiom 4.27: this stronger axiom is the Axiom of 2-Extendible Cardinals. And by climbing up further through the type theoretic hierarchy in the same way, one arrives at the Axiom of α-Extendible Cardinals, for any given ordinal α. A cardinal number that has the property of being α-Extendible for *every* ordinal α is called an *Extendible Cardinal*. Being an Extendible Cardinal is (in some informal sense) a much stronger large cardinal property than being α-Extendible for some given α.

Another important concept in large cardinal theory is the property of *supercompactness*, which again has an elementary embedding characterisation [**Kan94**, p. 298]:

DEFINITION 4.28 (SC). A cardinal κ is γ-supercompact (SC) if there is an inner model M and an embedding $j : V \mapsto_1 M$ such that κ is the critical point of j, $\gamma < j(\kappa)$, and M is closed under sequences of length γ of elements of M.

Analogous to the case of extendibility, a cardinal is said to be *supercompact* if it is γ-supercompact for every ordinal γ. The least supercompact cardinal is larger than the least α-extendible cardinal (for every ordinal α), whereby supercompactness is a very strong large cardinal property, but it is smaller than the least extendible cardinal.

[29]The concept of extendible cardinal traces back to [**Rei74**].

To conclude, we arrive at one of the strongest large large cardinal Axioms that is believed to be consistent with ZFC, namely *Vopenka's Principle* (VP). This Axiom is not formulated as an embedding principle ([**Kan94**, p. 335]):

Definition 4.29 (VP). For every proper class of first-order structures $\mathcal{C}$, there are structures $A, B \in \mathcal{C}$ such that A is isomorphic to a substructure of B,

where a structure is just what one would expect, i.e., a set with some operations on it.

(VP) is a very strong large cardinal principle because it entails the existence of many extendible cardinals, for instance.

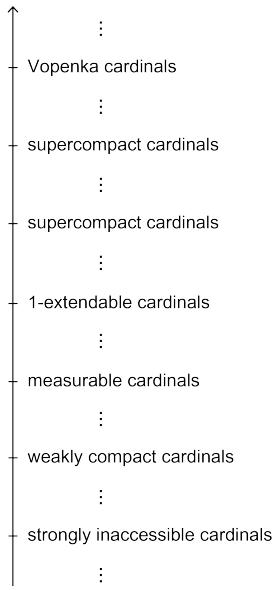

Figure 2

Many set theorists find all the above large cardinal Axioms plausible. Moreover, large cardinal Axioms are proof-theoretically strong. In particular, they decide many questions about small initial segments of V. *Gödel's program* consists in solving as many set theoretic questions that are independent of ZFC using large cardinal Axioms as possible by using large cardinal Axioms. This program has so far been *moderately* successful, as is witnessed by the history of research on the Continuum Hypothesis. It has been shown, by essential appeal to large cardinal Axioms, that at least some natural weakening of the Continuum Hypothesis, which is nevertheless highly nontrivial, is true ([**Woo01**]). However, it has also proved to be highly unlikely that large cardinal Axioms can ever decide the full Continuum Hypothesis, even though CH is of course a question about the rank $V_{\omega+2}$.

4.2.3.3. *Choiceless cardinals.* Reinhardt observed that natural ultimate limit of this process (of embedding V into inner models M that look more and more like V) is to postulate a non-trivial embedding from V into *itself*:

Axiom 4.30 (R). There is a non-trivial elementary embedding from V into V.

But this is was then soon found to be incompatible with ZFC [**Kun71**]:

THEOREM 4.31. *ZFC* $\vdash \neg\exists$ *non-trivial* $j : V \mapsto_1 V$.

It has been observed that the proof of Theorem 4.31 makes *essential* use of the Axiom of Choice. Indeed, a rich structure theory of "choiceless cardinals" is currently being developed in the context of ZF (*without* the Axiom of Choice). In particular, various strengthenings of Axiom R have been proposed, and in the absence of the Axiom of Choice, they appear (so far!) to be consistent. This gives the appearance of there being a whole *realm* of cardinals beyond the "choicy" cardinals.

4.3. Probability

In this concluding section, we begin by formalising Kolmogorov's classical theory of probability. However, the main aim of the concluding section of this chapter is to present two formal theories of probability. The first of these theories intends to capture concepts of finitely additive probability. The second theory contains a version of the principle of σ-additivity. In a later section,[30] we will use these theories as background for our discussion of *probabilistic reflection principles*.

4.3.1. Kolmogorov. We start with the standard or 'classical' theory of probability, which we will call K. It was first formulated by Kolmogorov in [**Kol33**]. Usually, Kolmogorov's theory is presented only semi-formally.[31] Here, we give a precise axiomatisation of it.

Probability values are real numbers between 0 and 1. So the intended domain of discourse is $\mathbb{R}$, and the language of probability theory includes a first-order language $\mathcal{L}_R$ that contains symbols for certain elementary relations and operations on the real numbers (such as $\leq$, for instance). In particular, we will assume that $\mathcal{L}_R$ contains a predicate (N) that holds of all and only the natural numbers. This allows us to assume that *formulas* of the language $\mathcal{L}_R$ belong, coded as natural numbers, to the domain of discourse. The background language may of course also contain empirical predicates and constants. But they will not play a role in what follows—*Schmieröl!*—so we will pretend that some such are included in $\mathcal{L}_R$, but make no specific assumptions about them.

The formal language of probability theory $\mathcal{L}_{\mathrm{Pr}}$ takes probability to be a property of *closed sentences*. It therefore contains a function symbol Pr, which takes codes of sentences (i.e., natural numbers) as arguments, and delivers real numbers as function values. We set $\mathcal{L}_{\mathrm{Pr}} = \mathcal{L}_R \cup \{\mathrm{Pr}\}$.

Kolmogorov's probability theory is a *typed* theory, in the sense that Kolmogorov's probability axioms are about statements in which the concept of probability does not occur. So they are *about* sentences of a background language $\mathcal{L}_R$ in which the probability symbol Pr does not occur, and are stated *in* the language $\mathcal{L}_{\mathrm{Pr}}$.

Kolmogorov's probability theory (K) is then formulated in the language $\mathcal{L}_{\mathrm{Pr}}$. It contains some standard theory R of the real and the natural numbers: the exact

[30] See Section 6.7.

[31] This is primarily due to the fact that the axiom that 'all *necessary* events have probability 1' is introduced without giving a formal theory of the concept of necessity.

details do not matter in what follows.[32] In addition, K contains the following axioms governing the probability predicate:

K1 Pr is a function such that $\forall \phi \in \mathcal{L}_R : 0 \leq \Pr(\phi) \leq 1$;
K2 $\forall \phi, \psi \in \mathcal{L}_R : \Pr(\phi \vee \psi) = \Pr(\phi) + \Pr(\psi) - \Pr(\phi \wedge \psi)$;
K3 $\forall \phi \in \mathcal{L}_R : \Pr(\neg\phi) = 1 - \Pr(\phi)$;
K4 $\forall \phi(x) \in \mathcal{L}_R : \Pr(\exists x \in N : \phi(x)) = \lim_{n\to\infty}(\phi(1) \vee \phi(2) \vee \ldots \vee \phi(n))$;
K5 For all $\phi(x) \in \mathcal{L}_R$:

$$\frac{\vdash \phi}{\vdash \Pr(\phi) = 1}.$$

Conditional probability can be defined in the usual way:

DEFINITION 4.32. If $\Pr(\psi) \neq 0$, then

$$\Pr(\phi \mid \psi) = \frac{\Pr(\phi \wedge \psi)}{\Pr(\psi)},$$

and $\Pr(\phi \mid \psi)$ is undefined otherwise.

From David Lewis' triviality result, which we state somewhat imprecisely here, we know that conditional probability cannot be equated with probability of a material conditional:

LEMMA 4.33 (Lewis). *For most probability functions p, there are sentences $\phi, \psi \in \mathcal{L}_R$ such that*

$$p(\phi \mid \psi) \neq p(\psi \to \phi).$$

Axioms K1–K3 are *basic* axioms of Kolmogorov's theory. Axiom K4 is a version of the principle of σ-additivity. The schematic rule 'Necessitation' rule K5 approximates the informal Kolmogorov axiom that *all necessary statements have probability 1*. An important subsystem of K, which we call K^-, is obtained by removing the principle of σ-additivity (K4) from K. The theory K^- is the standard system of *finitely additive probability*.

PROPOSITION 4.34. *The theories K and K^- are consistent and have standard models,*

where 'standard' means that the domain is $\mathbb{R}$, and the symbols $+, \cdot, \ldots$ are interpreted in the normal way.

Basic laws of classical probability theory can be formally proved in the system K. For instance, we can prove the following useful *substitution rule*:

PROPOSITION 4.35. *For all $\phi, \psi \in \mathcal{L}_R$:*
if $\mathrm{K}^{(-)} \vdash \phi \leftrightarrow \psi$, *then* $\mathrm{K}^{(-)} \vdash \Pr(\phi) = \Pr(\psi)$.

PROOF. We reason in $\mathrm{K}^{(-)}$ from a proof (in $\mathrm{K}^{(-)}$) of $\phi \leftrightarrow \psi$. By K5, we infer that $\Pr(\phi \leftrightarrow \psi) = 1$. A simple calculation in $\mathrm{K}^{(-)}$ shows that for all $\phi, \psi \in \mathcal{L}_R$, $\mathrm{K}^{(-)}$, using K2 and K3, proves that $\Pr(\phi \to \psi) = 1 \to \Pr(\phi) \leq \Pr(\psi)$. So the fact that $\Pr(\phi \leftrightarrow \psi) = 1$ allows us to conclude, in $\mathrm{K}^{(-)}$, that $\Pr(\phi) = \Pr(\psi)$. □

In fact, it is easy to see that against the background of the axioms K1–K4, Proposition 4.35 is *equivalent* to the Necessitation rule K5.[33]

[32] Details are given in [**Las09**, Section 3.2].

[33] Laskey's axiom system for Kolmogorov probability in [**Las09**] is therefore equivalent to our system K: her theory is like ours, except that she has the inference rule Proposition 4.35 instead of our inference rule K5.

4.3.2. Typefree probability. We have seen that the theories K and K^- are *typed*: they contain no principles that impose restrictions on the probability of statements that themselves contain the concept of probability. We can turn K and K^- into *untyped* or *typefree* theories K_u and K_u^- by simply removing the type restrictions that are built into their axioms. The procedure of 'untyping' axiom K2, for instance, yields the following principle:

$\mathrm{K2}_u$ $\forall \phi, \psi \in \mathcal{L}_{\mathrm{Pr}} : \mathrm{Pr}(\phi \vee \psi) = \mathrm{Pr}(\phi) + \mathrm{Pr}(\psi) - \mathrm{Pr}(\phi \wedge \psi)$

The motivation for the principles of K is supposed to extend to form a motivation of the principles of K_u. This holds also for the Necessitation rule. The Necessitation rule of K is motivated by arguing that provable mathematical sentences are not only true, but even *necessarily* true. Similarly, the Necessitation rule of K_u is motivated by arguing that provable statements (possibly about the notion of probability) express conceptual truths about mathematics and probability, and such conceptual truths are necessarily true. The motivation is then completed by Kolmogorov's claim that necessary truths have probability 1.

For K_u and K_u^-, an analogue of Proposition 4.35 can be proved (and in the same way):

Proposition 4.36.

For all $\phi, \psi \in \mathcal{L}_{\mathrm{Pr}}$*: if* $\mathrm{K}_u^{(-)} \vdash \phi \leftrightarrow \psi$*, then* $\mathrm{K}_u^{(-)} \vdash \mathrm{Pr}(\phi) = \mathrm{Pr}(\psi)$.

Kolmogorov's axioms are often taken to govern both concepts of subjective and concepts of objective probability. We want to do the same for our systems K_u and K_u^- of typefree probability. We will be especially occupied with the interpretation of Pr as *typefree rational subjective probability*.

The distinction between typed and untyped theories carries over to *truth theories* in the obvious way. The history of axiomatic truth theory has taught us that care should be taken when 'untyping' typed truth theories. The same holds for untyping probability theory. Indeed, just as the diagonal lemma allows us to construct liar sentences for truth theories, the diagonal lemma allows us to construct *probabilistic liar sentences* for axiomatic theories of probability, i.e., sentences λ_p such that for instance

$$\lambda_p \leftrightarrow (\mathrm{Pr}(\lambda_p) < 1).$$

Thus the sentence λ_p "says of itself" that it has probability < 1.

The good news is ([**CHL22**, Theorem 3]):

Theorem 4.37. *The theory* K_u^- *is arithmetically conservative over the background theory and has standard models.*

This does not mean that K_u^- is an *uncontroversial* basic system of self-referential subjective probability. Indeed, in the area of type-free subjective probability, there is very little established common ground. That being said, I do know of any theorems of the system K_u^- that are clearly objectionable.

However, when we add σ-additivity, we come dangerously close to inconsistency ([**CHL22**, Theorem 1]):

Theorem 4.38. *The theory* K_u *is consistent but* ω*-inconsistent.*[34]

[34]A theory S is ω-inconsistent if S proves $\exists x \in \mathbb{N} : \varphi(x)$ for some formula φ, but at the same time for every $n \in \mathbb{N}$ proves $\varphi(n)$.

For this reason, we will later, in our investigation of probabilistic reflection principles, mostly adopt the system K_u^-, rather than K_u, as background theory.

There is a sense in which we do not have much freedom to extend even K_u^-. Simple diagonal arguments show at even the probabilistic versions of the weak principles[35] of *positive introspection* (PI) and *converse positive introspection* (CPI) cannot be consistently added to K_u^- ([**CHL22**, Proposition 3, Proposition 5]):

PROPOSITION 4.39.

(1) *The positive introspection principle*

$$\Pr(\varphi) = 1 \to \Pr(\Pr(\varphi) = 1) = 1$$

cannot be consistently added to K_u^-*;*

(2) *The converse positive introspection principle*

$$\Pr(\Pr(\varphi) = 1) = 1 \to \Pr(\varphi) = 1$$

cannot be consistently added to K_u^-*.*

In [**SH22**], the notion of (type-free) *justified belief*[36] is axiomatised. The properties of the resulting basic system of justified belief are similar to those of K_u^-. In particular, adding introspection principles to this basic system typically results in inconsistency. These results can perhaps be taken to cast doubt on the ultimate coherence of certain scenarios of *perfectly* reflective agents, which were discussed in Section 3.9.

[35] See Section 3.8.

[36] The notion of justified belief was discussed in Section 1.2.

CHAPTER 5

Axiomatic truth and deflationism

In this Chapter, we review basic facts about theories that describe how truth can be used as a primitive concept in our reasoning. Such theories are developed in the field of axiomatic truth theory.

In axiomatic truth theory, truth is conceived of as a property that some sentences (of a given language) have, and that other sentences lack. Truth is then formally expressed by a primitive predicate that is governed by *truth axioms*. These axioms are considered against the background of axioms that describe the syntax of sentences. For instance, a truth theory might claim, of a given sentence φ, that φ is true if its negation is not true. Then the theory must be able to recognise, for instance, the syntactical fact that from prefixing a negation sign to a sentence, another sentence results.

Theories of syntax can be coded in first-order arithmetical theories. Whereas the former are investigated in branches of linguistic mostly semi-formally, arithmetical theories have been intensively investigated as fully formal theories. For this reason, the background theory of syntax is in axiomatic truth theory mostly expressed as a formal theory of arithmetic. We have seen in the previous chapter that there are many first-order theories of arithmetic that differ in mathematical strength. For definiteness, we will in this chapter mostly take one fixed weak arithmetical theory as the background theory of syntax. But it turns out that almost all of the basic theoretic and meta-theoretic features of axiomatic truth theories are insensitive to mild perturbation of the background syntax theory. We will see that it is only when a second-order theory (or set theory) is taken as background theory over which the truth axioms are formulated, that the truth axioms must take slightly different forms.

Axiomatic truth theories can be classified along two dimensions. On the one hand, there is a distinction between disquotational and compositional truth theories. On the other hand, there is a distinction between typed and untyped truth theories.

Disquotational truth theories roughly contain only axioms of the form

$$\text{`}A\text{' is true if and only if } A.$$

Compositional truth theories, on the other hand, contain axioms stating that truth commutes with the familiar logical particles. For instance, a compositional truth theory may contain an axiom that says that any conjunction is true if and only if both of its conjuncts are true.

In typed truth theories, the truth axioms attribute the property of truth only to sentences that do not themselves contain the truth predicate. The purpose of this stratagem, which is due to Tarski,[1] is to immunise against the argument of the

[1] See [**Tar83**].

liar paradox. Type-free truth theories, on the other hand, attribute truth also to certain sentences that themselves contain the concept of truth, whilst still guarding against liar-like paradoxes.

The debate about what might be our best axiomatic theory of truth—if there is such a thing—is intimately intertwined with the contemporary *philosophical* debate about the *nature* of the property that is expressed by the truth predicate. Substantivists about truth claim that the truth predicate expresses a substantial and philosophically pivotal property. Deflationists, on the other hand, deny this and maintain that the truth predicate merely expresses a "metaphysically light" property, the only function of which is to express and reason with certain generalisations that we could not otherwise express or reason with.

In this Chapter I will mostly take PA as a background theory over which truth theories are formulated, simply because this is what is mostly done in recent literature about axiomatic truth and deflationism.

5.1. Disquotational theories

We have seen in the previous chapter that for every (sufficiently strong) language, the notion of truth is undefinable in that language.[2] So if we want to make use of the notion of truth in our reasoning, we have to add a new predicate for it explicitly to the language. For instance, if we want to use the notion of truth in a discussion of arithmetic, we have to add it as a new primitive predicate (T) to the language. Thus we have to extend the language $\mathcal{L}_{PA}$ with the predicate T, yielding a larger language $\mathcal{L}_T$.

5.1.1. Unrestricted disquotation. We have to add axioms that regulate the behaviour of the truth predicate in order to ensure that it behaves properly. For example, from the assumption that the Twin Prime Conjecture is true, we want to derive the Twin Prime Conjecture. In this context, a fundamental observation is that truth is a device of semantic ascent and of semantic descent.[3] It seems that from $T(\phi)$, we ought to be allowed to infer ϕ, and, conversely, from ϕ, we ought to be able to derive $T(\phi)$. This *disquotational* nature of the truth predicate is captured in the *Tarski-biconditionals*, which are the statements of the form:

$$T(\phi) \leftrightarrow \phi.$$

Truth theories in which the principles regulating the logical behaviour of the truth predicate are Tarski-biconditionals are called *disquotational truth theories*.

However, when *all* Tarski-biconditionals are added to a minimally strong background theory (such as Robinson Arithmetic (Q)), then the argument of the *liar paradox* shows that the resulting theory (call it L), is inconsistent. In fact, something slightly stronger can easily be proved ([**KM60**]):

LEMMA 5.1. *The theory that consists of PA, formulated in the extended language $\mathcal{L}_T$, plus the T-Out scheme*

$$T(\varphi) \rightarrow \varphi,$$

and the Necessitation rule

$$\frac{\vdash \phi}{\vdash T(\phi)},$$

[2] See Theorem 4.5.

[3] See [**Qui86**, Chapter 1].

is inconsistent.

PROOF. We reason in this theory. By the diagonal lemma, there is a liar sentence L such that $PA \vdash L \leftrightarrow \neg T(L)$. Suppose $T(L)$. By T-Out, this yields L, which is equivalent to $\neg T(L)$. So we have reached a contradiction, and can reject the assumption, i.e., we have $\vdash \neg T(L)$, whereby $\vdash L$. Then by Necessitation, we also have $\vdash T(L)$, whereby we have reached a contradiction. □

This Theorem was intended to generate liar-like *intensional paradoxes*. After all, if we interpret the predicate T not as truth but as necessity, for example, then, intuitively, Necessitation and T-Out appear plausible basic principles. Nonetheless, the resulting theory is inconsistent. At any rate, Theorem 5.1 shows *a fortiori* that the unrestricted Tarski-biconditionals cannot all be consistently added to a background theory of syntax.

In response to this problem, Horwich proposed that our best theory of truth should collect as many jointly consistent Tarski-biconditionals as possible ([**Hor98**, p. 42]). Unfortunately, this strategy does not work. Not only is there no *unique* maximal consistent set of Tarski-biconditionals; in addition, no maximal consistent set of Tarski-biconditionals is recursively enumerable ([**McG92**]). This shows that the task of formuilating an optimal disqutational truth theory is non-trivial.

5.1.2. Typed disquotational truth. The Tarski-biconditional that is used in the argument for the liar paradox is a sentence that itself contains the truth predicate. Tarski realised that if only those Tarski-biconditionals that do not themselves contain the truth predicate are added as axioms regulating the truth predicate, then no inconsistency can be derived. The resulting theory is called TB. To be precise, it consists of PA, with the new truth predicate T allowed in the induction axiom, plus all the following Tarski-biconditionals:

$$T(\phi) \leftrightarrow \phi \qquad \text{with } \phi \text{ a closed sentence of } \mathcal{L}_{PA}.$$

These are called the *restricted* Tarski-biconditionals.

TB is a *basic* theory of truth. A slight strengthening is obtained from TB if in the Tarski-biconditionals free variables are allowed to occur, which are assumed to be universally quantified over from the outside. These are called the *uniform* Tarski-biconditionals, and the theory is called UTB. The theory UTB explicates the notion of a predicate being *true of* a sequence of objects. In other words, it is a theory of the *satisfaction* relation. The concept of truth is then a limiting case of the concept 'true of', namely it can be seen as being true of a sequence of 0 objects. A slight weakening of TB, called TB^-, is obtained by restricting the induction scheme to formulas that do not contain occurrences of the truth predicate T. The theory UTB^- is defined in a similar way: it is just like UTB, except that the truth predicate is not allowed in the induction scheme.

Theories of truth that do not claim truth of any sentence that itself contain the truth predicate, are called *typed truth theories*. TB and UTB are thus typed truth theories.

As adumbrated in the introduction to this Chapter, Tarski's restriction of the instances of the Tarski-biconditionals to truth-free sentences is motivated by a desire to avoid contradictions caused by liar paradox-like reasoning. This strategy is clearly successful. TB is consistent, and indeed has a standard model in the natural numbers.

Indeed, a stronger property holds: In typed disquotational truth theories, no 'new' mathematical sentences can be derived, i.e., TB is proof-theoretically conservative over its background theory PA for purely arithmetical sentences ([**Hal14**, Theorem 7.5, p. 55]):[4]

THEOREM 5.2. *The theories $TB^{(-)}$ and $UTB^{(-)}$ are proof-theoretically conservative over PA for the background language $\mathcal{L}_{PA}$.*

One should not be tempted to conclude from this that disquotational truth theories are *always* conservative over the background theory. For instance, $UTB[S_1^2]$—i.e., the UTB axioms added to the system S_1^2—is not arithmetically conservative over its background theory S_1^2,[5] and we will later see that also some typefree disquotational theories are not proof-theoretically conservative over their background theory for arithmetical sentences.

Beside the concept of proof-theoretic conservativeness, there is also the notion *semantic* notion of conservativeness, which may be of equal philosophical importance:

DEFINITION 5.3. Theory S' in language $\mathcal{L}'$ is semantically conservative over theory S in language $\mathcal{L}$ if and only if every model of S can be expanded to a model of S'.

Proof theoretic conservativeness will play an important role in what follows, whereas philosophical issues regarding semantic conservativeness will largely left aside.

An old observation says that proof theoretic deflationism and semantic deflationism do not coincide:

THEOREM 5.4. *Proof theoretic conservativeness implies semantic conservativeness, but the converse does not hold.*[6]

For instance, UTB^- is proof-theoretically conservative over PA, but UTB^- is not semantically conservative over PA.

5.1.3. Stay positive. Tarski's culling of the Tarski-biconditionals is draconian. Banning *all* Tarski-biconditionals in which ϕ contains occurrences of T seems an overreaction to the liar paradox. Indeed, the liar sentence is a very *special* T-containing sentence. Consider, in contrast to the liar sentence:

It is true that the Twin Prime Conjecture is true
if and only if the Twin Prime Conjecture is true.

This Tarski-biconditional (and many others like it) is completely innocuous: it has no untoward consequences. Yet it is not provable in TB.

Halbach observed that Tarski-biconditionals that do lead to problems, the formula φ contains *negative* occurrences of the truth predicate, where the notion of a negative occurrence of T in ϕ is defined as follows:

DEFINITION 5.5. An occurrence of T in ϕ is *positive* if it occurs in the scope of an even number of negation signs (where 0 is counted as an even number); otherwise the occurrence is *negative*.

---

[4]The concept of proof-theoretic conservativeness was defined on p. 110.

[5]See [**NP19**, Proposition 3].

[6]For a discussion of the relation between semantic and proof theoretic conservativeness, see for instance [**Cie15**].

Halbach then proposed to restrict the Tarski-biconditionals to sentences in which the truth predicate only occurs positively.[7] The resulting theory is called PTB (for: *Positive Tarski-Biconditionals*). As before, we can define the variants PTB^{-}, and $\mathrm{PUTB}^{(-)}$ (Positive Uniform Tarski-Biconditionals). PTB and variants on it are *untyped* truth theories: they easily prove truth iterations such as "it is true that it is true that 0=0".

The theory PTB again proves no 'new' arithmetical sentences ([**Cie17**, Theorem 6.0.5, p. 92]):

THEOREM 5.6. *PTB is proof-theoretically conservative over PA for the background language* $\mathcal{L}_{PA}$.

However, this is not the case for PUTB ([**Hal09**]):[8]

THEOREM 5.7. *PUTB is proof-theoretically (highly) non-conservative over PA for the background language* $\mathcal{L}_{PA}$.

Just to give an example, PUTB proves the consistency of PA. This is surprising: one would not expect a theory of a philosophical concept such as truth to have new *mathematical* consequences! However, there is no immediate reason for concern, for all the new arithmetical consequences of PUTB are *true* arithmetical statements.

A natural way of extending Halbach's theory PTB is by expanding the language $\mathcal{L}_T$ 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 Tarski biconditionals $T(\varphi) \leftrightarrow \varphi$ with φ restricted to $\mathcal{L}^{+}_{T,F}$, which are called the *truth biconditionals*, plus the *falsity biconditionals* $F(\varphi) \leftrightarrow \overline{\varphi}$, where $\overline{\varphi}$ is the *dual* of φ. Here we define duals recursively as follows [**HL17**, section 9.2]:

DEFINITION 5.8. 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.

PA plus these two sets of biconditionals is called TFB (for: *Truth-Falsity-Biconditionals*). It is then an easy exercise to work out that TFB proves statements such as $TF(0 = 1)$. As expected, we have ([**HL17**, Theorem 12, p. 226]):

THEOREM 5.9. *TFB is proof-theoretically conservative over PA for the background language* $\mathcal{L}_{PA}$.

5.1.4. Partial. There is yet another natural way to deal with the liar paradox in a disquotational manner. The core of the response is to preserve the inferential versions of the full Tarski-biconditionals, i.e., the following two inference rules:

T-In
$$\frac{\phi}{T(\phi)}$$

T-Out
$$\frac{T(\phi)}{\phi}$$

[7] This follows immediately from [**Hal09**, Theorem 5.1, p. 792].

[8] Like all truth theories with restricted induction that I know of, PUTB^{-} is proof-theoretically conservative over its background mathematical theory.

The motivation for this is the thought that truth is no more than a device for quotation and disquotation.

The *deduction theorem* holds for classical logic:

THEOREM 5.10. $\phi \vdash \psi \Leftrightarrow \vdash \phi \to \psi$.

Thus, over classical logic, the left-to-right direction of the deduction theorem ensures that the combination of T-In and T-Out is equivalent to the full Tarski-biconditionals. In other words, merely replacing the unrestricted Tarski-biconditionals by their inferential counterparts does not block the liar paradox.

The logical rule that plays the central rule in the proof of the left-to-right direction of the deduction theorem is the rule of *Conditionalisation* (or $\to$-In):

$$\frac{\begin{array}{c}\phi\,(Hyp)\\ \vdots \\ \psi\,(Hyp)\end{array}}{\phi \to \psi}.$$

In order to block the liar paradox, the rule $\to$-In must be restricted. In the present context, this does not seem entirely unreasonable, for the following reason. A trivial application of Conditionalisation yields the conclusion that $\vdash L \to L$, or, equivalently, $\vdash \neg L \vee L$. The argument of the liar paradox shows that either of the two disjuncts in this statement yield a contradiction if the inferential version of the unrestricted Tarski-biconditionals are accepted. This is a motivation for refraining from asserting the law of excluded middle for L, and therefore to restrict Conditionalisation. In other words, we are moved to reasoning in *partial logic* rather than in classical logic.

On the other hand, we should have no qualms about applying Conditionalisation to formulas for which we know that excluded third holds. So the rule $\to$-In should be replaced by the following rule of *Restricted Conditionalisation*:

$$\frac{T(\phi) \vee T(\neg\phi) \quad \begin{array}{c}\phi\,(Hyp)\\ \vdots \\ \psi\,(Hyp)\end{array}}{\phi \to \psi}.$$

In order to give the rule of Restricted Conditionalisation teeth, we assert as an axiom scheme that the law of excluded middle holds for all formulas that do not contain occurrences of the truth predicate.

When details are further filled in, one arrives at a system of partial logic that is called *Basic De Morgan Logic* (BDM).[9] In the framework of BDM, theories of arithmetic can then be formulated. For instance, the axioms of Elementary Arithmetic can be laid down.[10] When we then also add the inference rules T-In and T-Out, we arrive at the basic disquotational theory that is called TS_0.[11]

The theory TS_0 is consistent and arithmetically sound, and of course again conservative:[12]

[9]For a precise presentation, see [**FHN21**, Section 2.1].

[10]Some care should be taken in the formulation of the induction scheme: see [**FHN21**, Section 2.2].

[11]See [**FHN21**, Section 2.2].

[12]See [**FNH17b**, Section 2.4].

THEOREM 5.11. *In partial logic, TS_0 is proof-theoretically conservative for the background arithmetical language over Elementary Arithmetic.*

In sum, according to TS_0, truth is a very simple notion. It is governed by completely unrestricted semantic ascent and descent rules. The price for this "transparency" of the concept of truth is that we can then, on pain of contradiction, no longer fully endorse the laws of classical logic.

5.2. Compositional theories

We now turn to the discussion of truth theories that take principles that state that the concept of truth commutes with logical operations to be basic truth axioms. We distinguish between typed and untyped compositional truth theories.

5.2.1. Typed compositional truth. Davidson famously argued that truth is a *compositional* concept: the truth predicate distributes over the logical connectives.[13] Yet disquotational theories do not recognise this fact uniformly.[14] Consider, as an example, the basic disquotational theory TB. It can prove, for instance, the distributivity of the truth predicate over negation pointwise; but it cannot prove this fact in full generality:

PROPOSITION 5.12.

(1) *For all $\phi \in \mathcal{L}_{PA}$: $TB \vdash T(\neg\phi) \leftrightarrow \neg T(\phi)$;*
(2) $TB \nvdash \forall \phi \in \mathcal{L}_{PA} : T(\neg\phi) \leftrightarrow \neg T(\phi)$.

A reaction to this phenomenon has been to take the general principles stating the compositionality of truth to be *basic* truth axioms. Given that the collection of true atomic arithmetical formulas is arithmetically definable by a formula val^+,[15] the standard axioms for typed compositional truth (over arithmetic) look like this [**Hor11**, p. 71]:

CT1 $\forall$ atomic $\phi \in \mathcal{L}_{PA} : T(\phi) \leftrightarrow val^+(\phi)$;
CT2 $\forall \phi \in \mathcal{L}_{PA} : T(\neg\phi) \leftrightarrow \neg T(\phi)$;
CT3 $\forall \phi, \psi \in \mathcal{L}_{PA} : T(\phi \wedge \psi) \leftrightarrow (T(\phi) \wedge T(\psi))$;
CT4 $\forall \phi(x) \in \mathcal{L}_{PA} : T(\forall x \phi(x)) \leftrightarrow \forall x T(\phi(x))$.

The typed compositional truth theory CT consists of adding these truth axioms to PA, where the truth predicate is allowed to occur in instances of the induction axiom.

The idea behind CT is thus straightforward. Since an explicit definition of the class of true atomic arithmetical sentences can be given by means of the arithmetical formula val^+, truth for complex arithmetical sentences can be reduced to truth of atomic arithmetical formulas through the compositional truth axioms CT2–CT4.

The theory CT^- is obtained by restricting the induction scheme to formulas in which the truth predicate does *not* occur. Just as TB is the most important typed disquotational theory of truth, CT is the most important typed compositional truth theory.

The theory CT is at least as strong as the theory TB ([**Hor11**, Proposition 29, p. 75–76]):

[13] See [**Dav67**].
[14] See [**Hor11**, Section 6.1].
[15] See Section 4.1.1.

PROPOSITION 5.13. *TB* $\subseteq$ *CT.*

The second part of Proposition 5.12 shows that this inclusion is proper.
In fact, CT is *much* stronger than TB. We will now see why, unlike typed disquotational truth theories, typed compositional truth theories tend to be arithmetically non-conservative over the mathematical background theory unless occurrence of truth predicates in the induction scheme is severely restricted.

A *fundamental observation* concerning CT is the following [**Hal11**, Theorem 8.39]:

THEOREM 5.14. $CT \vdash \forall \varphi \in \mathcal{L}_{PA} : \mathrm{Bew}_{PA}(\varphi) \rightarrow T(\varphi)$.

PROOF. (*Sketch.*)
This theorem is proved, in CT, by an induction on the length of proofs in PA.
For the basis case, we must show that the logical axiom schemes are true and that the arithmetical axioms of PA, including the axiom scheme of mathematical induction, are true. Let us look only at the proof that the arithmetical axiom that states that there is a least natural number is true, and at the proof that all instances of mathematical induction are true.
(a) By Proposition 5.13,

$$\neg\exists y(0 = s(y)) \leftrightarrow T(\neg\exists y(0 = s(y)))$$

is a theorem of CT. Also, since $\neg\exists y(0 = s(y))$ is an axiom of PA, the theory CT proves it also. So CT indeed proves $T(\neg\exists y(0 = s(y)))$.
(b) The following is an instance of the induction axiom of CT:

$$\forall \phi \in \mathcal{L}_{PA} : [T\phi(0) \wedge \forall y(T\phi(y) \rightarrow T\phi(y+1))] \rightarrow \forall x T\phi(x).$$

By the compositional truth axioms of CT, the truth predicate can be moved to the front:

$$T\{\forall \phi \in \mathcal{L}_{PA} : [\phi(0) \wedge \forall y(\phi(y) \rightarrow \phi(y+1))] \rightarrow \forall x \phi(x)\},$$

which is what we wanted to prove.
If we take Modus Ponens to be the only logical rule of inference, then the inductive case is easy. By the inductive hypothesis, we may assume that we have a CT-proof of $T(\phi)$ and a CT-proof of $T(\phi \rightarrow \phi)$. By the compositional truth axioms of CT, there will then also be a proof in CT of $T(\phi)$. □

We will see later how this proof plays an important role in the contemporary definition about our epistemic warrant for proof-theoretic reflection principles.[16]

By instantiating ϕ by $0 = 1$ in Theorem 5.14, we see that Theorem 5.14 entails that CT is arithmetically non-conservative over PA. The precise arithmetical proof theoretic strength of CT is that of the system ACA of predicative second-order arithmetic ([**Hal11**, Theorem 8.42, p. 108]):

THEOREM 5.15. *CT and ACA prove the same first-order arithmetical sentences.*

Combined with Theorem 4.7, this entails that CT is arithmetically non-conservative over PA.

For set theory, there is a similar connection between ZFC and a subsystem of ZFC^2 ([**Fuj12**, p. 1507]):

[16] See Section 9.4.

THEOREM 5.16. *The theories ECA and the CT[ZFC] prove the same sentences of* $\mathcal{L}_{ZFC}$*,*

where CT[ZFC] is the theory that is obtained by adding the typed compositional truth principles to ZFC. There is actually a wrinkle in the definition of theories such as CT[ZFC], to which we now turn.

One special feature of arithmetic is that every object that arithmetic is about—i.e., every natural number—has a standard name (a standard arabic numeral, for instance). This is what enables theories such as CT to *define* truth for atomic arithmetical sentences. But there are also theories that are about objects that do not all have standard names. Set theory is one such theory. When a truth theory has one of the latter theories as background theory, it must take a slightly more complicated form.

One possibility is to work not with a primitive truth predicate but with a primitive satisfaction predicate $Sat(x, y)$, which intuitively says "formula x is *true of* set y". Then analogues for the truth axioms are formulated for the satisfaction predicate, and a truth predicate is *defined* in terms of satisfaction:

$$T(x) \equiv \text{Sentence}(x) \wedge \forall y Sat(x, y),$$

which says that a sentence is true if and only if it is—vacuously, because it is a closed formula—"satisfied by" (or true of) each set.

A second possibility is to work with an internally coded infinitary language of set theory, roughly as follows. Let $\mathcal{L}_\epsilon$ be the first-order language of set theory, with ϵ and $=$ as its only non-logical symbols, and let $\mathcal{L}_\epsilon^T$ be the extension of this language with a primitive truth predicate. Now let $\mathcal{L}_V^T$ to be the extension of $\mathcal{L}_\epsilon^T$ with a constant $\dot{a}$ for each set a in V. Then it is not hard to see that ZFC can develop an internal syntax theory for this highly uncountable language $\mathcal{L}_V^T$. Using this internal syntax theory, the compositional truth axioms can be expressed. For instance, we can then express that a universal statement φ in the finitary language $\mathcal{L}_\epsilon^T$ if and only if all instantiations of it are true in the highly uncountable language $\mathcal{L}_V^T$.[17] For definiteness, we assume that, if the problem arises (such as in the case of CT(ZFC)), we deal with domains that contain un-named objects in this second way.

The situation is different when we restrict the induction scheme of CT ([**KKL81**]):

THEOREM 5.17. *The theory* CT^- *is arithmetically conservative over its background arithmetical theory PA.*

Nonetheless, adding a minimum amount of induction for the extended language $\mathcal{L}_T$ makes CT^- non-conservative. More precisely, if we define CT_0 as the theory that results from adding mathematical induction for Δ_0 formulas *of the extended language* to CT^-, then we have [**WL17**, Theorem 3.1]:

THEOREM 5.18. CT_0 *is arithmetically non-conservative over its background arithmetical theory PA.*

In Section 4.2.2, we saw that predicative comprehension can be *iterated*, yielding ever stronger systems RA_α of predicative analysis. In a similar way, we can construct ever stronger typed compositional truth theories in stages. The first stage after the construction of CT looks as follows. We add a new truth predicate T_1

[17] For details, see [**Fuj12**, Section 2].

to $\mathcal{L}_T$, yielding a more expansive language $\mathcal{L}_{T,T_1}$. Then we formulate PA in this more expensive language, allowing also the new truth predicate T_1 to occur in the induction axiom and in the logical axioms. The truth axioms of CT are added to govern the 'old' truth predicate T. But also new typed compositional truth axioms can be added that govern the new truth predicate T_1:

CT1_1 $\forall$ atomic $\phi \in \mathcal{L}_{PA} : T_1(\phi) \leftrightarrow val^+(\phi)$;
CT2_1 $\forall \phi \in \mathcal{L}_T : T_1(\neg\phi) \leftrightarrow \neg T_1(\phi)$;
CT3_1 $\forall \phi, \psi \in \mathcal{L}_T : T_1(\phi \wedge \psi) \leftrightarrow (T_1(\phi) \wedge T_1(\psi))$;
CT4_1 $\forall \phi(x) \in \mathcal{L}_T : T_1(\forall x \phi(x)) \leftrightarrow \forall x T_1(\phi(x))$.

All these principles together constitute the truth system CT_1, of which the theory CT is of course a sub-theory. It can then again be verified that CT_1 is a well-behaved theory, and also that it is arithmetically non-conservative over CT. In a similar way, a next stronger compositional theory CT_2 can then be constructed, and so on, where at limit stages we simply take the union of all truth theories that have already been constructed. At some stage, we will run into problems with coding transfinite ordinals as natural numbers, but we will not worry about this here, and consider this iteration process only up to 'small' countable ordinal stages.

Then the following generalisation of Theorem 5.15 can be proved:

THEOREM 5.19. *Up to relatively small transfinite countable ordinals α, the theories RA_α and CT_α prove the same first-order arithmetical sentences.*

PROOF. As in the case of Theorem 5.15, the proof actually shows that the two theories can be relatively interpreted in each other in a way that leaves arithmetical sentences unchanged. □

Since such theories RA_α are predicatively acceptable, Theorem 5.19 shows that the hierarchy of typed compositional truth theories CT_α give an alternative description of at least the arithmetical content of predicative analysis. There is a school of thought that argues that when two theories can be relatively interpreted in each other, they have the same mathematical content. If that is so, then the hierarchy of typed truth theories can be seen as an alternative way of spelling out the content of predicative analysis.

We will see later that there is a tight connection between proof theoretic *reflection principles* and transfinite induction.[18] Similarly, there ia also a tight connection between *truth principles* and transfinite induction [**Lei16**, Theorem 1.6, Lemma 3.11]:[19]

THEOREM 5.20.
(1) $CT \vdash TI(\phi, < \varepsilon_{\varepsilon_0})$ *for all* $\phi \in \mathcal{L}_{PA}$;
(2) $CT \vdash TI(\phi, < \varepsilon_0)$ *for all* $\phi \in \mathcal{L}_T$.

The above are not only lower bounds, but also upper bounds.

5.2.2. Untyped compositional truth. We have seen how in the case of disquotational truth theories, there are two kinds of untyped truth theories: theories framed in classical logic, and theories based on non-classical logic. The same holds for compositional truth theories. We now discuss each kind in turn.

[18] See Section 6.2.1.

[19] The principle of transfinite induction up to α for the formula ϕ ($TI(\phi, < \alpha)$) was defined on p. 114.

5.2.2.1. *Classical.*

The most influential untyped theory of typefree truth is the theory KF ('Kripke-Feferman'), which was formulated and investigated by Feferman in his article [**Fef91**]. It can be seen as an attempt to formalise externally, in a classical context, a certain type of *partial* models for $\mathcal{L}_T$. But the theory KF can also be motivated, simply as a natural consistent classical axiomatic typefree truth theory, without specific reference to these models.

In our discussion of probability,[20] we saw that at least the principles of typed *finitely additive* probability can safely be 'untyped', but that most probability 1-introspection principles cannot then consistently be added as extra axioms. General truth iteration principles, in contrast, seem very plausible. From a proof theoretic point of view, Feferman's theory KF can be seen as an attempt to validate these principles while at the same time coming as close to fully 'untyping' CT as possible.

The theory KF consists of adding the following compositional truth axioms to the background arithmetical theory PA:

KF1 $\forall$ atomic $\phi \in \mathcal{L}_{PA} : T(\phi) \leftrightarrow val^{+}(\phi)$;
KF2 $\forall$ atomic $\phi \in \mathcal{L}_{PA} : T(\neg\phi) \leftrightarrow val^{-}(\phi)$;
KF3 $\forall\phi \in \mathcal{L}_T : T(\neg\neg\phi) \leftrightarrow T(\phi)$;
KF4 $\forall\phi,\psi \in \mathcal{L}_T : T(\phi \wedge \psi) \leftrightarrow (T(\phi) \wedge T(\psi))$;
KF5 $\forall\phi,\psi \in \mathcal{L}_T : T(\neg(\phi \wedge \psi)) \leftrightarrow (T(\neg\phi) \vee T(\neg\psi))$;
KF6 $\forall\phi(x) \in \mathcal{L}_T : T(\forall x\phi(x)) \leftrightarrow \forall y T(\phi(y))$;
KF7 $\forall\phi(x) \in \mathcal{L}_T : T(\neg\forall x\phi(x)) \leftrightarrow \exists y T(\neg\phi(y))$;
KF8 $\forall\phi \in \mathcal{L}_T : T(T(\phi)) \leftrightarrow T(\phi)$;
KF9 $\forall\phi \in \mathcal{L}_T : T(\neg T(\phi)) \leftrightarrow T(\neg\phi)$;
KF10 $\forall\phi \in \mathcal{L}_T : \neg(T\phi \wedge T\neg\phi)$.

Thus KF is a strongly compositional type-free theory of truth (KF1–KF7) that includes natural truth iteration axioms (KF8–KF9).

Axiom KF10 expresses the consistency of the extension of the truth predicate. It is a bit of an odd duck. Unlike the other truth axioms, it does not reduce the truth of statements to the truth of other statements or to elementary facts: it is not "inductive". For this reason, Axiom KF10 is sometimes left out. Axiom KF10 is for instance left out of the official version of KF in [**Fef91**]. Indeed, a number of *variants* of KF are discussed in the literature.[21] For many purposes, the differences between variants do not matter much—leaving out Axiom KF10 does not diminish the mathematical strength of KF, for instance.[22] But it is a natural axiom, and if we include it, then KF is able to prove that truth is closed under Modus Ponens:

PROPOSITION 5.21. $KF \vdash \forall\phi,\psi \in \mathcal{L}_T : [T(\phi) \wedge T(\phi \to \psi)] \to T(\psi)$

The theory KF proves one half of the Tarski-biconditionals:

PROPOSITION 5.22. *For all* $\phi \in \mathcal{L}_T$*:* $KF \vdash T(\phi) \to \phi$.

PROOF. This is proved by a straightforward induction on the complexity of ϕ. □

From Proposition 5.22, together with our earlier Lemma 5.1, an asymmetry with typefree Kolmogorov probability immediately follows:

[20] See Section 4.3.

[21] In Section 6.3.1.2 we will encounter one such variant of KF.

[22] This is shown in [**Can89**].

COROLLARY 5.23. *The theory that results from adding the Necessitation rule for the truth predicate to KF is inconsistent.*

The reader will have noticed that *even disregarding* Axiom KF10, the theory KF is not *fully* compositional. In particular, KF does not contain an axiom that states that the truth predicate commutes with negation. It is in fact very easy to verify that adding the untyped version of the negation Axiom CT2 to the theory KF yields an inconsistent system. But the omission from this Axiom from KF can be motivated. Concerning paradoxical sentences such as the Liar, it seems as wrong to assert that they are false—i.e., that their negation is true—as it would be to assert that they are true. But if the untyped version of CT2 is added as an axiom, then the resulting system proves that for *every* statement either it or its negation is true.

The axiomatic theory KF was developed by Feferman[23] as a natural formalisation of Kripke's[24] *semantic* theory of truth, to which we now turn. The reasoning of the liar paradox appears to show that assuming the liar sentence to be true leads to a contradiction and that assuming it to be false also leads to a contradiction. This led Kripke to propose that the liar sentence has no truth value at all: it is neither true nor false. In the literature, this is often expressed by calling the liar sentence "gappy".

This means that in order to reason with paradoxical sentences such as the liar paradox, we have to reason not in classical logic, where every sentence is taken to have exactly one of two truth values ('true' and 'false'), but in partial logic, where sentences are also allowed to be gappy. A system of partial logic tends to propose an evaluation scheme similar to the familiar truth clauses from classical logic for the logical connectives. One of the most natural and popular evaluation schemes for partial logic, and the scheme that we will work with here, is the *strong Kleene* evaluation scheme. The truth clauses for strong Kleene logic are given by the following tables (where $\star$ denotes the absence of a truth value, i.e., gappiness):

p	¬p
1	0
*	*
0	1

p	q	p ∨ q
1	1	1
1	*	1
1	0	1
*	1	1
*	*	*
*	0	*
0	1	1
0	*	*
0	0	0

p	q	p ∧ q
1	1	1
1	*	*
1	0	0
*	1	*
*	*	*
*	0	0
0	1	0
0	*	0
0	0	0

FIGURE 1

[23]See [**Fef91**].
[24]See [**Kri75**].

The material implication is defined in terms of $\neg$ and $\vee$ exactly as in classical logic. The clauses for existential and universal generalisations are just the infinitary analogues of the clauses for conjunction and disjunction. Given the ordering

$$0 < \star < 1,$$

we take the truth value of a formula of the form $\forall x\varphi(x)$ to be the *infimum* of its instances, and we take the truth value of a formula of the form $\exists x\varphi(x)$ to be the *supremum* of its instances.

Concerning atomic sentences of the form Ft, a strong Kleene model assigns an *extension* and an *anti-extension* to F, which are required to be *disjoint*, and takes Ft to be true if the denotation of t is in the extension of F, false if the denotation of t is in the anti-extension of F, and gappy if the denotation of t is neither in the extension, nor in the anti-extension of F.

So the case of the atomic sentences together with the inductive cases together define a semantic evaluation relation that holds between partial models $\mathfrak{M}$ and sentences φ of a given language. We denote this relation as $\mathfrak{M} \models_{sk} \varphi$. We say that a partial model makes a whole theory S true, i.e., $\mathfrak{M} \models_{sk} S$, iff $\mathfrak{M} \models_{sk} \varphi$ for every $\varphi \in S$.

For our purposes, the truth predicate T will be the only non-classical predicate, i.e., it will the only predicate such that the union of its extension and its anti-extension does not exhaust the domain of discourse. Beside the truth predicate, our models will of course have to interpret the arithmetical vocabulary. But they will always interpret the arithmetical vocabulary in the standard way in the standard natural numbers. So all our partial models $\mathfrak{M}$ will be of the form

$$\langle \mathbb{N}, \langle \mathfrak{M}^+, \mathfrak{M}^- \rangle \rangle,$$

where $\mathfrak{M}^+$ is the extension of T according to $\mathfrak{M}$, and $\mathfrak{M}^-$ is the anti-extension of T according to $\mathfrak{M}$.

Kripke intends to construct a 'good' or intended model for typefree truth in *stages*, which are indexed by ordinals. In each subsequent stage of the process that Kripke describes, an improved model is produced. The intended model for $\mathcal{L}_T$ is then generated as a 'limit' of these approximations.

The way of producing ever better models for typefree truth is supposed to mirror the way in which the concept of truth is learned by young children. The little daughter proudly announces that she has learned that 1+2=3. Then her mother introduces the concept of truth by adding that therefore *it is true that* 1+2=3. And she continues that *that* in turn implies that it is true that it is true that 1+2=3. *And so on.* Somewhat later, the child learns that the natural numbers go on indefinitely, and therefore the "and so on" means that when for any finite number n, n copies of "it is true that" are prefixed to "1+2=3", the result is still a true statement. And then she learns that even *that* sentence is true. And so on.

This process is abstractly expressed by Kripke in his inductive definition of a particular model of typefree truth. He starts with a model in which the extension and anti-extension of the truth predicate are empty. At every successor stage, the sentences that are made true (in the partial, strong Kleene sense) at the previous stage are added to the extension of the truth predicate, and the sentences that are made false (in the strong Kleene sense) are added to the anti-extension of the truth predicate. At limit stages, unions are taken. Formally, the extensions and anti-extensions of the approximations—and therefore the approximations themselves,

since they are determined by their extensions and anti-extensions—are simultaneously defined inductively as follows:

DEFINITION 5.24.

$$\mathfrak{M}_0^+ \equiv \mathfrak{M}_0^- \equiv \emptyset$$

$$\mathfrak{M}_{\alpha+1}^+ \equiv \{\varphi \mid \mathfrak{M}_\alpha \models_{sk} \varphi\}$$

$$\mathfrak{M}_{\alpha+1}^- \equiv \{\varphi \mid \mathfrak{M}_\alpha \models_{sk} \neg\varphi\}$$

$$\mathfrak{M}_\lambda^+ \equiv \bigcup_{\beta<\lambda} \mathfrak{M}_\beta^+ \qquad \text{whenever } \lambda \text{ is a limit ordinal}$$

$$\mathfrak{M}_\lambda^- \equiv \bigcup_{\beta<\lambda} \mathfrak{M}_\beta^- \qquad \text{whenever } \lambda \text{ is a limit ordinal}$$

It is clear that the sentence $0 = 0$ first enters the extension of the truth predicate at stage 1, the sentence $T(0 = 0)$ first enters the extension of the truth predicate at stage 2, and so on. In this way, the inductive process captures the way in which the truth concept is learned by the young child.

For elementary cardinality reasons, this inductive procedure of generating more and more truths and falsehoods closes off at some stage:

THEOREM 5.25.

There is an ordinal κ such that $\mathfrak{M}_\kappa^+ = \mathfrak{M}_{\kappa+1}^+$ and $\mathfrak{M}_\kappa^- = \mathfrak{M}_{\kappa+1}^-$.

In other words, $\mathfrak{M}_\kappa = \mathfrak{M}_{\kappa+1}$: from stage κ onwards, no *improved* models are generated. The ordinal κ is called the *closure ordinal* of Kripke's inductive definition.

The model $\mathfrak{M}_\kappa = \langle \mathbb{N}, \langle \mathfrak{M}_\kappa^+, \mathfrak{M}_\kappa^- \rangle \rangle$ is the intended model (or 'limit model') of $\mathcal{L}_T$ that we have been looking for. $\mathfrak{M}_\kappa$ is called the *minimal fixed point model* of Kripke's inductive definition of models of $\mathcal{L}_T$.

It turns out that according to the minimal fixed point model $\mathfrak{M}_\kappa$, the liar sentence is indeed gappy—it belongs neither to the extension nor to the anti-extension of T,—which is a pleasing result. Similarly, the negation of the liar sentence turns out to be gappy. These two facts together imply that the logical law of excluded middle does not fully hold in $\mathfrak{M}_\kappa$, whereby $\mathfrak{M}_\kappa$ is not a model of full classical logic.

The judgements by $\mathfrak{M}_\kappa$ concerning the truth status of sentences of $\mathcal{L}_T$ appear to be impeccable. In particular, no sentences have ever been found that are made true by $\mathfrak{M}_\kappa$ in the strong Kleene sense of the word (or equivalently, that belong to $\mathfrak{M}_\kappa^+$) that do not also intuitively appear to be true.

Then Kripke goes on to define the *'closed off version'* of $\mathfrak{M}_\kappa$, which is the *classical* model $\mathfrak{M}_\kappa^\star = \langle \mathbb{N}, \mathfrak{M}_\kappa^+ \rangle$. In other words, according to the classical model $\mathfrak{M}_\kappa^\star$, the anti-extension of the truth predicate is the union of the anti-extension of the truth predicate according to $\mathfrak{M}_\kappa$ and the collection of the sentences that are gappy according to $\mathfrak{M}_\kappa$. This of course implies that $\mathfrak{M}_\kappa^\star$ is a model of classical logic.

Like $\mathfrak{M}_\kappa$, the model $\mathfrak{M}_\kappa^\star$ turns out also to be a very special model for the language $\mathcal{L}_T$. It is a natural model for our highly compositional, typefree, and classical theory of truth:

THEOREM 5.26. $\mathfrak{M}_\kappa^\star \models KF$.

So KF is a consistent theory, and it has standard models. Moreover, the typed compositional theory CT is a sub-theory of KF.

In fact, KF is a much stronger theory than CT, as we will now see.

DEFINITION 5.27. Theory S_1 is *proof-theoretically equivalent* to theory S_2 for a class of formulas Φ if and only if there are primitive recursive functions f_1, f_2 such that f_1 transforms any S_1-proof of a formula in Φ into an S_2-proof of that same formula, and f_2 transforms any S_2-proof of a formulas in Φ into an S_1-proof of that same formula.

Then Feferman has shown that ([**Fef91**, Theorem 4.1.1, p. 23]):

THEOREM 5.28. *The theories KF and RA_{ε_0} are proof-theoretically equivalent for sentences in $\mathcal{L}_{PA}$.*

Some take this to mean that KF is just another way of presenting a hierarchy of predicatively acceptable systems of analysis. In fact, this appears to be how Feferman himself saw the import of theorem 5.28.[25] Moreover, given theorem 5.15, this can also be taken to show that KF can be seen as a succinct way of expressing the whole hierarchy of typed compositional truth theories up to level ε_0.

Theorem 5.28 indicates that the connection between truth principles and transfinite induction also persists for typefree truth theories ([**Lei16**, Theorem 1.6]):

THEOREM 5.29.

(1) $KF \vdash TI(\phi, < \varphi_{\varepsilon_0}(0))$ *for all* $\phi \in \mathcal{L}_{PA}$*;*
(2) $KF \vdash TI(\phi, < \varepsilon_0)$ *for all* $\phi \in \mathcal{L}_T$*.*

So even though CT and KF prove the same amount of transfinite induction for the whole language, the typefree axioms of KF boost this much more for arithmetical induction for the underlying purely arithmetical language than the typed truth axioms for CT do. Moreover, given Theorem 5.28, we see that KF proves a substantial fragment of predicative analysis.

Feferman also defined the *schematic version* KF(P) of the Kripke-Feferman theory. In KF(P), a schematic predicate P is added to the background language $\mathcal{L}_{PA}$, yielding the expanded background language $\mathcal{L}_P$. Moreover, the theory of typefree truth KF(P) that is built over this background language, and is formulated in the language $\mathcal{L}_{P,T}$, contains the axioms of KF and is closed under the *schematic substitution rule* Sub, which is defined as [**Fef91**, p. 21]:

DEFINITION 5.30.

$$\frac{\vdash \Phi(P)}{\vdash \Phi(B)},$$

with $\Phi(P)$ any formula in $\mathcal{L}_P$, and B any formula in $\mathcal{L}_{P,T}$.

The system KF(P) is mathematically substantially stronger than KP ([**Fef91**, Theorem 5.1.1, p. 30]):

THEOREM 5.31. *The theories KF(P) and RA_{Γ_0} are proof-theoretically equivalent for sentences in $\mathcal{L}_{PA}$.*

Let us go back to the theory KF. In Proposition 5.26 we are using the modelling relation in the *classical* sense of the word, and not in the strong Kleene sense of the

[25] See [**Fef91**, p. 3].

word. Observe that this is a bit odd. Why does Kripke switch from partial models to a classical model at the very end of the approximation process?

Most probably Kripke's motivation for 'closing off' the minimal fixed point model $\mathfrak{M}_\kappa$ was a desire to uphold classical logic. Moreover, we have seen that $\mathfrak{M}_\kappa^\star$ satisfies a very natural compositional theory of typefree truth. Indeed, KF can be seen as a natural axiomatisation of $\mathfrak{M}_\kappa^\star$. Nonetheless, the switch to the closed off model does cause problems, to which we now turn.

A serious conceptual problem with KF is that it is in a sense self-undermining. There are sentences that it claims (i.e., proves) but at the same time claims (i.e.) to be untrue:

LEMMA 5.32. *$KF \vdash \lambda \wedge \neg T(\lambda)$, where λ is the liar sentence.*

This is of course not something that we expect of a theory of truth. It seems to violate a law of assertion: *only assert what you know to be true*. This problem with KF, and therefore also with $\mathfrak{M}_\kappa^\star$, was first isolated in [**Rei86**]. Moreover, in this article Reinhardt also sketches and defends a possible solution.[26]

Reinhardt urges us to focus not on KF, but on what he calls the *inner logic* of KF (IKF), and which is defined as follows:

DEFINITION 5.33. $IKF \equiv \{\varphi \mid KF \vdash T\varphi\}$

Clearly IKF is a recursively enumerable and therefore axiomatisable collection of sentences. In fact, IKF proves exactly the same arithmetical sentences that KF does:

PROPOSITION 5.34. *For all $\varphi \in \mathcal{L}_{PA}$: $IKF \vdash \varphi \Leftrightarrow KF \vdash \varphi$.*

And unlike the theorems of KF, the theorems of IKF are all unobjectionable:

PROPOSITION 5.35. *$IKF \subseteq \mathfrak{M}_\kappa^+$.*

Of course this implies that IKF is not closed under full classical logic.

One challenge that Reinhardt formulated in his article, is to give a *natural axiomatisation* of IKF.[27] Indeed, one may wonder at this point if IKF can be seen as a *natural* truth theory at all. Reinhardt argues that in the absence of a natural reasoning system for IKF, we can still rely on KF to a limited degree. It follows from Proposition 5.32 that we cannot fully trust KF, whereas it follows from Proposition 5.35 that IKF is unobjectionable. For this reason, Reinhardt proposed that we use KF only as an engine for generating theorems of the form $T\varphi$, i.e., elements of IKF, which we then fully and unconditionally accept. In other words, Reinhardt advocates an *instrumentalist* stance towards KF.

5.2.2.2. *Partial.*

We have seen that KF is not *fully* compositional. If we wants to make KF more fully compositional by stipulating that truth commutes with negation, then, on pain of contradiction, the truth predicate cannot be governed by classical logic. But if we retreat to partial logic as described by BDM,[28] then we can arrive at a fully compositional theory of typefree truth, which is called PKF.[29]

---

[26] See [**Rei86**].

[27] We return to Reinhardt's challenge in Section 5.2.2.2.

[28] See Section 5.1.4.

[29] This truth theory is described and investigated in [**HH06**].

The truth principles of PKF are based on the *rule-counterparts* of the truth axioms of KF. To be precise, aside from a rule-version of the axioms of Peano Arithmetic, PKF contains the rules:

PKF1

$$\frac{val^+(t_1 = t_2)}{T(t_1 = t_2)} \quad \frac{T(t_1 = t_2)}{val^+(t_1 = t_2)}$$

PKF2

$$\frac{T(\phi) \wedge T(\psi)}{T(\phi \wedge \psi)} \quad \frac{T(\phi \wedge \psi)}{T(\phi) \wedge T(\psi)}$$

PKF3

$$\frac{T(\phi) \vee T(\psi)}{T(\phi \vee \psi)} \quad \frac{T(\phi \vee \psi)}{T(\phi) \vee T(\psi)}$$

PKF4

$$\frac{\forall x T(\phi(x))}{T(\forall x \phi(x))} \quad \frac{T(\forall x \phi(x))}{\forall x T(\phi(x))}$$

PKF5

$$\frac{\exists x T(\phi(x))}{T(\exists x \phi(x))} \quad \frac{T(\exists x \phi(x))}{\exists x T(\phi(x))}$$

PKF6

$$\frac{T(\phi)}{T(T(\phi))} \quad \frac{T(T(\phi))}{T(\phi)}$$

PKF7

$$\frac{\neg T(\phi)}{T(\neg \phi)} \quad \frac{T(\neg \phi)}{\neg T(\phi)}$$

Observe that Axiom PKF7 ensures that truth fully commutes with negation.

It is important to note that in the inference rules of PKF, t_1, t_2, ϕ and ψ function as *variables* ranging over terms and formulas, respectively. So as in the truth axioms in CT and unlike the truth axioms of TB, the inference rules universally quantify over terms and formulas. It is just that in PKF this is done implicitly, using free variables which are always treated as universally quantified over, while in CT this is done explicitly.

PKF contains a natural tyefree disquotational theory of truth:

PROPOSITION 5.36. $TS_0 \subset PKF$.[30]

Unlike the disquotational theory TS_0, however, the system PKF is not conservative over its background theory [**HH06**, Theorem 39]:

THEOREM 5.37. *The theories PKF and RA_{ω^ω} prove the same arithmetical sentences.*

This means that PKF is mathematically stronger than the typed compositional theory CT, but weaker than KF.

It follows from Proposition 5.34 and Proposition 5.37 that the theories PKF and IKF do not coincide. Nonetheless, the theories PKF and IKF are closely related. Nicolai has shown that [**Nic18**, Theorem 1]:

THEOREM 5.38. *IKF = PKF + transfinite induction up to ε_0 for $\mathcal{L}_T$.*

[30]The system TS_0 was described in Section 5.1.4.

In other words, the only thing preventing IKF from coinciding with PKF is that it does not contain enough transfinite induction; as far as *truth laws* go, the two coincide. He thus meets Reinhardt's challenge that was discussed at the end of Section 5.2.2.1.

5.3. Deflationism

Let us now turn to *philosophical* views about truth and their relations to axiomatic theories of truth. We concentrate on what is called truth theoretic deflationism. The core of this view is the rather nebulous thesis that truth is not a substantial notion and does not isolate a substantial property. In particular, being true is not a substantial *philosophical* property, and truth is not a deep philosophical concept.

Philosophical views about truth intend to address questions about the nature of the property of truth, and on the content and function of the concept of truth. These views can be divided in *deflationist* and *substantivist* philosophical theories of truth. The issues involved are subtle, and I will not be able to do justice to the complexities of the relevant philosophical considerations involved.

5.3.1. The correspondence theory. Traditionally, truth has been thought of as one of the most weighty, central, and complicated philosophical concepts. The *correspondence theory of truth* is one of the views that subscribes to this view. For more than a millennium, it has been the most popular substantivist theory of truth. Because it takes truth to be a weighty, central, and complicated philosophical concept, it is called a *substantivist* theory of truth.

According to this view, truth is a relation between propositions or maybe thoughts on the one hand, and facts or perhaps states of affairs on the other hand. The core idea is that truth is *correspondence* between a proposition or thought on the one hand, and a fact or state of affairs on the other hand: *adaequatio rei et intellectus*, in Thomas Aquinas' terms.

One of the reasons why, according to the correspondence theory, truth is deep and complicated, is that the concepts proposition, thought, fact, and state of affairs are unquestionably deep and complicated. Indeed, nothing like a consensus was ever reached about what a satisfactory philosophical theory of these concepts would more or less look like. Another reason why, according to the correspondence theory, truth is a complicated matter, is that no consensus could ever be reached about the approximate nature of the correspondence relation involved. The upshot is that there are wildly different views about the content of the correspondence theory of truth.

In the semantic work of Tarski, we witness an attempt at arriving at a more tractable and exact theory. Instead of taking propositions of thoughts as truth bearers, Tarski takes *sentences* (of a given language) to be the bearers of the concept of truth. Moreover, he dispenses with facts and states of affairs altogether. Instead, he makes use, in his definition of truth, of a collection $\mathcal{D}$ of objects (a domain of discourse),[31] plus a collection of sub-collections of $\mathcal{D}$ and a collection of relations on $\mathcal{D}$. All these notions are relatively well-understood: grammars are good theories

[31]Tarski took this domain of discourse to be *fixed* and consisting of all the objects that exist. Later, this assumption was relaxed so that any collection of objects is admissible as a possible domain of discourse.

of the collection of sentences of languages, and set theory is a good theory of collections and relations. In terms of these tractable concepts, Tarski defines the notion of truth with mathematical exactness. Whether the resulting theory can be seen as a form of the correspondence theory of truth, is open to debate. Tarski himself thought so ([**Tar44**]); some other philosophers demur.[32]

The lack of precision is not the only problem with the correspondence theory of truth. It is the ambition of the correspondence theory to give a *real definition* of truth, i.e., an explicit definition of truth where the definiens expresses the *essence* of truth. But it is not clear that this ambition can be realised. It might be that truth, like aesthetic beauty, and moral goodness perhaps, is one of the most *basic* philosophical concepts, that admits of no reductive definition in terms of more primitive notions. Tarski's own work in semantics seems to point in this direction, since his explicit definition of truth, for a given language, is formulated in an essentially richer metalanguage. If an explicit definition of truth, let alone a 'real' definition of truth, is unattainable, then it seems advisable to concentrate on uncovering the *laws of truth*. This is of course exactly what is done in the field of axiomatic theories of truth.

5.3.2. Minimalism. In this subsection, we discuss Horwich's version of deflationism about truth, as he develops it in [**Hor90**].

Horwich believes truth to be an undefinable, primitive concept. So, for him, a truth theory *in the strictest sense* consists of a collection of basic principles. More in particular, he holds that the correct theory of truth consists of a collection of unproblematic instances of the Tarskian equivalence schema:

$$\langle p \rangle \text{ is true} \Leftrightarrow p,$$

for p ranging over all *propositions*, and where $\langle p \rangle$ stands for a standard name for p. He calls this the *minimal theory of truth* (MT).

There are two reasons why Horwich's theory MT is only *semi-formal*. First, the Tarski-biconditionals of MT are not formulated "on top of" a formal theory of propositions. This is unlike the case of modern axiomatic theories of truth, where the background arithmetical (or, equivalently, syntacticall) theory is made fully explicit. In the latter case, the naming machinery can be explicitly defined, using a coding scheme. In the former case, the unclarity surrounding the notion of proposition also infects the naming machinery. (Does every proposition have a standard name?) Secondly, it is not clear exactly which Tarski-biconditionals are in, and which are out. The semantic paradoxes show that not all of them can be in, and we have seen in earlier Sections in this Chapter that it is a highly nontrivial task to decide what the best disquotational truth theory is.

Based on his semi-formal minimal theory of truth MT, Horwich then develops a philosophical view about the nature and function of the concept of truth. This can be called the *minimalist conception* of truth. We will see how Horwich's view about the nature of truth flows from his view about the function of the concept of truth.

Horwich believes that the *content* of the concept of truth is given by a simple and natural collection of Tarski-biconditionals. This doctrine, which has been held

[32]For a discussion of this question, see for instance [**Sch98**].

by a number of authors at a number of times,[33] is called *disquotationalism*. So Horwich's minimalist conception of truth is a specific form of disquotationalism.

The question of the nature of truth is deeply related to the question what role the notion of truth plays in our intellectual endeavours. This is a question about the range of applications of the concept of truth. Horwich's view about the function of the concept of truth has a positive aspect and a negative aspect. The positive aspect describes how and where the concept of truth plays a helpful role, and the negative aspect tells us where the concept of truth cannot play a helpful role (and why that is so).

Everyone agrees that the concept of truth is a useful conceptual device for expressing generalisations. The concept of truth allows us to assert an infinity of sentences in one finite statement. Without the concept of truth, we could not assert assert all the consequences of Newton's theory of motion, for there are infinitely many of them. The concept of truth, however, allows us to do it:

All consequences of Newton's theory of motion are true.

Thus truth allows us to express certain kinds of infinite conjunctions (and disjunctions). This role of being a tool for expressing complex propositions is one that it shares with the familiar *logical* words such as 'not', 'or', and 'some'.

Expressing truth generalisations is one thing; reasoning with them is another. Quine ([**Qui86**]), and Horwich in his footsteps ([**Hor90**]), maintain that all that is needed, is licencing quotational and disquotational inferential moves, i.e., inferring φ from $T(\varphi)$ and *vice versa*. Thus, they argue, a suitable set of Tarski-biconditionals suffices to govern our reasoning with truth generalisations. We have seen in Section 5.1.1 that it is a highly non-trivial task to specify *exactly* which collection of Tarski-biconditionals is optimal for performing this task.

But logic is neutral in substantive disputes, whereas truth has traditionally been thought to play a vital role in philosophical debates. For instance, one important philosophical question is the question of scientific realism, i.e., whether what is entailed by our best scientific theories is likely to be true. There seems at first blush no hope of answering this question without enquiring into the nature and laws of truth. Similar points can be made for other philosophical debates. To give another example, how are we to evaluate the 'traditional' theory of knowledge, according to which knowledge is true justified belief, without asking deep and difficult questions about truth?

Horwich rejects this view, and reasons more or less as follows. Controversial philosophical theses are indeed for convenience's sake formulated using the concept of truth. Take, for example:

Our best scientific theories are approximately true.

If we formalise the predicate 'best scientific theory' as $B(x)$, and for simplicity's sake forget about the qualification 'approximately', then this statement has the following form:

$$\forall x : B(x) \to T(x).$$

However, Horwich would insist that truth theory does not play a substantive role in establishing or refuting this thesis. Roughly, according to Horwich, we should concentrate on the corresponding *truth-free* scheme

$$B(\varphi) \to \varphi \qquad \text{for all } \varphi \text{ belonging to the 'language of science'.}$$

[33] See for instance [**Hal01a**], [**HL17**].

Since the truth predicate does not play a role in this scheme, we can establish or refute it using background theories of philosophy of science and / or metaphysics, without drawing on the laws of truth at all. So truth theory is not needed after all. Other examples of controversial philosophical theses in which the truth predicate occurs can be analysed in a similar way, according to Horwich.

This is the negative part of Horwich's view of the function of the concept of truth. A corollary of this line of reasoning is that truth is not a philosophically substantial notion. Therefore the minimalist conception of truth can be classified as a species of truth theoretic deflationism.

One question that is not answered by Horwich's 'negative' argument, however, is whether the truth generalisation $\forall x : B(x) \rightarrow T(x)$ can in such a situation be established. Halbach showed that even though a truth theory consisting of natural Tarski-biconditionals can prove (over a background theory) the instances of this truth generalisation (by simple applications of the 'quotational direction' of the truth axioms), it cannot prove the truth generalisation itself ([**Hal99**]).[34] Thereby Horwich's contention that a suitable set of Tarski-biconditionals suffices to govern our reasoning with truth generalisations is open for debate.

This issue is called the *truth generalisation problem* ([**Hal14**, p. 57]). We will come back to it later.[35] But for now, let it suffice to say that Horwich would probably maintain that *for the philosophical debate* about realism, all that matters is the truth-free schematic assertion.

5.3.3. Conservativeness deflationism. One can generalise from Horwich's line of reasoning to the following general philosophical conservativeness claim:

Our best truth theory is proof theoretically conservative
for the background language over any reasonable background theory.

This is a form of truth theoretic deflationism, for it maintains that truth does not play an essential role in philosophical debates: the *only* function of the truth predicate is to express truth generalisations.

Let us call this view *conservativeness deflationism*.[36] It has been explicitly or implicitly endorsed by several authors. Horsten and Leigh defend a type-free disquotational truth theory as our fundamental theory of truth, and thus indirectly endorse conservativeness deflationism.[37] Waxman explicitly endorses conservativeness deflationism in the following passage:

> Is there any reasonable scope for denying that a deflationist theory of truth must be conservative? [...] the transition [from the claim that truth is insubstantial to the claim that truth is conservative] has considerable intuitive force, for it seems extremely uncomfortable to maintain that truth is an insubstantial or non-robust property if the addition of truth principles leads one to rule out what were previously considered to be live possibilities concerning a (truth-free) subject matter. Perhaps the best

[34] See Theorem 5.2 above.

[35] Cfr infra, Section 9.6.1.

[36] There are also authors who advocate a semantic version of this conservativeness claim, but we will not discuss it here.

[37] See [**HL17**].

> way of understanding the transition is as a proposed explication: the informal notion of metaphysical insubstantiality is to be (possibly partially) explicated in terms of the formal criterion of conservativeness. It is striking, and a mark in favour of the plausibility of this understanding, that the conservativeness requirement has attracted considerable support among deflationists themselves. [**Wax17**, p. 445–446]

Horwich defends conservativeness deflationism for *philosophical* background theories. The background theories that he finds most relevant in this context are theories in metaphysics, epistemology, philosophy of science, ethics, action theory, and philosophy of language. Most theories in those disciplines are expressed in an informal manner, and are hard to formalise in an uncontroversial way. Thus, even though Horrwich's version of truth conservatism is proof theoretic at heart, it is not only due to its sensitivity to the question what our best truth theory is that this claim is in practice hard conclusively to establish or refute.

One may also wonder whether proof theoretic deflationism holds for scientific and mathematical background theories. In the literature on truth theory, only questions of proof theoretic conservativeness of truth theories over formal *arithmetical* theories has received much attention. Such questions are seen as tractable test cases for conservativeness deflationism. Moreover, such questions have a wider import. Since not only many 'higher' mathematical theories but also many highly theoretical scientific theories (such as relativity theory, for instance) have arithmetic embedded into them, non-conservativeness results for truth theories over arithmetic carry over to non-conservativeness results for truth theories over these theories. And since at least weak theories of arithmetic are interpretable in grammars, such non-conservativeness results also carry over to many linguistic theories. And one may wonder whether highly developed philosophical theories of meaning (such as compositional theories of meaning) might not also have grammars somehow built into them or at least presupposed. If so, then non-conservativeness results might also carry over to theories in philosophy of language, in which case Horwich's claim of conservativeness deflationism for philosophical theories would at least not be true across the board.

But in order to arrive at truly precise proof theoretical questions, we must move away from Horwich's minimal theory MT and return to the setting of previous Sections, where truth was seen as a predicate of *sentences* of a language, and where a truth theory is always formulated against the backdrop of a fully explicit theory of syntax (arithmetic). TB and PTB are then precise candidate counterparts of Horwich's somewhat vague theory MT.

We have seen in Section 5.1 that in the typed setting, disquotational theories are usually conservative over arithmetical theories. We have also seen that compositional truth theories tend to be non-conservative, and that in general, compositional truth axioms cannot be derived from disquotational axioms. Since standard typed compositional truth axioms are unobjectionable, it seems that compositional truth theories are *better* than disquotational theories. Hence, it has been argued, conservativeness deflationism does not hold for arithmetic ([**Hor95**], [**Sha98**], [**Ket99**]).[38] And this has caused many philosophers to take a somewhat dim view of conservativeness deflationism in general. Nonetheless, it should not be forgotten that

[38]For a recent discussion of conservativeness deflationism, see [**MR20**].

Horwich could still be *in large part* right. It might be that in most philosophical theories, arithmetic is not interpretable, and that as a result thereof, proof theoretic conservativeness even of compositional truth theories holds for those background theories.

5.3.4. Truth as a logico-linguistic concept. The fact that many philosophers had second thoughts about conservativeness deflationism has not resulted in truth theoretic deflationism becoming less popular. Rather, it has resulted in attempts to generate alternative reconceptualisations of the philosophical insubstantiality of the concept of truth. There are many such, but in this section, and in this monograph more generally, we briefly consider only one of them.

We have seen in Section 5.3.3 how the truth predicate is a tool for expressing certain forms of infinite conjunctions and disjunctions. This suggests the view that the truth predicate is closely related to the familiar logical connectives. Indeed, Field suggested that the concept of truth is a *logical* notion ([**Fie99**, p. 534]). According to the theory of *inferentialism* in the philosophy of logic, as championed for instance by Prawitz,[39] the meaning of the logical constants is given by introduction and elimination rules. A disquotational truth theory may be taken to contain a truth introduction (ascent) and a truth elimination (descent) component. Thus, if one takes disquotational axioms to give the *meaning* of the truth predicate, as some philosophers do, then the truth predicate seems closely related to the logical constants. One is thus led to the claim that truth is a *logico-linguistic* concept ([**HH02**]).

Since truth is a property of sentences, sentences are governed by grammar, and grammar is intertranslatable with arithmetic, it is no wonder that truth theory interacts with arithmetic. And then it is perhaps not surprising that truth is nonconservative over arithmetic. Thus the conception of truth as a logico-linguistic concept seeks to *demystify* the non-conservativeness of truth.

Given the machinery of coding, the natural numbers play a double role in formalised truth theories. They are not only what arithmetic is about, but also play (as codes) the role of what grammar is about. In the formalisation of truth theories, it is possible to disentangle syntax and arithmetic by keeping the entities that play the syntactical roles separated from the entities that serve as the subject of arithmetic. Formally, one distinguishes between two sorts of variables: variables that range over natural numbers on the one hand, and variables that range over linguistic expressions on the other hand. If these two sorts are kept separated by the formal theory, then the nonconservativeness phenomena disappear.[40]

This development can be seen as pointing to consilience between conservativeness deflationism on the one hand, and the doctrine of truth as a logico-linguistic doctrine on the other hand. But it need not be seen in this way, for the core of the doctrine of truth as a logico-linguistic notion is divorced from specific proof theoretic conservativeness claims. Indeed, perhaps the view of truth as a logico-linguistic concept should not commit itself to the conservativeness of truth. Even if one distinguishes between variables ranging over the natural numbers and variables ranging over linguistic expressions, non-conservativeness is not far away. If one

[39]See [**Pra83**].
[40]See [**LN13**].

adds certain natural bridge principles connecting these two sorts of entities, then the non-conservativeness phenomena reappear.

Since truth is in part a linguistic notion, it comes as no surprise that truth plays a substantive role in philosophy of language, as is witnessed by many decades of philosophical work on truth conditional theories of meaning. At the same time, it still seems reasonable that the notion of truth does not figure in an essential manner in philosophical disciplines in which arithmetic does not play a role.

The thought is that there then still is a sense in which truth is philosophically 'light', and the conception of truth as a logico-linguistic concept may still be called in a sense deflational. Whether that is right, is not quite clear, however. This is mainly due to the *linguistic* aspect of truth on this conception. Meaning is also a linguistic notion, for instance. But it is clearly a philosophically substantive one.

Nonetheless, at this point in the dialectic, it may be thought that the doctrine that takes truth to be a logico-linguistic notion is the better view, since it is not affected by the nonconservativeness of truth phenomena. In Chapter 9, we will see that this is not the end of the story: a connection between truth and proof theoretic reflection adds an interesting twist to it.

CHAPTER 6

Reflection principles in the mathematical sciences

In chapter 3, we distinguished between ontological and epistemic reflection relations and processes, and traced their philosophical history. Now we are ready to discuss the role that *principles* that describe these forms of reflection play in the mathematical sciences.

There are two main types of reflection principles in the mathematical sciences: *proof theoretic reflection principles*, which describe forms of ontological reflection, and *set theoretic reflection principles*, which describe forms of epistemic reflection. Both types of reflection play a role in *reducing incompleteness*: proof theoretic reflection principles reduce Gödelian incompleteness, whereas set theoretic reflection principles play a role in reducing incompleteness that stems from other sources, such as forcing. Of course we know since Gödel (Theorem 4.3) that in the mathematical sciences incompleteness can never be *completely* eliminated.

We have seen how epistemic reflection is related to mental processes that take place in time, which makes them *iterable* in a natural way. Accordingly, we will see that proof theoretic reflection principles are as a rule iterable in natural ways, and what the properties of iterations of proof theoretic reflection principles are. We have seen in Section 3.3 that certain forms of repetition are also a feature of ontological reflection. The direction of ontological reflection is "inward" into reality, so to speak, whereas the direction of epistemic reflection can rather be seen as "upward".

First, we will discuss proof theoretic reflection, which is an important theme in proof theory. We will distinguish purely mathematical reflection principles from reflection principles that can only be expressed using non-mathematical concepts such as truth and rational belief. Secondly, we discuss set theoretic reflection principles. We also address the question about the relation between set theoretic and proof theoretic reflection principles. In particular, we are interested in the question whether there is a strict dichotomy between the two kinds of mathematical reflection principles, or whether there is some sort of gradual spectrum in which reflection principles are situated. Thirdly, we have a look at reflection principles in probability theory. We will see that this is presently largely unknown territory. Nonetheless, I will argue that the subject holds promise: I recommend it for further investigation. All this then serves as preparation for later chapters, where we turn to philosophical questions of epistemic warrant for mathematical reflection principles.

6.1. Proof theoretic reflection principles

We are interested in the iteration of proof-theoretic reflection principles over formal 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*. This way of formalising soundness was already articulated by Kreisel and Levy in 1960 [**KL68**, p. 98]:

> 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.*

Such reflection principles cannot straightforwardly be formulated in mathematical settings [**KL68**, p. 98]:

> 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
>
> $$\forall p \forall a[Prov_S(p, a) \to T_S(a)].$$
>
> Such a truth definition T_S, does not exist [...]

This difficulty was circumvented by *approximating* the intended reflection principle by means of purely arithmetical principles, as we will shortly see. 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 $\mathcal{L}_T = \mathcal{L}_{PA} \cup \{T\}$, 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, over the past decade the resulting formal systems were related to a philosophical discussion about the function or role of the concept of truth.

We have seen how one important role for the concept of truth is to express and reason with generalisations over statements.[1] Exactly this function of the truth predicate is what allows us to express Kreisel's formalisation of soundness statements directly in the object language. Suppose that S is an arithmetical theory. Then, if we a standard provability predicate Bew_S for a given theory S, this *global reflection principle* for S can be expressed in the language $\mathcal{L}_T$ as follows [**KL68**, p. 98]:

GRF(S) **Global Reflection Axiom**:

$$\forall \text{ sentence } \varphi \in \mathcal{L}_T : \mathrm{Bew}_S(\varphi) \to T(\varphi).$$

Clearly, for this to have any operational meaning, typed Tarski biconditionals have to be added to the background theory S.

We have actually already encountered GRF(PA): Theorem 5.14 told us that GRF(PA) is non-conservative over PA for the language $\mathcal{L}_{PA}$.[2]

By Tarski's theorem on the undefinability of truth (Theorem 4.5), the language of arithmetic does not contain its own truth predicate, as Kreisel says in the quotation above. So *in the language of arithmetic* this guiding idea can only be *approximated* to varying degrees, and this is exactly what *arithmetical proof theoretic reflection principles* do. We can distinguish the following types of reflection principles for S:

[1] See Section 5.3.2.

[2] This theorem easily generalises for all theories $S \supseteq PA$.

Con(S) **Consistency**:

$$\neg \mathrm{Bew}_S(0 = 1)$$

Rfn(S) **Local Reflection Scheme**: For all closed sentences φ:

$$\mathrm{Bew}_S(\varphi) \to \varphi$$

RFN(S) **Uniform Reflection Scheme**: For all formulas $\varphi(x)$:

$$\forall x : \mathrm{Bew}_S(\varphi(x)) \to \varphi(x)$$

Here Con(S) is a (weak) proof theoretic reflection principle because it is (trivially) equivalent to $\mathrm{Bew}_S(0 = 1) \to 0 = 1$, i.e., it can be regarded as a weak soundness assertion.

Restricted versions for these principles are also considered: one can consider Rfn(S) (RFN(S)) for sentences (formulas) of a specific syntactic complexity. Δ_0-Rfn(S), for instance, is the local reflection principle for the Δ_0 fragment of S, and is equivalent to Con_S.

A particular "intermediate" proof theoretic reflection principle that has received some attention in the history of proof theory is the principle of *ω-consistency* of a theory S (ω-Con(S)):

$$\forall \varphi \in \mathcal{L} : \mathrm{Bew}_S(\exists x \varphi(x)) \to \neg \forall x \mathrm{Bew}_S(\neg \varphi(x)).$$

This principle is strictly stronger than Con(S): by the second incompleteness theorem, it is easily seen that $\neg\omega$-Con(PA), for example, is consistent but ω-inconsistent. That this principle can indeed be seen as a proof theoretic reflection principle follows from the following theorem of Smorynski [**Smo77**, p. 851]:

THEOREM 6.1. *For every finitely axiomatizable theory S in the language of arithmetic: S is ω-consistent if and only if Σ_2-Rfn(S+PA) holds.*

We concentrate on theories S that are formulated in the language of first-order arithmetic or an extension thereof. Moreover, we concentrate on theories that contain a good theory of its own syntax. In practice, this means that we shall mainly be dealing with theories that are at least as strong as Elementary Arithmetic (EA).[3]

It is also clear that:

$$(*) \qquad S + Con(S) \subseteq S + Rfn(S) \subseteq S + URF(S)$$

It then follow from Theorem 4.3 that for all minimally strong arithmetical theories S, all of Con(S), RFN(S), URF(S) are arithmetically non-conservative over S. Moreover, we will see later that all the inclusions in (*) are proper.[4]

Under mild assumptions, the global reflection principle is the strongest of them all:

PROPOSITION 6.2. *If S contains UTB, then $S + URF(S) \subseteq S + GRF(S)$.*

Uniform reflection is related to Hilbert's *ω-rule*, which we will call ωR ([**Hil96**, p. 1154]):

DEFINITION 6.3. An application of the *ω-rule* is an inference from the premises $\vdash \varphi(n)$ for each natural number n, to the conclusion $\vdash \forall x \varphi(x)$.

[3] Sam Buss's theory S^1_2 (see p. 109) is also strong enough for our purposes.

[4] See Corollary 6.16 below.

Unlike the familiar rules of inference, ωR is an *infinitary* rule: it has an infinite set of premises.

The ω-rule is much more powerful than the reflection principles that we have considered so far. If ωR is added to a modest theory of arithmetic, then the result is *true arithmetic*:

THEOREM 6.4. *For every true arithmetical sentence ϕ:*

$$EA + \omega R \vdash \phi.$$

PROOF. Straightforward mathematical induction on the complexity of ϕ. □

Unfortunately, the ω-rule for a (consistent) system S (extending EA) is *non-effective*, since there is no decision procedure for deciding, for arbitrary formulas $\varphi(x)$, whether for each natural number n, the formula $\varphi(n)$ is provable in S.

The rule-versions of local and uniform reflection will later also play a role:

RfR(S) **Local Reflection Rule**:

$$\frac{\vdash \varphi}{\vdash \mathrm{Bew}_S(\varphi)}$$

RFR(S) **Uniform Reflection Rule**:

$$\frac{\vdash \forall x \mathrm{Bew}_S(\varphi(x))}{\vdash \forall x \varphi(x)}$$

It is not hard to see that the local reflection rule is conservative. For any of the standard minimally strong arithmetical theories S that we have been considering, it is easy to see that [**Cie17**, Fact 13.1.1, p. 238]:

LEMMA 6.5. *S+RfR(S) is proof theoretically conservative over S.*

PROOF. For an arbitrary sentence φ, assume that $\mathrm{S} + \mathrm{RfR(S)} \vdash \varphi$. Since S is sound for Σ_1 sentences, for all ψ, if $\mathrm{S} \vdash \mathrm{Bew}_S(\psi)$, then $\mathrm{S} \vdash \psi$. So, for any proof of φ in S+RfR(S), we can systematically eliminate the uses of RfR(S), and transform it into a proof of φ in S. □

Surprisingly, the uniform reflection rule, in contrast, is equivalent to the uniform reflection principle [**Fef62**, Theorem 2.19]:[5]

THEOREM 6.6. *For any extension S of EA, S + RFR(S) is equivalent to S+ RFN(S).*

PROOF. The right-to-left direction is obvious, so we concentrate on the left-to-right direction.

We first show that for any formula $\varphi(x)$:

$$(^*) \qquad S \vdash \forall x, y : \mathrm{Bew}_S[\mathrm{Proof}_S(y, \varphi(x)) \to \varphi(x)].$$

We reason in S.
On the one hand, we have:

$$\mathrm{Proof}_S(y, \varphi(x)) \Rightarrow \mathrm{Bew}_S(\varphi(x)) \Rightarrow \mathrm{Bew}_S[\mathrm{Proof}_S(y, \varphi(x)) \to \varphi(x)].$$

[5]We here give Beklemichev's proof of this Theorem in [**Bek05**, Proposition 2.1].

On the other hand, we have

$$\begin{aligned}\neg\mathrm{Proof}_S(y,\varphi(x)) \Rightarrow\ & \text{(by } \Sigma_1\text{-completeness)}\\ & \mathrm{Bew}_S[\neg\mathrm{Proof}_S(y,\varphi(x))] \Rightarrow\\ & \mathrm{Bew}_S[\mathrm{Proof}_S(y,\varphi(x)) \to \varphi(x)].\end{aligned}$$

To conclude, applying the law of excluded third in S, we establish (*).

Now we reason in S + RFR(S). Applying RFR(S) to (*), we obtain

$$\mathrm{Proof}_S(y,\varphi(x)) \to \varphi(x).$$

From this, the desired result immediately follows. □

This surprising fact, which is often called *Feferman's little reflection theorem*, will play a role in later discussions.[6]

It is commonly thought that, over set theory, large cardinal principles are *much* stronger than proof theoretic reflection principles. Gödel, for instance, expresses this view as follows [**G̈46**, p. 151]:

> Any proof of a set-theoretic theorem in the next higher system above set theory (i.e. any proof involving the concept of truth [...]) is replaceable by a proof from such an axiom of infinity.

But this is too quick. As far as *consistency strength* goes, strong axioms of infinity are indeed generally much stronger than large cardinal principles. But as far as outright implication goes, this is not generally the case. For the Axiom of measurable cardinals,[7] for instance, we have:[8]

THEOREM 6.7. $ZFC + (MK) \nvdash Rfn(ZFC)$.

6.2. Iterating proof theoretic reflection

In this section we concentrate on *arithmetical* reflection principles. We postpone the discussion of global reflection until later.

6.2.1. Iterating reflection. 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 $\mathcal{R}(S)$ we denote the result of adding $\mathcal{R}(S)$ to S as follows:

DEFINITION 6.8. $\mathcal{R}[S] = S + \mathcal{R}(S)$.

Then we can define *iteration* of adding reflection principles thus:

DEFINITION 6.9.

(1) $\mathcal{R}^0[S] = S$;
(2) For α a successor ordinal, $\mathcal{R}^{\alpha+1}[S] = \mathcal{R}[\mathcal{R}^\alpha[S]]$;
(3) For λ a limit ordinal, $\mathcal{R}^\lambda[S] = \bigcup_{\alpha<\lambda} \mathcal{R}^\alpha[S]$.

In working with reflection iterations, use is made of the ordinal notation systems that were defined in Section 4.1.5.

Already for finite iterations of reflection principles, interesting phenomena emerge. One such phenomenon concerns the *relation between mathematical induction and*

[6] In particular in Section 7.4.2 and in Section 7.4.3.

[7] See p. 119.

[8] Thanks to Karl-Georg Niebergall for pointing this out to me.

uniform reflection. Firstly, there is an intimate connection between restricted mathematical induction and restricted uniform reflection [**Bek05**, Theorem 7]:

THEOREM 6.10. *For every* $n \in \mathbb{N} : I\Sigma_n = EA + RFN_{\Sigma_{n+1}}[EA]$.

This immediately entails an extremely tight connection between full mathematical induction and full uniform reflection:[9]

COROLLARY 6.11. $PA = EA + RFN[EA]$.

We will see that substantial fragments of second-order arithmetic extend this phenomenon, by proving transfinite iterations of uniform reflection.

Moreover, there are systematic relations between iterated consistency extensions, iterated local reflection extensions, and iterated uniform reflection extensions. Concerning the relation between consistency extensions and uniform reflection extensions, we have *Schmerl's theorem* ([**Sch79**], [**Bek95**, p. 27]):

THEOREM 6.12. *For ordinals* $\alpha \geq 1$:

$$PA + RFN^{1+\alpha}[PA] = PA + Con^{\varepsilon_\alpha}[PA].$$

So, in particular:

COROLLARY 6.13. $EA + RFN[EA] = EA + Con^{\varepsilon_0}[EA]$.

Concerning the relation between consistency extensions and local reflection extensions, Beklemishev has shown [**Bek95**, Theorem 1]:[10]

THEOREM 6.14. *For ordinals* $\alpha \geq 1$:

$$PA + Rfn^{\alpha}[PA] = PA + Con^{\omega^\alpha}[PA].$$

So, in particular:

COROLLARY 6.15. $EA + Rfn[EA] = EA + Con^{\omega}[EA]$.

Together, the previous theorems show that the inclusions in equation (*) on p.153 above are proper, i.e.:

COROLLARY 6.16.

$$S + Con(S) \subsetneq S + Rfn(S) \subsetneq S + URF(S).$$

Concerning the relation between uniform reflection and global reflection, we have ([**Lel23**]):

THEOREM 6.17. *The collection of purely arithmetical consequences of* $CT^- + GRF[PA]$ *coincides with the theory* $PA + RFN^{\omega}[PA]$.

So over a conservative compositional theory of truth (CT^-), global reflection is strictly the strongest reflection principle of them all.

[9] See [**KL68**].

[10] Actually, this is a special case of Beklemichev's Theorem 1, which is much more general.

6.2.2. Progressions. We now turn to the question how far reflection iterations can be extended, and to the question to what extent arithmetical incompleteness can be thereby reduced. In this effort, we again make use of the ordinal notation systems that were described earlier.[11]

Given a proof theoretic reflection principle $\mathcal{R}$, and given Kleene's ordinal notation system $\mathcal{O}$, we define the notion of a progression as follows:

DEFINITION 6.18. An *$\mathcal{R}$-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) $S_{lim(a)} = \bigcup_{b<a} S_b$.

Any progression thus 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 the relevant reflection principle for 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 three main results:

(1) Turing's completeness theorem for consistency progressions;
(2) Feferman's completeness theorem for uniform reflection progressions;
(3) Feferman's results about autonomous progressions.

Turing used consistency progressions in an attempt to reduce incompleteness in arithmetic. He proved the following theorem [**Tur39**]:

THEOREM 6.19. *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.*

At first sight this looks impressive, but, unfortunately the epistemological import of Turing's completeness theorem is limited. Theorem 6.19 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 already pointed out ([**Fra04b**, §6]), 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.[12]

One could try to get around this problem by insisting that only *natural* ways of defining S_ω are permitted. The question of how to characterise natural ways of defining S_λ, for λ a limit ordinal, occupied proof theorists for decades since the

[11] See Section 4.1.5.

[12] Turing and Feferman were acutely aware of this problem. For more on the philosophical significance of the use of non-canonical definitions see [**Fra04b**].

1960s. No satisfactory answer to this question has been obtained, and today it is widely seen as intractable.

The *intensional* aspect of defining S_ω is responsible, in Turing's theorem, for the complicated branching pattern at the limit stage.

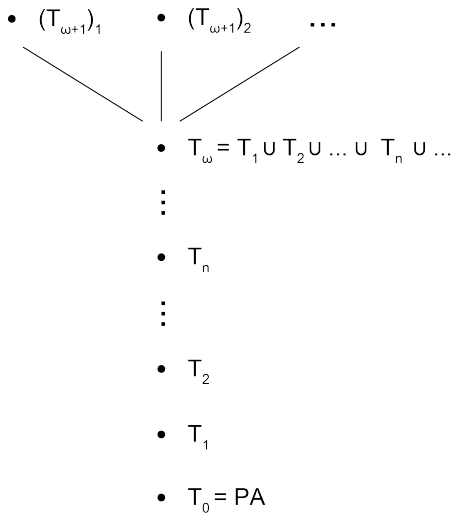

FIGURE 1

This branching phenomenon raises epistemic interpretation questions of its own. A natural idea, which we will explore more fully later, is that explicit acceptance of proof theoretic reflection principles is typically the result of an idealised mental reflection process. Reflection is thus a process that takes place in time, and time is *linear*. Against this, one can adopt a *branching time picture* of the relation between time and modality.[13] The basic idea here is that at branching points, time *could* continue in different ways. The interpretation would then be that for each true Π^0_1 sentence φ, the ideal epistemic agent could have recognised the truth of φ in $\omega + 1$ reflective acts. Nonetheless, the epistemic problem concerning the limit theories that we have discussed above still remains.

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 [**Fef62**]:

THEOREM 6.20. *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 $S_a \vdash \varphi$.*[14]

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

[13] See [**BMP23**].

[14] Feferman's completeness theorem can be strengthened. Using the notion of *smooth progression* developed in [**Bek95**] it can be shown that the length of this path can be shortened to ω^{ω^2+1}. For the idea of the proof of this improvement see [**Fra04b**].

the notations in this path is arithmetically complete. So here we do not have to entertain branching time scenarios.

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.[15]

6.2.3. 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 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 $\mathcal{O}$.[16] A progression that satisfies this additional criterion is called an *autonomous progression*.

Let us start by considering autonomous uniform reflection progressions over first-order Peano Arithmetic. Recall the hierarchy of systems RA_α of ramified analysis.[17] The following is a typical result, which is apparently "folkore":[18]

THEOREM 6.21. *The autonomous uniform reflection progression based on PA is the first-order fragment of the system of RA_ω, and the length of this progression is $\varphi_2(0)$.*

Hierarchies of systems of ramified analysis themselves also form progressions. Here, the engine is not made up of of proof theoretic reflection principles, but of successive "reifications" as sets of numbers of definitions in prior systems. We will investigate later whether, like proof theoretic reflection hierarchies, ramified analysis hierarchies can also be seen as obtained by processes of reflection.

Feferman and Schütte investigated autonomous progressions of predicative theories of analysis. They were able to express the length of the autonomous progression of systems of predicative analysis in terms of the Veblen hierarchy ([**Fef64**], [**Sch64**], [**Sch65**]):[19]

THEOREM 6.22. *The length of the autonomous progression of systems of ramified analysis is Γ_0.*

The second order system $RA_{<\Gamma_0}$ is therefore of special significance: Feferman claimed that it captures what one is *implicitly predicatively committed to* when one accepts PA. The Feferman-Schütte ordinal Γ_0 marks the limit of predicative reasoning, and is often referred to as the proof theoretic ordinal of predicativity.

Theorems 6.21 and Theorem 6.22 are from an epistemological point of view more significant than the earlier completeness theorems of Turing and Feferman (Theorem 6.19 and Theorem 6.20). In contrast to the non-autonomous progressions,

---

[15] It is also known that completeness depends on the choice of the path in $\mathcal{O}$. Feferman and Spector have shown in [**FS62**] that there are even paths *through* $\mathcal{O}$ such that the corresponding uniform reflection progression does not even prove every true Π^0_1 sentence.

[16] This notion of autonomous progression traces back to [**Kre60**].

[17] See p. 111.

[18] Thanks to Kentaro Fujimoto for pointing this out to me.

[19] See Section 4.1.5.2.

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, results such as Theorem 6.21 and Theorem 6.22 show *what we can come to know* in reflection progressions. Of course a strong idealisation is involved here: *we humans* are only able to go through a (small) finite number of stages of an autonomous progression before we die.[20]

6.3. Reflecting on truth

We 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 predicate (and perhaps also a falsity predicate).

6.3.1. Reflection, induction and compositionality. In this subsection, we will see how against the background of a weak disquotational theory of truth, proof theoretic reflection, mathematical induction, and truth are intimately related to each other.

6.3.1.1. *Typed.*

To start with, we assume full classical logic, and work with a typed concept of truth. After that, we will consider the leading questions in the framework of typefree truth and in the framework of partial logic.

We have discussed a connection between truth and proof theoretic reflection already: Theorem 5.14 tells us that adding compositional truth axioms to a background theory S enables one to prove $GRF(S)$.

In Section 6.2.1 we have also seen that, against the background of weak arithmetical theories, there is a close connection between proof theoretic reflection and mathematical induction. This connection also holds against the background of weak truth theories. One of the weakest disquotational truth theories that one can think of is $TB^-[EA]$, which has EA as its background arithmetical theory,[21] where the truth predicate is not allowed to occur in the bounded induction scheme, and which contains only the typed Tarski-bicontitionals. Then we have [**HL17**, Theorem 2]:

THEOREM 6.23. $RFN[TB^-[EA]] \vdash Ind(\mathcal{L}_T)$,

where $Ind(\mathcal{L}_T)$ is the *full* induction scheme for the language of truth $\mathcal{L}_T$. This can of course be seen as an extension to truth theory of the phenomenon that we have seen in the context of arithmetic.[22]

One application of uniform reflection, applied to $TB^-[EA]$, gives us even more. It gives us in addition the *uniform* typed Tarski-biconditionals [**HL17**, Theorem 2]:

LEMMA 6.24. $RFN[TB^-[EA]] \vdash UTB[PA]$.

This is the first indication that, when applied to a weak truth theory, adding uniform reflection makes *new truth laws* provable.

---

[20]For a discussion of the role of idealisation in the epistemological discussion of transfinite progressions of formal theories, see [**AMH19**].

[21]The theory EA was introduced on p. 109.

[22]See Corollary 6.11.

Against a disquotational background, one round of uniform reflection does not give us all the compositional truth axioms. Halbach has observed, however, that applying uniform reflection to UTB gives us CT [**Hal01a**, section 4]:

LEMMA 6.25. $RFN[UTB[PA]] \vdash CT$.

From this and Lemma 6.24 we then conclude that iterating uniform reflection over TB^- *twice* recovers typed compositional truth :

THEOREM 6.26. $RFN^2[TB^-[EA]] \vdash CT$.

In fact, we see that slightly more is true: if we start from $TB^-[EA]$, and apply uniform reflection twice, we obtain the full theory CT (over PA).

We have seen that there is a tight correspondence between systems of ramified analysis on the one hand, and systems of iterated compositional truth on the other hand (Theorem 5.19). This means that Theorem 6.22 on the implicit commitment of predicativity also has significance for the implicit commitment of *acceptance as true*, when truth is understood in a typed compositional sense.[23] The system $CT_{<\Gamma_0}$ can then be seen as expressing what one is implicitly committed to when one accepts PA as true (in a typed sense).

6.3.1.2. *Type-free.*

Turning to the classical typefree truth framework, we see that this phenomenon persists [**HL17**, Theorem 7]:

THEOREM 6.27. $RFN^2[TFB] \vdash Pos(KF)$,

where Pos(KF) ('*positive KF*') is a variant of Feferman's system KF.[24] Even though Pos(KF) and KF can for many purposes be seen as interchangeable—for instance, KF and Pos(KF) prove the same class of arithmetical statements—they are not outright equivalent. In $Pos(KF)$, the compositional axioms are restricted to the *positive fragment* of the language $\mathcal{L}_T$, whereas in Feferman's system KF the compositional axioms are completely unrestricted.

Returning to the classical setting, we observe that iterating reflection does not only recover compositional principles from disquotational ones. Indeed, we know that iterating the process of reflection also increases the amount of provable transfinite induction. It follows immediately from Corollary 6.11 and Theorem 6.12 that $RFN^2[EA]$ is a proper supertheory of PA.

Predicative analysis can be autonomously reached from KF, using uniform reflection as an engine:[25]

THEOREM 6.28. *The length of the autonomous reflection progression based on KF or Pos(KF), with uniform reflection as an engine, is* Γ_0.

COROLLARY 6.29. *The length of the autonomous reflection progression based on the disquotational theory TFB, with uniform reflection as an engine, is* Γ_0.

PROOF. This follows immediately from the previous theorem and Theorem 6.27. □

[23] Franzén's notion of 'accepting as sound' (or 'accepting as true') was introduced on p. 59.

[24] For a precise description of the axioms of Pos(KF), see [**HL17**, p. 225].

[25] This was proved by Fujimoto in unpublished work.

This Theorem and its Corollary is significant for the following reason. *Even if* one accepts the Quinean dictum *to be is to be the value of a variable* and is not prepared to reify classes of numbers as *sui generis* objects, one can still reach predicative analysis in an autonomous way starting from a positive disquotational theory of typefree truth.

Now suppose again 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 the following general theorem [**Lei16**, theorem 1.4]:

THEOREM 6.30. *For all $\kappa \in \mathbf{O}$ with $\kappa > 0$:*

(1) $CT_{\varepsilon_\kappa} = RFN^{1+\kappa}[TB_0]$;
(2) $Pos(KF)_{\varepsilon_\kappa} = RFN^{1+\kappa}[TFB_0]$.

Moreover, if we look at the consequences of these theories for the restricted language $\mathcal{L}_{PA}$, then we have the following connection with transfinite induction [**Lei16**, theorem 6.24]:[26]

THEOREM 6.31. *For all $\kappa \in \mathbf{O}$ with $\kappa > 0$:*

(1) *If A is an $\mathcal{L}_{PA}$-formula provable in $RFN^{1+\kappa}[TB_0]$, $RFN^{\kappa}[CT]$, or CT_{ε_κ}, then A is a theorem of $EA + TI(< \varepsilon_{\varepsilon_\kappa})$.*
(2) *If A is an $\mathcal{L}_{PA}$-formula provable in $RFN^{1+\kappa}[TFB_0]$, $RFN^{\kappa}[Pos(KF)]$, or $Pos(KF)_{\varepsilon_\kappa}$, then A is a theorem of $EA + TI(< \varphi_{\varepsilon_\kappa}(0))$.*

More in general, one may wonder what one is implicitly committed to when one accepts *as true* a mathematical theory such as PA, where 'true' is understood in a type-free compositional sense. One proposal would be to say that this implicit commitment is given by the following autonomouos progression of type-free truth theories, as follows. As a first step, one is then *explicitly* committed to KF (over PA). As a second step, one is *implicitly* committed to accepting KF *as sound* (or: *as true*). The latter involves the introduction of a second truth predicate (T_1), which is also self-applicative, and governed by the KF axioms. The third step is a repetition of step two, but for a new type-free truth predicate T_2, and so on.

The question what the autonomous ordinal of such a progression is, has been answered by work of Strahm and of Fujimoto ([**Str00**], [**Fuj11**, Theorem 16]):

THEOREM 6.32. *The autonomous ordinal of the above progression is $\varphi 200$, and the theory that is generated by this autonomous progression is $KF_{\varphi 200}$,*

where φ is the *ternary* Veblen function. I will not explain the ternary Veblen function in detail here.[27] Let it suffice to say that it is a natural generalisation of the ordinary ("binary") Veblen function: it gives us a nice ordinal notation system for larger countable ordinals than the ordinary Veblen function can deal with. In particular, the ordinal $\varphi 200$ is much larger than Γ_0, the ordinal of predicative analysis. All this is of relevance to the foundational discussion about predicativism, as we will now see.

[26] The arithmetical formalisation of principles of transfinite induction was discussed in Section 4.1.5.4.

[27] For the precised definition of the ternary Veblen function, see [**Str00**, Section 2].

We know that adding a next level in a Ramified Analysis hierarchy means recognising definable sets of natural numbers *as objects*. Given the tight correspondence between the Ramified Analysis hierarchy on the one hand, and the hierarchy of Tarskian compositional truth predicates on the other hand,[28] the same can be said about adding a next level in a Tarskian hierarchy. The system KF, however, is a stronger engine. Adding a next level in a KF-hierarchy adds more sets of numbers, since the notion of definability at play is more liberal, since it is "untyped". So if we let an autonomous definability hierarchy be powered by type-free truth in the sense of KF rather then by definability in the sense of Ramified Analysis, then we can obtain (truth-theoretic equivalents of) stronger subsystems of Analysis than $RA_{<\Gamma_0}$.

From a semantic point of view, KF asserts the existence of a *fixed point* of a particular monotone inductive operator.[29] Iterating KF along an autonomous path can then be seen as asserting the existence of *iterated* fixed points. This is formally captured by a theorem of Strahm, which says that this autonomous progression generates an impredicative subsystem of second-order analysis: it corresponds to the second-order axiom that states that for any set X, if X is a well-ordering, then the transfinite iteration of fixed points along this well-ordering X exists [**Str00**, Theorem 3].[30]

Fujimoto proposes to liberalise the strictures of predicativism in such a way that the kind of self-referentiality that is encapsulated in iterations of KF is also allowed. In his article on these matters ([**Fuj19**]), he concentrates on theories of second-order set theory rather than on theories of second-order arithmetic. But we have seen that the relevant issues in second-order number theory are perfectly analogous to the relevant issues in second-order set theory. The upshot is then, that according to *liberalised predicativism*, a larger fragment of PA^2 is justified than what is captured by $RA_{<\Gamma_0}$.

The situation in the non-classical (BDM) setting is structurally similar to that in the classical setting. Let EA_P be "*Elementary Arithmetic in partial logic*": it is formulated in the language with the truth predicate $\mathcal{L}_T$, formulated in *BDM* logic, containing all arithmetical axioms of EA except that it has an induction *rule* for Δ_0-formulae.[31] Clearly in the context of partial logic not only an induction axiom has to be replaced by the corresponding rule, but also uniform reflection axiom have to be replaced by the corresponding rule PRFN. In the context of Gentzen-style formalisation of logic,[32] this rule PRFN looks like this:

$$\frac{\Rightarrow \mathrm{Bew}^*_S(\Gamma(\dot{x}) \Rightarrow \Delta(\dot{x}),\, \Phi(\dot{x}) \Rightarrow \Psi(\dot{x})) \quad \Gamma(x) \Rightarrow \Delta(x)}{\Phi(x) \Rightarrow \Psi(x)} \quad (PRFN)$$

where the Bew^*_S expresses in the object language that the rule from $\Gamma(x) \Rightarrow \Delta(x)$ to $\Phi(x) \Rightarrow \Psi(x)$ is an admissible rule of S.

In [**FNH17a**, Section 3.3, Proposition 3] it is shown that two applications of uniform reflection over EA_P proves the principle of transfinite induction for the language $\mathcal{L}_T$ for all ordinals up to and including ω^ω:

---

[28] See Theorem 5.19.

[29] See Theorem 5.25.

[30] This second-order axiom can be seen as a strengthening of the system ID_1, which was briefly discussed on p. 112.

[31] See [**FNH17a**, Section 2.2] for more details.

[32] For an introduction to Gentzen-style formalisations of logic, see [**TS96**, Chapter 3].

THEOREM 6.33. $PRFN^2[EA_P] \vdash TI_{\mathcal{L}_T}(\omega^\omega)$

Iterating reflection into the transfinite proves even more transfinite induction, as it is shown in [**FNH17a**, Subsection 3.3, Corollary 3]:

THEOREM 6.34. $PRFN^\omega[EA_P] \vdash TI_{\mathcal{L}_T}(< \omega^{(\omega^2)})$

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. The reason for this is that EA_P is formulated in the non-classical logic BDM.

Moreover, when not EA_P but TS_0 is taken as a starting point, then we have the following general connection between reflection and transfinite induction [**FNH21**, Proposition 2]:

THEOREM 6.35. $PRFN^{\omega_n+1}[TS_0] \vdash TI_{\mathcal{L}_T}(\omega_n)$.

The recovery of compositionality through reflection also extends to the type-free non-classical context of partial logic [**FNH17b**, corollary 1, section 3.2]:

THEOREM 6.36. $PRFN^2[TS_0] \vdash PKF$.

In sum, PKF proves less transfinite induction than KF. The situation is probably different when we consider the *schematic version* PKF(P) of PKF, which is defined from PKF in the same way as KF(P) is defined from KF. In unpublished work, Fischer sketches an argument for the following:[33]

THEOREM 6.37. $PKF(P) \vdash TI_{<\Gamma_0}$.

Given Theorem 5.31, we then see that the first-order mathematical strength of PKF(P) coincides with that of KF(P).

6.3.2. Global reflection over a truth theory. ??

We have seen that the pioneers of proof theoretic analysis of proof theoretic reflection principles concentrated on consistency and uniform reflection statements rather than on global reflection.[34] I speculated that one reason for this was that making use of a primitive truth predicate was seen as an appeal to a philosophical notion that mathematical logicians should strive to avoid. But we will now see that there are other reasons why global reflection poses problems.

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, since it makes iterated truth ascriptions provable. But from a type-free perspective, $GRF[S]$ may be a plausible way of making the commitment that is implicit in accepting type-free truth theory S explicit.

We have seen earlier in this chapter that KF can coherently be closed under repeated applications of uniform reflection, and how this has given rise to research into transfinite iterations of uniform reflection on KF. On the other hand, it follows immediately from Lemma 5.32 that KF cannot consistently be closed under global reflection:[35]

THEOREM 6.38. $GRF[KF] \vdash \bot$.

[33] As I write this, some details of his argument remain to be verified.

[34] See Sections 6.2.2 and 6.2.3.

[35] This is a folklore result. For a proof, see [**FNH17a**, Footnote 11, p. 2638].

This has been seen as a damning feature of KF.

We saw earlier that two applications of uniform reflection yield the system Pos(KF), which is a version of KF (Theorem 6.27). It turns out that there is an important theoretical difference between Pos(KF) and KF. Zicchetti has shown that, in contrast to KF, the system Pos(KF) can be coherently closed under (repeated applications of) global reflection ([**Zic23**, Theorem 1]):

THEOREM 6.39. *$GRF[Pos(KF)]$ is arithmetically sound.*

This phenomenon might deserve more attention than it has received thus far. Perhaps it can be taken as a reason for preferring Pos(KF) over KF. The system $Pos(KF) + GRF[Pos(KF)]$ is formulated in classical logic. Therefore, by an application of the diagonal lemma, it still proves:

$$(\lambda \wedge \neg T(\lambda)) \vee (\neg\lambda \wedge T(\lambda)),$$

where λ is again the liar sentence. As we have seen earlier, this may still be regarded as objectionable. But every axiomatic theory of truth that is based on classical logic proves it, it is more palatable than proving a contradiction!

As opposed to in the classical framework, in the partial framework BDM everything works smoothly and naturally. Of course, in the partial framework, proof theoretic reflection has to be expressed as rules rather than as axioms. But if we do that,[36] then over a weak disquotational theory such as TS_0, uniform reflection and global reflection coincide [**FNH17b**, Proposition 1]:

THEOREM 6.40. $RFN[TS_0] = GRF[TS_0]$.

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 *fully coherent* with its implicit commitment.

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. In what follows, we will discuss a different procedure to make the implicit acceptance of a theory explicit.

6.4. Reflecting on believability

In his book *The lightness of truth* ([**Cie17**]), Cieśliński aims at justifying reflection principles, not by using the concept of truth and principles that govern it, but by using principles governing the notion of *believability*, which is an epistemic notion. Intuitively, the expression 'φ is believable' means that there is a good reason to accept φ [**Cie17**, p. 251]. After the publication of his book, Cieśliński revised the basic principles governing believability in a Corrigendum ([**Cie20**]). Our discussion will therefore be based on the version of the theory of believability that is found in the Corrigendum.

Let the background theory be PA. The aim is to find a believability theory *over PA*, which we will call Bel(PA).[37] The language of believability theory ($\mathcal{L}_B$) is $\mathcal{L}_{PA}$, extended with a new predicate B (for believability). The theory Bel(PA) then contains the following believability axioms and rules [**Cie20**, p. 3–4]:

B1 PA, formulated in the extended language $\mathcal{L}_B$;

[36] For the details, see [**FNH17b**, Section 2.4].

[37] In Cieśliński's terminology of [**Cie20**], this theory would be called $Bel^*(PA)$.

B2 $\forall \varphi \in \mathcal{L}_B : Bew_{PA}(\varphi) \rightarrow B(\varphi)$;
B3 $\forall \varphi, \psi \in \mathcal{L}_B : [B(\varphi) \wedge B(\varphi \rightarrow \psi)] \rightarrow B(\psi)$;
B4

$$\frac{\vdash B(\forall x B(\varphi(x)))}{\vdash B \forall x \varphi(x)}.$$

B5

$$\frac{\vdash \psi}{\vdash B(\psi)}.$$

The idea is that each of these principles should respect a basic aspect of the content of the notion of believability. For instance, Axiom B2 expresses that all theorems of Peano Arithmetic are believable, i.e., there are good reasons to believe all of PA.

Believability is intended to be a *defeasible* notion. It may be the case that there are not only good reasons to believe φ, but also good reasons for believing $\neg\varphi$. In such a case, it is not always rational to proceed from a belief in the believability of φ to a belief of φ [**Cie17**, p. 251].

The believability theory over PA model-theoretically behaves as it should [**Cie20**, Theorem 5, p. 4]:

THEOREM 6.41. *There is a standard model $\mathcal{M}$ for $\mathcal{L}_B$ such that for all sentences $\varphi \in \mathcal{L}_B$ such that $Bel(PA) \vdash B(\varphi)$, $\mathcal{M} \models \varphi$.*

In analogy with Reinhardt's truth theory IKF,[38] the *internal theory* of Bel(S) is defined as follows:

DEFINITION 6.42. $IBel(S) \equiv \{\varphi \mid Bel(S) \vdash B(S)\}$

The believability theory for PA does not prove any new arithmetical statements:

THEOREM 6.43. *Bel(PA) is proof theoretically conservative over PA for $\mathcal{L}_P A$.*

But the internal theory of Bel(PA) proves reflection principles for PA [**Cie20**, Theorem 8, p. 5]:

THEOREM 6.44. *For every $n \in \mathbb{N} : RFN^n(PA) \in IBel(S)$.*

In words: IBel(PA) proves for every instance of finitely iterated uniform reflection for PA that it is believable. Maciej Głowacki and Mateusz Łełyk have recently shown in unpublished work that this is also the *exact* mathematical strength of IBel(PA) ([**GL23**]):

THEOREM 6.45. *For any $\varphi \in \mathcal{L}_{PA}$:*

$$Bel(PA) \vdash B(\varphi) \Leftrightarrow \exists n \in \mathbb{N} : RFN^n(PA) \vdash \varphi.$$

In the absence of any good reasons against iterated uniform reflection principles, then, it is rational to come to believe uniform reflection principles on the basis of the proof of their believability in Bel(PA).

Believability theory over background theories other than PA also yield interesting results. For instance, believability theory over disquotational truth with restricted induction proves the believability of the full compositional theory of truth [**Cie20**, Theorem 10, p. 7]:

THEOREM 6.46. $Bel(TB^-) \vdash B(CT)$.

[38] IKF was introduced on p. 142.

Again, in the absence of reasons that speak against CT, this means that from a proof of $B(CT)$ in $Bel(TB^-)$, the full typed theory of compositional truth may be inferred. Similarly, one can consider the believability theory over a weak *untyped* disquotational theory, such as PUTB.[39] Then it is not surprising that one obtains, in analogy with theorem 6.46, the following phenomenon [**Cie17**, Theorem 13.4.18, p. 266]:

THEOREM 6.47. $Bel(TFB^-) \vdash B(KF)$.

This means that the believability of a natural untyped compositional truth theory is entailed by the belief theory over a natural weak set of untyped disquotational axioms.

6.5. Set theoretic reflection

Set theoretic reflection is an informal concept: there exists no precise characterisation of the concept of set theoretic reflection and of set theoretic reflection principle. That being said, a set theoretic reflection states a form of ontological reflection. It says that the set theoretic universe V as a whole is similar to one or more small parts P of V. Here the similarity relation is made precise in a *semantic* way: in terms of the relation of truth in a structure, where the structure is V, or some part of V. The "smallness" of the reflecting structure is made precise as meaning *set*-sized, or without loss of generality, as the reflecting structure being an initial segment V_α of V. Martin expresses the modern concept of reflection as follows [**Mar76**, p. 85–86]:

> Reflection principles are based on the idea that the class ON of all ordinal numbers is so large that, for any reasonable property P of the universe V, ON is not the first stage α such that $[V_\alpha]$ has P.

This concept of set theoretic reflection seems to be widely accepted. A variant of it is found in [**Ber61**, p. 6], and variants of it can also be found in the contemporary literature.[40]

The concept of set theoretic reflection principle finds a clear expression in the principle of *Montague-Levy reflection*. If we denote the relativisation of the quantifiers of a formula φ to V_α as φ^{V_α}, then this schematic principle for sentences of $\mathcal{L}_{ZFC}$ can be expressed as follows

(ML) $$\varphi \rightarrow \exists \alpha : \varphi^{V_\alpha}$$

Montague and Levy showed that this principle is provable in standard set theory ([**Lév60a**], [**Mon61**]):

THEOREM 6.48. $ZFC \vdash \varphi \rightarrow \exists \alpha : \varphi^{V_\alpha}$.

Nonetheless, Montague-Levy reflection has hidden strength. Over ZCF *minus* the axiom of infinity and the axiom of replacement (call this theory ZC^-), ML is equivalent to the remainder of the axioms of ZFC:

THEOREM 6.49. $ZC^- \vdash ML \Leftrightarrow (\ \textit{Infinity} \ + \ \textit{Replacement})$.[41]

[39] See section 5.1.3.

[40] See for instance [**Inc16**, p. 163] or [**BT23**, p. 5].

[41] The Axiom of Choice play no role in the proof of this equivalence.

The weakest well-known large cardinal principle, the Axiom of Inaccessible Cardinals (IC),[42] can then be rephrased as a set theoretic reflection principle in this sense, namely as follows:

For every sentence $\varphi \in ZFC^2$:
if $(V, \epsilon, \mathcal{C}) \models \varphi$, then there is an ordinal α such that $(V_\alpha, \epsilon, V_{\alpha+1}) \models \varphi$.

(The reason for the equivalence with the standard expression of IC is of course that we know the antecedent of the implication to be true.) We have seen that IC is independent of ZFC. This shows that set theoretic reflection principles can have strength even against the background of ZFC.

A natural second-order (i.e., class-theoretic) analogues of (ML) can straightforwardly be formulated. Bernays formulated the following principle (BR, for "Bernays reflection") for sentences φ of $\mathcal{L}_{ZFC^2}$ ([**Ber61**]):

(BR) $$\varphi(X) \rightarrow \exists \alpha : \varphi(X \cap V_\alpha)^{V_\alpha, V_{\alpha+1}},$$

where, X is the only free variable occurring in φ, and the superscripts in $\varphi(X \cap V_\alpha)^{V_\alpha, V_{\alpha+1}}$ indicates that not only are the first-order quantifiers of $\varphi(X \cap V_\alpha)$ are restricted to V_α, but furthermore the second-order quantifiers of φ are restricted to $V_{\alpha+1}$.

The first-order fragment of $ZFC^2 + BR$ is significantly stronger even than $ZFC + IC$. Since the Axiom of Weakly Compact Cardinals is in effect just Π^1_1-BR,[43] it is immediate that Bernays reflection (Axiom BR) entails that there exist weakly compact cardinals (WCC):

Proposition 6.50. $ZFC^2 + BR \vdash WCC$.

Nonetheless, $ZFC^2 + BR$ does not prove the existence of *large* large cardinals.

Once Axiom BR is formulated as a natural second-order analogue of the first-order Montague-Levy principle, it is natural to consider the natural *third-order* analogue of Axiom BR. We will not state this third-order principle precisely here, for it has been shown to be inconsistent. Moreover, it has been shown that even semi-natural *consistent* higher-order strengthenings of Axiom BR do not prove the existence of large large cardinal principles.[44]

Nonetheless, Axiom BR also has significant *class theoretic* consequences.

Theorem 6.51. $NBG + BR \vdash MK$

In words: adding Axiom BR to the mild and predicative class theory NBG yields the impredicative class theory MK.

Moreover, Axiom BR yields a class theoretic strengthening of the Axiom of Choice, which is known as the principle of *Global Choice* (GC):

(GC) $$\exists \text{ class function } f : V \xrightarrow[1-1]{onto} Ord,$$

where Ord is the class of ordinal numbers.

Theorem 6.52. $NBG + BR \vdash GC$.

[42]The Axiom of Inaccessible Cardinals was discussed on p. 118.

[43]The Axiom of Weakly Compact Cardinals was discussed on p. 118.

[44]For a discussion of higher-order analogues of Axiom BR, see [**Tai05**] and [**Koe09**].

Thus Axiom BR implies that not just every set can be well-ordered, but that even a well-ordering of the set theoretic universe as a whole exists.

Gödel thought that *all* large cardinal principles can be derived from set theoretic reflection principles ([**Wan96**, Section 8.7.9]):

> All the principles for setting up the axioms of set theory should be reducible to Ackermann's principle: The Absolute is unknowable. The strength of this principle increases as we get stronger and stronger systems of set theory. The other principles are only heuristic principles. Hence, the central principle is the reflection principle, which presumably will be understood better as our experience increases.

In a different place, he describes a similar thought later as follows ([**Wan96**, Section 8.7.16]):

> Generally I believe that, in the last analysis, every axiom of infinity should be derivable from the (extremely plausible) principle that V is indefinable, where definability is to be taken in [a] more and more generalized and idealized sense.

Most contemporary set theorists find Gödel's claim fanciful. Koellner, for instance, concludes from the failure of Bernays reflection to extend in a natural way to much stronger set theoretic reflection principles, that Gödel's claim cannot be right. In particular, Koellner claims that there are no set theoretic reflection principles that entail the existence of large large cardinals ([**Koe09**]).

Recent work on set theoretic reflection principles shows, however, that Koellner's claim may be premature. Philip Welch has proposed the following set theoretic reflection principle, which is in the literature mostly called the *Global Reflection Principle*,[45] but which we will abbreviate as WR2 (WR stands for 'Welch Reflection'):[46]

Axiom 6.53 (WR2). There is an initial segment of the universe V_κ, and a nontrivial elementary embedding

$$j : (V_\kappa, \in, V_{\kappa+1}) \longrightarrow_2 (V, \in, \mathcal{C})$$

with critical point κ, and where e is an elementary equivalence relation.

Bernays' reflection principle BR states that the second-order theory of V is reflected *point-wise*, sentence by sentence. Welch Reflection, in contrast, postulates that there is a *single* initial fragment V_κ of V that reflects the entire second-order theory of V—hence the qualification "global". Axiom 6.53 postulates that V, together with all of its classes, is reflected in a *set-sized*, and therefore *small* part (V_κ) of the universe, together with all of its sub-sets ($V_{\kappa+1}$). Thus WR2 is a set theoretic reflection principle in the sense that we have been using the expression.

WR2 is as strong as the Axiom of 1-Extendible Cardinals (Axiom 4.27), which we know to be one of the strong large large cardinal principles—albeit by no means the strongest. Also, like BR, the principle WR2 proves all the axioms of the impredicative class theory MK, and it proves the existence of a global well-ordering of V.

[45] The term *global reflection* unfortunately has a prior use, as we have seen earlier: it is used to refer to the proof theoretic principle GRF that was introduced on p. 152.

[46] For a discussion of Welch Reflection, see [**WH16**].

It is important to note that j plays an essential role in WR: merely postulating elementary equivalence between V and some initial fragment V_κ does not yield significant large cardinal strength. On the basis of this, Koellner might claim that WR2 is not really a set theoretic reflection principle after all, because the class function j has no counterpart in our informal notion of set theoretic reflection principle.

In any case, against this, and without going into details, it can be remarked that WR2 can be "split" into a principle stating that there is a single rank V_κ that reflects the whole second-order theory of V (where no mention is made of j), and a very strong (third-order) choice principle on the other hand. Each of these two principles is weak in large cardinal strength on its own, but in combination they yield the strong principle WR2.

One might consider a somewhat weakened version of WR2, in which the notion of elementarity is first-order ($\longrightarrow_1$) rather than second-order. Let us call this principle WR1. Against the background of a predicative theory of classes such as NBG, the principle WR1 still entails the existence of large large cardinals. But unlike WR2, over NBG the principle WR1 does not prove the axioms of the impredicative class theory MK.

Welch regards the impredicative class theory MK with suspicion, and takes weaker class theories such as NBG to be more plausible. For this reason, he prefers WR1 over WR2. Against this, however, it can be said that WR1 still entails the existence of a global wellordering of V,[47] which is widely regarded as impredicative in nature.

In [**Rob17**], Sam Roberts seeks to generalise Bernays reflection in a way that the resulting reflection principle proves large large cardinal principles while not running into the generalisation problems identified by Reinhardt, Tait, and Koellner.

Roberts starts from an informal *general principle*, which we may call *Roberts reflection* (RR):

> For any formula φ: if φ is true in a (possibly proper class sized) structure $\mathcal{S}_1$ consisting of Φs and Ψs and Θs and..., then φ is also true in a *small* structure $\mathcal{S}_1$ consisting of Φs and Ψs and Θs and...

Observe that set theoretic reflection principles concern situations where the structures involved consist only of sets (the Φ) and of classes (the Ψs), and where the formulas φ range over the language $\mathcal{L}^2_\epsilon$. We obtain a first precise Robertsonian reflection principle RR1 from the general idea RR by focussing on such structures, such sentences φ, by (as usual) taking "small" structures to mean set-sized structures, by requiring the reflecting set to be a rank V_α, and by requiring the classes of the reflected structure $\mathcal{S}_1$ to be *plentiful* in the following sense:

Definition 6.54. A collection of classes $\mathcal{C}$ is plentiful for a set s if and only if for every $x \in s$, there is a class $C \in \mathcal{C}$ such that $x = C \cap s$.

With these choices made, we postulate:

Axiom 6.55 (RR1). For any formula $\varphi \in \mathcal{L}^2_\epsilon$, if φ is true in $\langle V, \mathcal{C}\rangle$, then there is a rank V_κ and a set of classes $\mathcal{C}'$ that is plentiful for V_κ, such that φ is true in $\langle V_\kappa, \mathcal{C}'\rangle$.

[47] Thanks to Sam Roberts for pointing this out to me.

Roberts then observes:

PROPOSITION 6.56. *$NBG \vdash RR1 \Leftrightarrow BR$.*

So far, so good: nothing spectacular is happening. Now suppose we extend the language over which φ ranges with a satisfaction predicate $Sat(x, Y)$ (intended to be interpreted as: "the formula x is satisfied by Y, where the class Y codes an assignment to first- and second-order variables") for the language $\mathcal{L}^2_\epsilon$, yielding the language $\mathcal{L}^2_{\epsilon,Sat}$. Then the set theoretic reflection principle RR1 can be straightforwardly generalised to the extended language yielding the set theoretic reflection axiom RR2:

AXIOM 6.57 (RR2). For any formula $\varphi \in \mathcal{L}^2_{\epsilon,Sat}$, if φ is true in $\langle V, \mathcal{C}\rangle$, then there is a rank V_κ and a set of classes $\mathcal{C}'$ that is plentiful for V_κ, such that φ is true in $\langle V_\kappa, \mathcal{C}'\rangle$.

Now let CT_{Sat} be the usual compositional satisfaction axioms for the language $\mathcal{L}^2_\epsilon$, i.e., the natural analogues for Sat of the typed compositional axioms for the truth predicate T. Then against the background of the modest class and truth theory $NBG + CT_{Sat}$, which is a very modest extension of NBG, the principle RR2 is surprisingly strong:

THEOREM 6.58. *$NBG + CT_{Sat} + RR2 \vdash WR2$.*

We even have:

THEOREM 6.59. *$NBG + CT_{Sat} + RR2$ proves that there is a proper class of 1-extendible cardinals.*

The latter theorem shows that RR2 is even a bit stronger than WR2.[48] It is not, however, *much* stronger than GRP, since from the existence of a 2-extendible cardinal, the consistency of RR2 can be proved.

In one important respect, the idea behind RR2 *differs* from the idea behind WR2. Firstly, according to WR2, formulas are reflected to a structure $\langle V_\kappa, V_{\kappa+1}\rangle$ of *sets*. According to RR2, formulas are reflected to structures $\langle V_\kappa, \mathcal{C}'\rangle$, where $\mathcal{C}'$ can easily contain proper classes.[49] In another, equally important respect, RR2 is very closely related to WR2. The *plentifulness requirement* on $\mathcal{C}'$ entails that $\langle V_\kappa, \mathcal{C}'\rangle$ is a "full" second-order structure, in the same sense as the requirement in WR2 that *all* subsets of V_κ are sent by j to classes over V. This fullness or plentifulness requirement plays a crucial role in the strength of WR2 and RR2.

One advantage of RR2 over WR2 is the absence of an embedding function j: in RR2, parameters are not reinterpreted at all. A second advantage of RR2 over WR2 is its *generality*. It is intended to work not just in our setting, where the Ψs are classes, but also in settings where WR2 would falter, for instance where the Ψs are *intensional* entities such as properties.[50]

The absence of an embedding function in RR2 makes it conceptually less close than WR2 is to the ontological reflection idea that goes back to Philo of Alexandria,[51] and dissociates it from the thought that the Absolute is unknowable. According to Philo's form of ontological reflection, the mathematical Absolute (i.e.,

[48]Perhaps this has to be taken with a grain of salt, for we are not comparing the two principles over the same background theory.

[49]After all, V, for instance, is allowed as a parameter in φ.

[50]It is easy to see that WR2 would "extensionalise" such entities.

[51]See Section 3.3.

proper class sized entities) are transcendent in themselves and are reflected in immanent entities (i.e.,, in pure sets).

Thus set theoretic reflection reaches the realm of 1-extendible cardinals. Can stronger principles of infinity be reached by set theoretic reflection principles?

Victoria Marshall has investigated a class of very *higher-order* (i.e., of higher order than second-order) set theoretic reflection axioms,[52] and Rupert McCallum has recently investigated related principles.[53] Certain versions of their reflection principles prove the existence of supercompact cardinals, and even go beyond supercompactness. I will not go into the details of these reflection principles here, even though their work certainly deserves more attention that it has received so far.

One philosophical question that immediately arises when one considers the reflection principles of Marshall and of McCallum, is: *How we can make philosophical sense of the these higher-order principles?* First, one must climb the type theoretic hierarchy above the sets to some degree, and rationally come to accept classes of αth order, for relatively small ordinals α. This is by no means an easy task.[54] Once one has achieved this,[55] it is perhaps natural to come to accept classes of αth order for *any* ordinal α. If one then also accepts the principle that *the mathematical realm abhors the potential*,[56] then it is natural to accept the existence of V with its classes of arbitrarily high orders as an actual infinity, whereby quantification over classes of every order is admissible. Furthermore, if the embedding formulation of supercompactness discussed earlier[57] serves as a guide for understanding the reflection principles of Marshall and of McCall that entail the existence of supercompact cardinals,[58] then it is clear that all this is necessary and sufficient for their principles to have a determinate meaning. Having a determinate meaning is one thing, truth is another. In order for their principles to be warranted, the precise reflection conditions that these reflections impose on V with its classes of all ordinal orders have to be made plausible. That again seems to me a highly non-trivial task, which has hitherto not been delivered upon (in my opinion).

It is not necessary to countenance *higher-order* classes of sets in order to formulate very strong principles that have been argued to qualify as plausible set theoretic reflection principles. The next principle that we encounter in our discussion of ever stronger set theoretic reflection concerns a form of reflection between *structures*.

Definition 6.60. A *relational set structure* is a set A belonging to V, equipped with a number of relations R_i for i belonging to some index set I.

Bagaria has proposed what he calls the *structural reflection principle* (SR) for classes of structures of the same type ([**Bag23**]):

Axiom 6.61 (SR). For every $\mathcal{L}_{ZFC}$-definable, possibly with parameters, class $\mathcal{C}$ of relational structures of the same type, there exists an ordinal α that reflects $\mathcal{C}$, i.e., for every $A \in \mathcal{C}$ there exists a $B \in \mathcal{C} \cap V_\alpha$ and an elementary embedding $j : B \mapsto_1 A$.

[52] See [**Mar89**].
[53] See [**McC21**].
[54] We will discuss this later in some detail: see Section 9.8.2.
[55] I have not.
[56] As I do: see p. 3.
[57] See p. 120.
[58] As it does, in fact.

Earlier we have defined Vopenka's Principle (VP) as a second-order axiom.[59] Just as the full impredicative induction axiom of second-order arithmetic is approximated by the first-order induction scheme of PA, the second-order axiom VP is approximated by the first-order axiom scheme VP^*, which restricts VP to *definable* classes of relational set structures. The definable version VP^* of Vopenka's Principle still has much more large cardinal strength than the Axiom of 1-extendible cardinals (Axiom 4.27).

Bagaria has shown that:

THEOREM 6.62. *The structural reflection principle SR is equivalent to VP^*.*

Again we can ask the question whether SR is a set theoretic reflection principle in our sense of the word. And again the answer to this question is no, but for a different reason. Recall that on the conception of set theoretic reflection principles with which we are operating, the similarity between V and the reflecting structure is cashed out in *semantic* terms: if V makes a sentence (belonging to some class Φ) true, then the reflecting structure does so, too. The principle SR interprets the similarity relation in *ontological* terms: the *existence* of any given class of relational set structures in V is reflected in the existence of a similar class of structures in some V_α. Bagaria is aware of all this: he argues for exactly this reason that the standard notion of set theoretic reflection principle should be somewhat liberalised [**Bag23**, Section 2].

At this point, one may start to wonder which principles count as set theoretic reflection principles. In order to assess this, it is important to keep the difference between reflection *within* the universe and reflection *of* the universe in mind. Also, we should not forget that it is part of our conception of set theoretic reflection principles that they express reflection in *small* parts of the universe.

Are perhaps all natural elementary embedding axioms (such as Axiom 4.22) are set theoretic reflection principles? According to the fairly widely accepted concept of set theoretic reflection principles that we have been working with so far, the answer is no. The reason is that in elementary embedding principles, the "reflecting structure" (the inner model M), is proper class sized, and therefore not "small". This difference between elementary embedding principles and genuine set theoretic reflection principles has mathematical consequences. Elementary embedding principles do not entail that there exists a global wellordering of V, for instance, whereas even mildly strong set theoretic reflection principles do.

In the corridors of set theory, one sometimes hears people expressing the view that the standard embedding formulations of strong principles of infinity all deserve to be called natural set theoretic reflection principles. I believe that these set theorists may well be right. There is something to be said for extending our concept of set theoretic reflection principle so as to include standard embedding axioms. So I will now briefly make a plea for dropping the "smallness" condition on the reflecting object from the concept of set theoretic reflection principle.

First of all, embedding principles express forms of *self-reflection* of the universe V: they express that V is reflected in a *part* of itself. Secondly, we know from our discussion of embedding principles that smallness (or immanence) of the reflecting object is not a necessary condition for mathematical strength. Thirdly, it has not been made clear, I believe, in which sense proper classes are less epistemically

[59] See p. 121.

accessible than (complicated) sets. We seem to have some grasp on the laws that govern proper classes, just as we have a grasp of the laws that govern sets. For this reason, the thought that set theoretic reflection principles express a reflection of the unknowable in the knowable does not fully convince.

I will not insist in what follows on this proposed extension of the "official" concept of set theoretic reflection principle in what follows. But there may be something to be said for it. At least the hope is that ontological reflection in set theory is something akin to a natural kind and that set theoretic reflection principles intend to capture this property. If so, then set theoretic embedding principles ought to be seen as reflection principles.

In the set theoretic literature, elementary embedding axioms are often "cut down to size" and thereby "firstorderized". Instead of postulating class embedding functions from V into some inner model M, these miniaturisations postulate the existence of a *set-sized* embedding function from a rank initial segment of V to a taller but thinner set-sized part of V. From a mathematical point of view, the miniaturisations do as well as their big sisters. It is clear, however, that these miniaturisations do not count as set theoretic reflection principles in the usual sense of the word. They describe reflection *in* the universe rather than reflection *of* the universe. Again we may ask: might such first-order miniaturisations not in the eyes of some who are sceptical of proper classes be candidates for being basic set theoretic principles? And should we not include such principles among the set theoretic reflection principles?[60]

6.6. From epistemic to ontological reflection

In our discussion of aspects of the evolution of the concept of reflection in the history of philosophy (chapter 3), we traced two philosophically important notions of reflection: *epistemic reflection* and *ontological reflection*.

What was called epistemic reflection gave rise, in the twentieth century, to the concept of proof theoretic reflection. In section 3.8 we saw how a *provability predicate* captures introspective powers of a mathematician.

But proof theoretic reflection principles go beyond merely capturing the result of an introspective process. They relate what a mathematician can prove to what is *true*. This means that our epistemic warrant for proof theoretic reflection principles will have to go beyond our warrant for introspection. To the vast majority of mathematicians and philosophers they seem to be a very safe bet. And yet I maintain that until recently the question of our warrant for epistemic reflection principles has not received the philosophical attention that it deserves. We will be much exercised by the question of warrant for proof theoretic reflection in the chapters that follow.

Set theoretic reflection principles are ontological reflection principles. Like proof theoretic reflection principles, they are in excellent mathematical standing. We have philosophical arguments that purport to justify them. These philosophical arguments have their root in theological principles going back to Antiquity, and that is also what is the problem with them. Many set theorists sympathise with the view that the set theoretic universe is far too complicated to pin down

[60]Something similar can be said about principles that postulate downward and upward Löwenheim-Skolem-Tarski numbers for strong logics, which are discussed below (p. 176). They, too, express forms of reflection *within* V.

in human mathematical language. Yet most philosophers and mathematicians no longer share the philosophico-theological background assumptions that underpin detailed philosophical accounts of why this is so. At any rate, as we have seen in section 1.8, the mathematicians' warrant for basic mathematical principles is not of a philosophical nature. Their warrant for such principles is a form of entitlement. Adoption of set theoretic reflection principles is a sound mathematical response to the phenomenon of deep set theoretic incompleteness. Reflection principles are a powerful and flexible instrument for reducing set theoretic incompleteness. Moreover, they reduce this incompleteness in an elegant way, and they contribute greatly to the economical organisation of our set theoretic knowledge.[61]

I have argued that one fundamental difference between set theoretic and proof theoretic reflection is that the latter contains an epistemic component, whereas the former does not. Is there more we can say about the question *how do set theoretic reflection and proof theoretic reflection relate to each other?* This question was raised in the very early days, when *the same* mathematical logicians (Montague, Levy, Kreisel,...) worked both on proof theoretic and on set theoretic reflection [**KL68**, p. 101]:

> The authors [i.e., Kreisel and Levy] cannot agree on whether [Montague's designation of set theoretic reflection principles as *reflection principles*] is merely a pun.

But this question was not pursued further: it seems that most people thought that there is no close relation between the two.

The only exception to the silence in the mathematical community that I am aware of was Reinhardt, who wrote 25 years later [**Rei74**, p. 193, footnote 3]:

> The connection between [proof theoretic reflection principles] and set theoretic reflection principles does not seem to me to be merely verbal. Here one considers what is true, and this is mirrored by what is provable. There is also an element of ostensive reflexivity: one's considerations are turned back upon themselves (e.g., one tries to be conscious of the formal system one is using). It seems, however, that *they always fall back on something less than themselves* [my emphasis]. This element also occurs in set theoretic reflection, but the ostensive reflexivity is more ontological: we reflect on the (mathematical) existence of that which we consider, as we consider mathematical existence.

That ontological reflection principles involve a 'dropping down to something that is smaller' seems plausible, in the light of what we have seen so far. That this is also true for epistemic reflection principles, is not *entirely* obvious. However, there is a sense in which this applies to epistemic reflection principles, also. Given the completeness theorem for first-order logic, the local reflection principle Rfn(S) for a first-order theory S, for instance, is equivalent to the scheme

$$\varphi \to \exists \mathcal{M} : \mathcal{M} \models S \text{ and } \mathcal{M} \models \varphi.$$

Moreover, given the Löwenheim-Skolem theorem, we may take this model $\mathcal{M}$ to be countable. Thus, if S is some set theoretic theory, then Rfn(S) can (somewhat loosely) be interpreted as saying that for every true φ, there is a *small* set in which S can also be taken to hold, possibly by reinterpreting logical vocabulary.

[61] Already Theorem 6.49 bears witness to this.

So we *can interpret* many standard epistemic reflection principles as ontological reflection principles. This should not blind us to the fact that, *viewed as epistemic reflection principles*, they have very different motivations from when they are viewed as ontological reflection principles.

This raises the question whether there are reflection principles that are somehow "intermediate" between standard proof theoretic reflection principles, viewed as ontological reflection principles, and standard set theoretic principles. And there are. By requiring the structure $\mathcal{M}$ to be more like small sub-structures of the intended structure, we obtain more strength. Pakhomov and Walsh consider the *ω-reflection principle* for a given theory S (ω-Rfn(S)):[62]

Axiom 6.63.

$$\varphi \to \exists\ \omega\text{-model } \mathcal{M} : \mathcal{M} \models S \text{ and } \mathcal{M} \models \varphi,$$

where an ω-model is a model in which the natural numbers are interpreted standardly, i.e., as an ω-sequence. They show that ω-Rfn(S) is equivalent to the claim that arbitrary iterations of uniform Π^1_1 reflection along countable ordinals are Π^1_1-sound [**PW21**].

ω-model reflection is a natural strengthening of the proof theoretic reflection principles that we have considered earlier. But since we cannot avail ourselves to a completeness theorem to give ω-Rfn(S) a straightforward epistemic interpretation, we have to interpret it as an ontological rather than as an epistemic reflection principle.

In line with our discussion at the end of Section 6.5, we may regard the statement of the Löwenheim-Skolem Theorem as a *reflection principle*. It expresses a form of reflection not *of* the universe but *in* the universe. Just as in the case of 'classical' reflection principles such as Bernays Reflection, and in the case of set theoretic embedding principles, it is natural to ask if the Löwenheim-Skolem might plausibly and naturally be *strengthened*. This question was addressed already by Magidor in some of his early work, and is the subject of some recent work in large cardinal theory.

Definition 6.64. Let τ be a fixed vocabulary. A *logic* L consists of:

(1) A set, also denoted by L, of "formulas" of L. If $\varphi \in L$, then there is a natural number n_φ, called the length of the sequence of free variables in φ;
(2) A modelling relation

$$\mathfrak{M} \models \varphi(a_0, \ldots, a_{n-1})$$

between models of vocabulary τ, sequences $(a_0, \ldots a_{n-1})$ of elements of the domain of $\mathfrak{M}$ and formulas $\varphi \in L$. It is assumed that this modelling relation satisfies the isomorphism axiom, that is, if $\pi : \mathfrak{M} \cong \mathfrak{M}'$, then $\mathfrak{M} \models \varphi(a_0, \ldots, a_{n-1})$ if and only if $\mathfrak{M}' \models \varphi(\pi(a_0), \ldots, \pi(a_{n-1}))$.

τ is called the *vocabulary* of the logic L.

Definition 6.65 (downward Löwenheim-Skolem-Tarski number). Let L be any logic. Then the Löwenheim-Skolem-Tarski number of L is the smallest cardinal

[62] In an analogous way, one could define β-model reflection, which should give us even more strength.

κ such that if $\mathfrak{M}$ is any L-structure, then there is a substructure $\mathfrak{M}_0$ of $\mathfrak{M}$ of cardinality strictly lower than κ such that $\mathfrak{M}_0 \longrightarrow_L \mathfrak{M}$.

Then the following is a reflection principle, which we may call *Magidor Reflection* (MR):

Axiom 6.66 (MR). Full second-order logic has a downward Löwenheim-Skolem-Tarski number.

It is clear that Axiom 6.66 is related to Bagaria's Structural Reflection Principle (Axiom 6.61). (This is attested further by the Theorem of Stavi that Vopenka's Principle holds if *every* logic has a Löwenheim-Skolem-Tarski number ([**MV11**, Theorem 6]).)

Magidor has shown that the reflection principle MR has considerable large cardinal strength indeed [**Mag71**]:

Theorem 6.67. *MR is true if and only if there exists a supercompact cardinal, in which case the downward Löwenheim-Skolem-Tarski number of second-order logic is the least supercompact cardinal.*

We can also define the notion of an *upward* Löwenheim-Skolem-Tarski number:

Definition 6.68 (upward Löwenheim-Skolem-Tarski number). Let L be any logic. Then the upward Löwenheim-Skolem-Tarski number of L is the smallest cardinal κ such that if $\mathfrak{M}$ is any L-structure of at least size κ such that $\mathfrak{M} \models \varphi$ for some sentence φ, then there are arbitrarily large L-structures $\mathfrak{M}'$ such that $\mathfrak{M}' \models \varphi$ and $\mathfrak{M}$ is a substructure of $\mathfrak{M}'$.

The following has recently been established [**GO24**]:

Theorem 6.69. *Second-order logic has an upward Löwenheim-Skolem-Tarski number if and only if there is an extendible cardinal, in which case it is the least extendible cardinal.*

6.7. Probabilistic reflection

In this section, we are concerned with the question whether there are plausible probabilistic reflection principles for typefree rational subjective probability.

Formal epistemologists have been interested in this question. Nevertheless, it is still largely an uncharted comain: few genuinely formal investigations have been carried out in this area so far. Nonetheless, I believe that this area holds promise. What follows can therefore be considered as n invitation to look deeper into these matters.

6.7.1. Expert principles and van Fraassen's reflection principle. Expert principles express constraints on rational subjective probability functions. They state that a subject's probability function should *defer* to "better" probability functions (*expert functions*) in certain specific ways.[63]

If Pr_1 be a subjective probability function, and Pr_2 is some other probability function, then expert principles are generally of the following form:

$$\text{For all } r \in \mathbb{R}, \text{ and for all } \phi \text{ such that } \mathrm{Pr}_1(\mathrm{Pr}_2(\phi) = r)) \neq 0,$$

$$\mathrm{Pr}_1(\phi \mid \mathrm{Pr}_2(\phi) = r) = r.$$

[63]For a good short introduction to expert principles in formal epistemology, see, [**Spo12**, Chapter 9].

The antecedent of this principle is of course required to ensure that the relevant conditional probability is *defined*. $\Pr_2$ is the probability function that is supplied by an *expert*. This expert can be one's own subjective probability function at some specific future time, but can also be the probability function of someone much more qualified (a human 'expert'), or a probability function that expresses objective *chance* (i.e., 'nature').[64]

David Lewis advocated the expert principle that results from taking $\Pr_2$ to express objective chance: this is Lewis' *principal principle* ([**Lew80**]). Gaifman investigated the expert principle that results from taking $\Pr_2$ to be the degrees of belief of an expert in a subject matter ([**Gai86**]). Van Fraassen advocated the expert principle that results from taking $\Pr_2$ to be the subject's own subjective probability function at some specific future time: this is *van Fraassen's reflection principle* ([**vF84**]). Variants of the latter are the subject matter of this section.

Since subjective probability is a concept of graded belief, it seems that Fraassen's reflection principle should express the result of a form of epistemic reflection. But given Theorem 4.33, we have to distinguish sharply between conditional probability and probability of a conditional. So van Fraassen's reflection principle should certainly not be seen as a converse positive introspection principle. Thus it is not immediately clear as what *kind* of epistemic reflection principle van Fraassen's reflection principle should be classified. Van Fraassen's reflection principle looks like a kind of coherence or *consistency* principle, which is also how it is often presented: the principle says that it should never be the case that coming to believe with probability 1 your future credence in ϕ forces you (by Bayesian updating) to change your current credence in ϕ.

All expert principles, including van Fraassen's reflection principle, are *controversial*. Indeed, van Fraassen himself argues that Ulysses, having good reasons to believe that his cognitive state will be adversely affected by the sirens soon, should not believe all instances of

$$\Pr_1(\phi \mid \Pr_2(\phi) = r) = r,$$

where $\Pr_1$ represents his current credences, and $\Pr_2$ represents his credences when he will be tied to the mast a few hours in the future [**vF95**].

The expert principles that we have discussed so far are *typed*, in the sense that the possibility that $\Pr_1 \neq \Pr_2$ is left open. Van Fraassen and his successors also discussed the *type-free* variant of van Fraassen's reflection principle, which we will call VFR:

Definition 6.70 (VFR).

For all $r \in \mathbb{R}$, and for all ϕ such that $\Pr(\Pr(\phi) = r)) \neq 0$,

$$\Pr(\phi \mid \Pr(\phi) = r) = r.$$

VFR is called the *synchronic version* of van Fraassen's reflection principle.

The principle VFR differs from the diachronic reflection principle in important ways. Firstly, VFR expresses an *internal* coherence constraint. Thus it is more similar to proof theoretic consistency, which we have recognised as *bona fide* reflection principle, than van Fraassen's diachronic reflection principle is. Secondly, VFR is widely regarded as less controversial than van Fraassen's diachronic reflection principle. Roush describes why [**Rou**, Section 2.1]:

[64]Hence $\Pr_2$ does not necessarily a subjective concept of probability.

> While I can sensibly imagine my future self to be epistemically compromised, unworthy of my deference, violating [VFR] would require regarding my current self as epistemically compromised, as having a degree of belief that should be other than it is. This appears to be something that doubt of my own judgment would call for [...]

Another reason why many find VFR appealing is that it has been shown to follow from introspection principles. This, however, seems not to be a convincing argument, since we have seen earlier that many introspection principles are incompatible with a basic theory of finitely additive typefree probability.[65]

The principle VFR expresses that for every sentence φ, there is a part o of the sample space Ω, where φ behaves as it does on the entire sample space Ω, i.e., such that φ is "reflected" in o. This can be seen as follows. Suppose that $\Pr(\varphi) = r$. Suppose also that $\Pr(\Pr(\varphi) = r) = s \neq 1$. Then there will be a proper part $o \subset \Omega$ of some size $s \in \mathbb{R}$ such that $\Pr(\varphi) = r$ holds everywhere in o. Then by VFR, as evaluated in the restriction of the sample space Ω to its part o, we will likewise have $\Pr(\varphi) = r$.

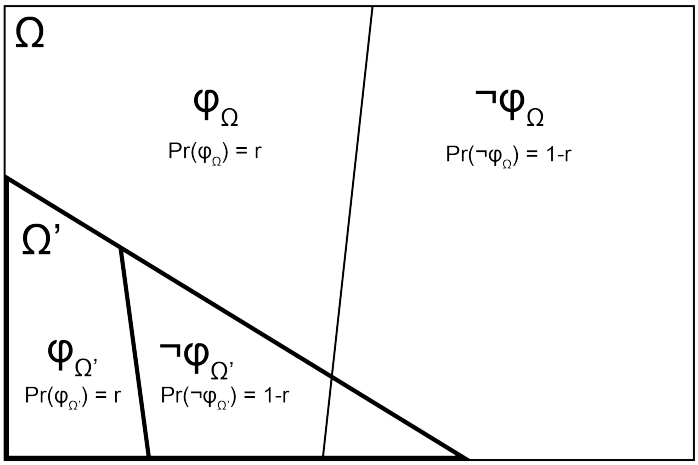

FIGURE 2

Thus, as far as φ is concerned, its behaviour is mirrored (*reflected*) in o. This looks like a probabilistic version of the Montague-Levy phenomenon, except for the requirement that o is "small" as compared to Ω.

6.7.2. An impossibility result. In contrast to van Fraassen's diachronic reflection principle, the principle VFR violates type restrictions: it is essentially type-free. In a type-free context, as we have witnessed on several occasions, one has to be mindful of liar-like arguments. Indeed, a straightforward diagonal argument shows ([**CHL22**, Proposition 6]):

THEOREM 6.71. *The principle VFR cannot consistently be added to the basic theory of typefree subjective probability* K_u^-.

[65] See Proposition 4.39.

So if we are prepared to accept K_u^- as a background theory for type-free subjective probability, then VFR is false.

QUESTION 6.72. Are there natural weakenings of VFR that can be consistently added to K_u^-?

Part III

CHAPTER 7

Epistemic Warrant for Proof-Theoretic Reflection

In this Part, the epistemic views developed in Part I are connected to the logical theories of proof theoretic reflection and truth that were discussed in Part II. We investigate the connections between justification, epistemic entitlement, and philosophical reflection on the one hand, and truth principles and proof theoretic reflection principles on the other hand. Our main question is: *What is our epistemic warrant for principles of truth and proof-theoretic reflection?*

Since mathematical proof is the workhorse of justification in mathematics,[1] it will not be a surprise that there is a connection between proof theoretic reflection and justification. That there should be a connection between proof theoretic reflection and the more touchy-feely notions (such as 'trusting that') that were discussed at the end of Chapter 2, sounds unlikely. Nonetheless, I will make a case for this also.

In the next two Chapters, the concept of truth is set aside. We will not be concerned with principles of truth or with the global reflection principle. Instead, we focus on questions concerning our epistemic warrant for purely mathematical proof theoretic reflection principles in which the concept of truth is not involved. The present chapter is concerned with statements expressing the consistency of a theory, local reflection, and especially *uniform reflection* on a theory, and with how such statements can be *justified*. In the next chapter, the connection between consistency and *epistemic entitlement* is explored. We will see that philosophical thought about these matters is currently still in its infancy.

7.1. Implicit commitment

Gödel's incompleteness theorems show that every consistent mathematical theory S that interprets a modest amount of first-order arithmetic is *incomplete*. There is a specific sense in which Gödel's proof of the incompleteness theorems is constructive. Gödel's proof gives a procedure for *finding* sentences that are independent of S: the Gödel sentence $G(S)$ for S, and an arithmetical sentence $\mathrm{Con}(S)$ canonically expressing, relative to some convenient numerical coding scheme, the consistency of S. Moreover, under fairly general conditions, these sentences do not express *absolutely unsolvable* problems, but only problems that are unsolvable in S.[2] Suppose we start with a mathematical theory S that we justifiedly believe. Then there are

[1] See Chapter 1.

[2] By proving that the Continuum Hypothesis is consistent with ZFC and by conjecturing that it is in fact independent of ZFC, Gödel also contributed to the question whether there are absolutely undecidable mathematical problems. According to quite a few set theorists, the Continuum Hypothesis is an absolutely undecidable mathematical problem.

ways of extending S that can be seen to be correct, such that the extended system decides $G(S)$ and $Con(S)$.[3]

The extended system will in fact *prove* Con(S) [**Göd51**, p. 309]:[4]

> It is [the second incompleteness theorem] which makes the incompletability of mathematics particularly evident. For *it makes it impossible that someone should set up a certain well-defined system of axioms and rules and consistently make the following assertion about it: All of these axioms and rules I perceive (with mathematical certitude) to be correct, and moreover I believe that they contain all of mathematics.* If someone makes such a statement he contradicts himself. For if he perceives the axioms under consideration to be correct, he also perceives (with the same certainty) that they are consistent. Hence he has an insight not derivable from the axioms.

So the theory S + Con is then a correct system that is *less* incomplete than S is. Gödel moreover remarked that the *resources* for reducing the incompleteness of a given believed mathematical system S are in some sense already *contained* in S [**G̈46**, p. 151]:

> It is well known that, in whichever way you make [the concept of demonstrability] precise by means of a formalism, the contemplation of this very formalism gives rise to new axioms which are exactly as evident and justified as those with which you started, and this process of extension can be iterated into the transfinite.

In this way, at the same time as proving that every sufficiently strong mathematical system is incomplete, Gödel found a practical way of *reducing* the incompleteness of every theory that is justifiedly believed.

Turing suggests that reflection principles that are added to a justified mathematical theory are *intuitively* seen to be correct [**Tur39**, p. 198]:

> 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.

However, the further history of the epistemology of proof theoretic reflection has—rightly, in my view—downplayed the role of intuition here.

Myhill may have been the first one to pick up on Gödel's thought that incompleteness can be reduced by making *explicit* what is in some sense implicitly contained in a theory that one already accepts [**Myh60**, p. 462]:

> [I]f a person who has been using certain methods for proving arithmetical theorems succeeds in making these methods explicit, he is *ipso facto* committed to the perfectly definite proposition that the use of those methods cannot lead to a false arithmetical statement, for example the statement that 0 is equal to 1. By Gödel's technique of arithmetization, which translates every

[3] "It can be shown that the undecidable propositions constructed here become decidable whenever appropriate higher types are added [to the system P]" [**Göd31**, p. 181, note 48a].

[4] It will also prove $G(S)$.

> statement of formal deducibility into a statement of arithmetic, any such person is compelled to admit a new arithmetical statement, namely the arithmetized version of the statement that his methods cannot lead to a proof of the statement that 0 is equal to 1. By Gödel's theorem, he could not have established this statement by his previous methods. Hence, as soon as a person makes explicit the tools which he has been using in the construction of arithmetical proofs, he is ipso facto in a position to obtain new arithmetical proofs which he could not have obtained by using those tools alone.

The procedure of adding the consistency of S to an already accepted mathematical theory S can obviously be iterated, as Gödel realised. We have seen in the previous chapter[5] how Turing took first steps in the *mathematical* investigation of transfinite processes of iterating consistency assertions [**Tur39**].

Taking consistency extensions provides *one* way of systematically reducing incompleteness; there are others. Kreisel saw the importance of concentrating on uniform reflection [**Kre60**, p. 289]:

> [N]ow consider finitist proof: if $P(\phi)$ has been recognized by finitist means to be the provability predicate of a (partial) formalization, say S_α, of finitist mathematics, and $P(\phi(n))$ has been established by finitist means then, on the intended meaning of free variables, $\phi(n)$ is finitistically established. In other words, S_α is incomplete and can be extended to S_β, in which $\phi(n)$ is provable.

Already at this stage, Kreisel seems to have had at least an inkling that increasing the strength of the engine (uniform reflection versus forms of local reflection) may not be 'swamped' by the length of the iteration process.

In his landmark paper about transfinite progressions of reflection principles [**Fef62**], Feferman took Kreisel's recommendation from 1958 to heart, by making uniform reflection, instead of consistency, the engine of his transfinite progressions. In the wake of these developments, Kreisel then posed the underlying question in full generality [**Kre70**, p. 489]:

> What principles of proof do we recognize as valid once we have understood (or, as one sometimes says, 'accepted') certain given concepts?

Feferman seems to have seen his work on transfinite progressions as answering Kreisel's question in the situation where the starting theory is Peano Arithmetic. Importantly, he states that our acceptance of a reflection principle for our base theory (and iterating this procedure) rests on our pre-theoretic *attitude* [**Fef62**, p. 261]:

> 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.

This emphasis on the dimension of trust in theory acceptance harmonises with the more philosophical views of acceptance and belief that we discussed in Section 2.7.

[5] See Section 6.2.2.

In the above quotes by Feferman, we observe also the absence of any appeal to mathematical intuition, which, as we saw, is present in Turing's remarks on our epistemic warrant for proof theoretic reflection principles. At the same time, it is clear that Feferman is deeply interested in questions concerning our epistemic warrant for reflection principles. Also in later work, Feferman continues to emphasise this epistemic dimension [**Fef91**, p. 2]:

> 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.* [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, 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'.

At this point, one expects the attention to shift to the question of the justification of $RFN(S)$ given that one has justification for S. But Feferman seems to say that, beyond the justification of S, *no further justification is needed* for $RFN(S)$. In one place, he writes [**Fef88**, p. 131, my emphasis]:

> The idea of an autonomous progression more nearly approximates the process of finding out what is implicit in accepting a basic system L1, i.e., of *what one ought to accept, on the same fundamental grounds,* when one accepts L1.

And along the same lines, in a later article he writes [**Fef96**, Section 2]:

> That idea, in the case of formal systems S in the language of arithmetic comes down [...] to one form or another of (proof-theoretic) reflection principle, that is a formal scheme to the effect that whatever is provable in S is correct. [...] The axioms Rfn_S, and more generally, RFN_S [local and uniform reflection, respectively] may indeed be considered 'exactly as evident and justified' as those with which one started.

In sum, Feferman is proposing what is labeled the *Implicit Commitment Thesis* (ICT):[6]

> *When you are justified in believing a mathematical theory S, then you ought to accept, on the same fundamental grounds, proof theoretic reflection principles for S.*

Here 'ought to accept' should be interpreted as *is rationally required to accept*, and 'on the same fundamental grounds' means *without giving more reasons*, i.e., without doing extra justificatory work.

Now this is odd. By the incompleteness theorems, $S+RFN(S)$ is stronger than S. So it seems that, beyond the justification of S, extra justfication is needed to be justified in also believing $RFN(S)$. In other words, ICT is a bold claim because

[6]This terminology is due to Dean: see [**Dea15**, p. 32].

if S is sufficiently strong (and consistent), then by the incompleteness theorems, proof theoretic reflection principles for S are logically independent of S. Later in this book, we will come back to the important question what exactly is going on here.

The boldness of Feferman's claim is often overlooked in the literature. Reflection principles for a theory S can of course also be *proved* in theories that are stronger than S. The way in which reflection principles can be proved in truth theories in particular, has been the subject of some of Feferman's influential research in this area [**Fef91**, p. 2]:

> [...] which statements in the base language L of S [...] ought to be accepted if one has accepted the basic axioms and rules of S? The answer is given in an ordinary theory Ref(S) formulated in a language $L(T, F)$ [...] where T and F are partial truth and falsity predicates which are self-applicable in the sense that they apply to (codes of) statements of $L(T, F)$ [...] Thus, for example, we may reason in Ref(PA) by induction on the truth of statements which contain the notion of truth, and thus arrive at statements of the form: $\forall x[Prov_{PA}(x) \rightarrow T(x)]$, and by iterating this kind of argument derive iterated reflection principles for arithmetic.

Ketland takes Feferman to argue in this passage that reflection principles need to be justified, and that one particular natural way to justify them is to derive them from truth axioms [**Ket05**, p. 79–80]. But this interpretation does not explain the earlier quoted passages, where Feferman argues that no new principles are necessary to be epistemically warranted in believing reflection principles ("essentially the same grounds"). Moreover, it is clear from Feferman's own comments on the passage just quoted that he does not take it to have fundamental epistemological significance. He merely believes that $KF(S)$ gives a particularly elegant way of expressing the implicit commitment of a theory S [**Fef91**, p. 3]:

> [...] the schematic notion of reflective closure meets among other things the aim to give a more perspicacious generation procedure for predicativity without the use of progressions of theories or *prima facie* impredicative notions such as ordinals or well-orderings.

Later in the book, we will come back to the important question to what extent Feferman's epistemological remarks can be taken at face value.

In what follows, we will suppose that your justification for S need not justify more than S. In particular, suppose that it does not justify the statement expressing a reflection principle for S.[7] That you can be in such a situation can be argued as follows. Suppose we had a solid argument for the thesis that for every recursively axiomatised theory S in the language of arithmetic, for you fully and justifiedly to believe S, you would in addition have to have a justified belief in a reflection principle for S. Then it would follow that for no recursively axiomatised theory S in the language of arithmetic, you could be justified in believing S and no more than that. I.e., then as far as arithmetic is concerned, your powers would outstrip

[7] That such a situation is possible (for a theory such as PRA, for instance), is argued for instance in [**Dea15**].

those of any Turing machine. But—pace Lucas and Penrose[8]—it is widely accepted that currently no such argument exists that carries conviction. On the contrary, the position that humans are at bottom equivalent to Turing machines as far as justified mathematical beliefs are concerned seems widely accepted.

Feferman, in the tradition of Kronecker, Poincaré, and Hermann Weyl, takes Peano Arithmetic to be a mathematically justified starting point. He thus takes every mathematician who thinks likewise, to be *implicitly committed* to successive strengthenings by uniform reflection of this evident starting point [**Fef91**, p. 2, my emphasis]:

> 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 the acceptance of the starting theory.*

In the previous Chapter, we have seen how Feferman then argued that the implicit commitment of PA stretches as far as the *autonomous* progression generated from PA reaches, and how he calculated the length of this iteration (Theorem 6.21). Thus, in Feferman's view, the autonomous closure of PA under uniform reflection is what the mathematician who justifiedly believes PA is committed to. This autonomous closure is much stronger than the starting theory.

The proof of the theorem that reveals that the scope and limitations of the autonomous progression of PA can only be carried out in a mathematical theory that is essentially stronger than the autonomous progression of PA. Thus the limits of what one is implicitly committed to when one is committed to PA, cannot even *implicitly* be recognised from PA. This is a familiar refrain from the foundations of mathematics: the limits of finitism cannot finitistically be recognised, the limits of predicativism cannot predicatively be discerned, and so on.

Whether ICT is correct is a significant question not only as a general epistemological question, but also from the point of view of the foundations of mathematics. So one would expect especially philosophers of mathematics to be interested in it. Until recently, however, this question was almost universally ignored in the philosophy of mathematics.[9] I consider this a collective dereliction of epistemological duty. Since about 2015, however, this situation has changed, as we will see in the remainder of this book.

It is tempting to take Feferman's epistemological claims to be at best no more than loose talk, and at worse philosophical nonsense. Indeed, I believe that privately, some philosophers of mathematics contrast Feferman the brilliant logician and mathematician with Feferman the epistemological simpleton. I do not accept this picture. On the contrary: I believe that Feferman's philosophical remarks are carefully thought through—albeit also very terse. In the next Chapter, I will argue that Feferman is basically right when he claims that reflection principles can be warranted without being justified. I will argue for this thesis by appealing to Burge's and Wright's work on epistemological entitlement. Roughly, I will attempt to show that epistemic warrant for reflection principles can be more a matter of

[8]See [**Luc61**], [**Pen89**], and [**Pen94**].

[9]From the early 1970s onwards until recently, mathematical epistemology was dominated by questions related to the epistemological access that we as spatiotemporal beings have to the world of abstract mathematical objects. The seminal paper here is [**Ben73**].

entitlement than of justification—although the details of the account and the issues involved are complicated.

In this context, it seems to me significant that very early on, there was a recognition of the connection between reflection extensions on the one hand, and concepts lying on the border between epistemology, pragmatics, and philosophy of mind on the other hand. As we have seen earlier, Feferman writes early on [**Fef62**, p. 261, my emphasis]:[10]

> 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* in the system of axioms A_k.

Being primarily interested in the mathematical problems to which this gives rise,[11] Feferman did not pursue this particular connection further. However, it seems to me that this admittedly cryptic passage points to a connection between epistemic warrant for reflection principles and fiducial trust, which was discussed at the end of Chapter 3. In particular, it seems to me that warranted belief in reflection principles can result from exercising our entitlement to reflect on fiducial trust.

7.2. Reflection as basic?

The contemporary debate about our epistemic warrant for reflection principles begins around the year 2000 in truth theory. As we have seen, according to Horwich's minimalist truth theory, correct theories of truth are neutral in substantive debates inside and outside philosophy [**Hor90**]. As a consequence of this, truth theories should then also be *mathematically neutral*, in the sense that truth theory should be proof-theoretically conservative over mathematical background theories [**Hor95**]. Shapiro and Ketland then argued that this consequence of Horwich's minimalism is incompatible with what *they* regard as a desideratum of a good truth theory A (over a background theory S), namely, that S+A proves certain sentences that are indepdentent of S, such as $G(S)$ and $Con(S)$ [**Sha98**], [**Ket99**]. Moreover, they point out that there are natural truth theoretic arguments for proving $G(S)$ and $Con(S)$, but these arguments exceed the power of natural disquotational truth theories. Thus, they conclude, disquotationalist truth theories, and in particular Horwich's minimalist truth theory, are unsatisfactory.

In reaction to Shapiro's and Ketland's arguments, Tennant argued that, if we are "living" in a theory S, truth laws are not *needed* for acquiring a warranted belief in $G(S)$ or $Con(S)$ [**Ten02**]. Instead, he argues, they can be *proved* from modest reflection principles, such as uniform reflection for S restricted to primitive recursive predicates [**Ten02**, p. 573]. This then shifts the problem to the question how such reflection principles are themselves warranted [**Ket05**, p. 85].

Tennant at one place seems to suggest that no warrant beyond the warrant for S is needed: an agent can "express (in [S + uniform reflection for S restricted to primitive recursive predicates]) his willingness, via the soundness principle, to assert any theorem of S" [**Ten02**, p. 574].

Ketland rightly points out that this last statement is not correct [**Ket10**, p. 430]:

---

[10] Some will view this as just more loose talk; I don't.

[11] In particular, Feferman was interested in the question: *how far can we get by iterating reflection principles?* [**Fra04a**, p. 228].

> It should be noted that this is a non-standard claim. Usually, a reflection scheme like $[Rfn_{PA}]$ is said to express the soundness of [PA]: that whatever [PA] proves is true. And being true is not the same as being accepted.

In other words, a person's acceptance of a system S is *expressed* rather by the inference rule

$$\frac{\vdash Bew_S(\varphi)}{\vdash \varphi}.$$

This rule is easily seen to be conservative over S [**Fra04a**, p. 216],[12] and therefore not equivalent to any of the standard reflection principles for S, which are ampliative. In particular, when this rule of inference is added to S, $G(S)$ and $Con(S)$ do not become provable.

In a later article, Tennant repeats his view that reflection principles need no further justification, but he also says that reflection principles are somehow the outcome of a process of reflection [**Ten05**, p. 92]:

> No further justification is needed for the new commitment made by expressing one's earlier commitments. As soon as one appreciates the process of reflection, and how its outcome is expressed by the reflection principle, one already has an explanation of why someone who accepts S should also accept all instances of the reflection principle.

This sounds intriguing, but of course then one wants to know what this process of reflection looks like, and how it leads to the acceptance of a reflection principle. The philosophical story is missing here.

7.3. A sceptical position

In section 7.1, we saw how Feferman claims that if one accepts a basic theory S, then one *ought* to accept reflection principles for S *on the same fundamental grounds* [**Fef88**, p. 131]. Let us put the puzzling "on the same grounds" aside for a moment, and concentrate on the "rational ought" in this claim. Dean is sceptical about this aspect of Feferman's views on implicit commitment [**Dea15**, p. 35]. He gives two arguments for the thesis that one can be perfectly rational, and still refuse to accept reflection principles for a theory S that one unreservedly accepts.

In his *first argument*, he recalls proof-theoretic results that show that, under fairly general circumstances, a reflection principle for an arithmetical theory S is equivalent to a principle of transfinite induction [**Dea15**, section 3].[13] Not only does a proof-theoretic equivalence hold between a typical reflection principle and some principle of transfinite induction, but the two are also *justificatorily equivalent* [**Dea15**, p. 47], presumably in the sense that one is also not *epistemically* prior to the other. If one has justification for S and no more, then warrant for accepting extra transfinite induction requires extra reasons. Why should we believe that, for any theory S, there are always compelling extra reasons for accepting this degree of transfinite induction?

[12] See also Theorem 6.5.

[13] See Theorem 6.10.

Note that, because of the claimed justificatory equivalence between reflection and transfinite induction, proof of transfinite induction from the reflection principle is not an acceptable answer for Dean. However, the claim that justificatory equivalence follows from proof theoretic equivalence at least in the case of reflection and transfinite induction, needs an argument. None is given in [**Dea15**]. That *in general* justificatory equivalence follows from proof theoretic equivalence, seems doubtful. The literature on reverse mathematics is full of examples of pairs of statements that are proof-theoretically equivalent [**Sim11**], where one of them appears to be epistemically more basic than the other.

Dean's *second argument* turns on the concept of *epistemically stable* theories. A mathematical theory S is epistemically stable if there appears to be nothing blameworthy about someone who accepts S and nothing more [**Dea15**, p. 53]. Dean takes PRA and PA to be epistemically stable theories. The reason is that there seem to be coherent rationales for accepting the theory PRA (the theory PA, respectively) and nothing more. In the case of PRA, this rationale is given by Hilbertian finitism ([**Hil26**], [**Tai81**]). In the case of PA, Isaacson has attempted to provide such a rationale [**Isa87**].[14] One might add Ramified Analysis up to level Γ_0 as another such epistemically stable theory: its coherent rationale is given by Feferman's flavour of predicativism [**Fef05**]. Note, incidentally, Dean's argument from epistemic stability requires him to adopt, as I do, a liberal conception in the sense of van Fraassen.[15] After all, on his view, it is rational to be a Hilbertian finitist, but it is also rational to be a Fefermanian predicativist.

In order to evaluate Dean's second argument, we have to distinguish cases.

Let us start with Ramified Analysis up to level Γ_0. For every $\alpha < \Gamma_0$, the theory RA_α is autonomously, i.e., predicatively, justifiable. But *that* this is so, requires an induction up to Γ_0, which is just beyond the reach of predicativism, at least as understood by Feferman. For each $\alpha < \Gamma_0$, the typical reflection principles for RA_α are provable in $RA_{\alpha+1}$, which itself is predicatively justified. In this sense, $\bigcup\{RA_\alpha \mid \alpha < \Gamma_0\}$ is closed under reflection principles: the predicativist can justify reflection principles for every theory that she justifiedly believes. So Dean's objection does not apply.

Next, consider PRA. As with predicativism, the Hilbertian finitist accepts all finite fragments of PRA. But she cannot justify PRA *as a whole*. That task goes beyond what the Hilbertian finitist is capable of. But in this case, there is also a prior difficulty. As we have seen in Chapter 6, PRA is formulated in a quantifier-free language (where variables are permitted as free parameters) [**Sko23**]. The standard provability predicate for PRA is essentially Σ_1, and therefore cannot be expressed in the language of PRA. From a foundational point of view, this is an expression of the fact that the Hilbertian finitist does not recognise such formulas as meaningful. Therefore the Hilbertian finitist does not recognise a typical reflection principle as meaningful. This points to what might be a reasonable restriction on the "rational ought" in Feferman's contention that we should accept reflection principles for the mathematical theories that we justifiedly believe. It might be said that someone who does not recognise that all concepts involved in a reflection principle are in

[14] It is more controversial whether Isaacson succeeded in giving a coherent rationale for accepting PA and no more than whether Hilbert succeeded in providing a coherent rationale for accepting PRA and no more. But be that as it may.

[15] See p. 67.

good standing, is under no rational obligation to accept reflection principles. Note, in this context, that such conceptual doubts might not only be targeted to the notion of provability in a formal system, but also to the concept of truth, which figures prominently in global reflection principles. Some might reply to this that it is just unreasonable not to accept at least provability in a theory as a legitimate concept. But, then again, perhaps such doubters are just really cautious thinkers: why would that be rationally unacceptable?

To conclude, let us look at PA. Concerning the determination of the intended model of arithmetic, Isaacson is a second-orderist. He believes that we fix the intended structure of arithmetic (up to isomorphism) by asserting Dedekind's *second-order* formalisation of Peano arithmetic (PA^2), where the second-order quantifiers range over *all* subsets of the domain of discourse. But when it comes to the question which arithmetical sentences can be recognised to be true merely on the basis of our intuition of the natural numbers (and elementary relations on them), Isaacson holds that this collection is captured *exactly* by *first-order* Peano arithmetic [**Isa87**, p. 166], [**Isa92**, p. 95]. In *this* sense, Dean argues, accepting PA and no more is an epistemically stable position [**Dea15**, section 4.2].

Of course Isaacson does not think that PA contains all true sentences of the language of first-order arithmetic. He recognises, for example, as most of us do, that all first-order arithmetical sentences that follow from PA^2, are true, and that proof-theoretic reflection principles for PA are true. However, he believes that what exceeds PA cannot be known only on the basis of our intuition of the natural numbers (and elementary relations on them): first-order arithmetical statements exceeding PA can only be seen to be true on the basis of *higher-order concepts*. For instance, we can recognise the truth of many sentences that are independent of PA by deriving them from PA^2. But the latter is recognised to be true on the basis not only of our intuition of the natural numbers, but also on the basis of our concept of *set* of natural numbers, which is a higher-order concept in Isaacson's sense. Concerning the particular case of proof-theoretic reflection principles, he writes [**Isa92**, p. 96]:

> In the case of the Gödel sentence for Peano arithmetic, the hidden concepts are provability in the formal system of Peano arithmetic and, most crucially, consistency of Peano arithmetic. That is, to perceive the truth of the Gödel sentence (presented purely in the first-order language of arithmetic) we must understand that it expresses the condition that this sentence is not provable in this given formal system and see that this formal system is consistent.

This indeed points to a question that we have not addressed until now. Isaacson is right that provability (in a formal system) is strictly speaking not an arithmetical concept. Bew_{PA} is merely a complex arithmetical Σ_1 formula, and as such expresses a complex property of natural numbers. We "interpret" it as a provability predicate, which is a predicate of arithmetical *sentences*, on the basis of some convenient Gödel coding scheme. This reading of Bew_{PA} as a provability predicate is presupposed in all accounts of why we ought to believe reflection principles for theories that we are justified in believing. Thus our epistemic warrant for believing such proof-theoretic reflection principles presupposes an epistemic warrant for reading Bew_{PA} in this "syntactic" way. It is then incumbent upon us to get clear about how this

warrant is acquired and on what it is based. This problem will occupy us in the next chapter.[16]

Summing up, our verdict on Dean's objections against Feferman's "ought" is, at least so far, mixed. His objections may be inconclusive, but they do throw up questions. The most important of these have to do with our understanding of certain arithmetical predicates as standard provability predicates. Moreover, there is a deeper problem. Until now we have not seen a *positive* argument for Feferman's "ought".

7.4. Justifying reflection

Most philosophers of mathematics believe that we *can* acquire warrant for proof theoretic reflection principles for theories that we are already warranted to believe. Let us now look at how such warrant can be acquired.

7.4.1. Higher-order concepts. There are ways of justifying proof theoretic reflection principles. Suppose, for instance, that you are a Hilbertian finitist who accepts PRA and nothing more. At some point, you come to adopt mathematical induction principles for good reasons. Then you can, for instance, prove the consistency of PRA.

Isaacson has argued that Peano Arithmetic is somehow "arithmetically complete" [**Isa87**]. Of course he does not mean to deny that Gödel's incompleteness theorems hold for PA. Rather, by this statement he means that all ways of proving arithmetical statements that exceed PA, require the use of *higher-order concepts*, i.e., concepts that go beyond the first-order logical concepts and the concept of being a natural number. In particular, someone who has a mathematically justified belief in PA and wants to prove proof-theoretic reflection principles for PA, needs to appeal to principles regarding concepts that exceed the language of (first-order) PA.[17] For instance, she may acquire the concept of *set of* natural numbers, and good reasons for accepting principles governing this new concept (principles of induction, principles of comprehension). Then she can prove proof theoretic reflection principles for PA. Alternatively, she may use principles of truth, and prove proof theoretic reflection principles in CT.[18] In such situations, the question of course arises how the new axioms involving the new concepts are warranted.

Whether Isaacson is right, is a moot question. Moreover, a further question is whether, if Isaacson is right concerning PA, a similar phenomenon might occur for ZFC.[19] Of course Con(ZFC) can be derived from large cardinal axioms. But large cardinal axioms are not *generally* taken by the mathematical community to be mathematically warranted. Moreover, as we have seen,[20] some large cardinal axioms that are expressible as a single axiom do not entail even the schematic principle Rfn(ZFC).

Adopting a proof theoretic reflection principle as a new fundamental axiom is not really an option. They do not satisfy the conditions for being a fundamental axiom. They do not sufficiently organise and unify mathematical knowledge. They do not "shed new light" on the natural number structure. Moreover, and just as

[16] See Section 8.3.3.

[17] Isaacson argues for this thesis in [**Isa92**].

[18] These strategies will be discussed in detail in Chapter 9.

[19] This question is discussed in [**Hor01a**].

[20] See Theorem 6.7.

importantly, they do not have the generality that is mostly expected of fundamental axioms, for they refer, via coding, to particular formal systems.

Even if proof theoretic reflection principles can never be fundamental axioms, perhaps epistemic warrant for them can still be obtained in a *direct way*, i.e., without using new concepts. The strategy is to argue, based on a fine-grained epistemic analysis, that the epistemic distance between the background theory (PRA, PA, ZFC,...) and reflection principles for them is somehow "very small". Two such attempts have recently been made: one by Martin Fischer [**Fis23**], and one by Mateusz Łełyk and Carlo Nicolai [**LN22**]. We will see that these proposals, while certainly not identical, are nonetheless deeply related to each other; we discuss each of them in turn.

7.4.2. The uniform reflection rule and informal arithmetic. Fischer's aim in [**Fis23**] is to show that we are warranted to believe $URF(S)$ for a mathematical theory S that we believe.

Let us restrict ourselves to arithmetic. Fischer distinguishes between the formal theories of arithmetic that we endorse on the one hand, and *informal arithmetic* on the other hand [**Fis23**, Section 4]. For definiteness, on the formal side, we will suppose that PA is endorsed. On the other hand, it is much more difficult to give a clear and informative account of informal mathematical—in our case arithmetical—knowledge.[21] Fischer writes [**Fis23**, Section 1.3]:

> We will call [implicit] mathematical knowledge [...] implicit, informal and intuitive. Implicit in the sense that we do not require awareness of this knowledge. Informal, because we think of it as not necessarily recursive in the relevant sense, but rather semi-formal. Intuitive only in so far as we think that the knowledge stems from a concept of natural number...

Fischer takes the implicit commitments of arithmetic to be captured by informal arithmetic; he wants to show that implicit (or informal) arithmetic is closed under *uniform reflection*.

There is something puzzling in talking about implicit *knowledge* in the way that Fischer does. On the one hand, knowledge entails belief. On the other hand, on most conceptions of belief, having a belief that p (and therefore also knowing that p) requires at least being *close to* being aware that p. Some say, for instance, that a good test of whether you believe that p is whether you would, when asked whether p, answer 'yes' after carefully considering the question (assuming that you are sincere and willing to answer the question on the occasion). On some views,[22] you may believe that p while not even be close to being aware of your believing that p ("*I am not in love with X!*"). However this may be, more detail about the conception of belief that is at play here, would not be amiss. However, I will sidestep this question here.

Fischer first considers the local reflection *rule* RfR(PA):[23]

$$\frac{\vdash Bew_{PA}(\varphi)}{\vdash \varphi}.$$

[21] See Section 1.5.

[22] See for instance [**Rad66**].

[23] See p. 154.

We have seen that RfR(PA) is an admissible inference rule of PA: adding it does not result in new theorems.[24] However, this rule is not *derivable* in PA. It can (trivially) be derived if, in addition to PA, we assume that PA is Σ_1-sound, i.e., the scheme

$$(\mathrm{Bew}_{PA}(\varphi) \wedge \text{“}\varphi \text{ is } \Sigma_1\text{”}) \to \varphi.$$

Now Fischer contends that it is reasonable to assume that the Σ_1-soundness of PA is part of our *informal* arithmetical knowledge [**Fis23**, Section 3.2]. So, “informally”, the rule RfR(PA) *is* derivable. Something like this cannot, at this stage, for the axiom scheme RFN(PA), for we have as yet no argument that it, too, belongs to informal arithmetic.

In the next stage of his argument, Fischer makes use of the informal derivability of RfR(PA). Moreover, he takes the outcome of the first stage of his investigations to suggest having a close look at the rule-form $RFR(PA)$ of the axiom scheme of uniform reflection:[25]

$$\frac{PA \vdash \mathrm{Proof}_{PA}(\varphi(x), f(x)) \qquad (\ f \text{ a primitive recursive function }\ )}{\vdash \forall x \varphi(x)}.$$

We have seen that, over PA, the rule RFR(PA) is equivalent to RFN(PA).[26] This means that the rule form of RFN(PA) behaves very different from the rule form of Rfn(PA): the former proves new theorems, when added to PA; the latter does not.

Fischer then sets out to argue that *informal* arithmetic is closed under the rule RFR(PA). The rough argument, which Fischer admits is not completely convincing as it stands, goes as follows [**Fis23**, Section 3.1]:

> If we look at the premise of the uniform reflection rule, we see that from $\vdash \mathrm{Bew}_{PA}(A(x))$ we can derive all the ‘standard’ instances $\vdash \mathrm{Bew}_{PA}(A(\overline{n}))$ for numerals $\overline{n}$. Applying the local reflection rule would give us $\vdash A(\overline{n})$ for all $n \in \mathbb{N}$. Assuming an understanding that the numerals exhaust all the natural numbers we could argue that $\forall x A(x)$ has to be correct.

Somewhat less sketchily, the argument goes as follows [**Fis23**, section 4.2]:

> [We assume that the premise of RFR(PA) holds, and] would like to have an inductive argument for the correctness of $\forall x(N(x) \to A(x))$. Let us try to argue informally in a metatheory and assume that $f(z)$ [where z is intended to denote the number 0] and $f(Sc(x))$ by $h(f(x))$ [where Sc stands for the successor function]:
>
> 1. $\vdash \mathrm{Proof}_{PA}(A(x), f(x))$
> 2. $\vdash \mathrm{Proof}_{PA}(A(\overline{z}), g(z))$
> 3. $\vdash A(z)$
> 4. $\vdash \mathrm{Proof}_{PA}(A(\overline{Scz}), h(g(z)))$
> 5. $\vdash A(Sc(z))$
>
> ...
>
> 6. $\vdash \mathrm{Proof}_{PA}(A(\overline{ScSc...z}), h(f(Sc...z)))$
> 7. $\vdash A(ScSc...z)$

[24] See Lemma 6.5.

[25] See p. 154.

[26] See Theorem 6.6.

> By 'reflecting' on the argument it appears that we can turn it into an informal inductive argument. The primitive recursive function allows us to argue inductively that $A(n) \to A(Sn)$. In contrast to a metatheoretic argument that establishes $A(n)$ for all $n \in \mathbb{N}$, we do not have to assume a specific understanding of $\mathbb{N}$, but the inductive structure allows us to argue for all structures of natural numbers that agree on the inductive build-up.

This argument is infinitary, and it is not immediately clear how the conclusion $\forall x A(x)$ is derived from 1.–7. Certainly an appeal to the ω-rule is not what is intended at this point. So the question arises: what does 'reflecting on this argument' mean?

In this context, Fischer takes the reflection to be a (finitary) argument in informal arithmetic, involving the following laws of truth [**Fis23**, section 5]:

(1) **T-In** $B(x) \to T(B(x))$, for B a Δ_0 arithmetical formula;

(2) **U-Inf** $\forall x T A(x) \to T \forall x A(x)$, for A an arithmetical formula;

(3) **Conec**

$$\frac{\vdash T(A)}{\vdash A},$$

for A an arithmetical sentence.

(In these principles T is a primitive truth predicate.) In other words, the ultimate warrant for RFR(PA) in informal arithmetic can be formalised as a derivation in a truth theory over PA.

At this stage, one may ask whether, since truth is a philosophical notion, this would not make the resulting warrant for URR(PA), and thus also for URF(PA), partly philosophical in nature. Indeed, the question arises how Fischer's account is at bottom different from views such as that of Shapiro and Ketland, who take reflection principles such as URF(PA) to be justified by deriving them in CT.[27]

Fischer anticipates this worry and replies that the truth principles that he uses are much weaker than CT. In contrast to the full compositional notion of truth that is captured by CT, his truth predicate plays a purely 'instrumental' or 'expressive' role [**Fis23**, section 5.1].

It is not completely clear what is meant by 'purely instrumental' or 'purely expressive' here. Moreover, vagueness aside, it is not clear to me that this can also be said for the principle U-Inf.[28] Somehow, like URR(PA), the principle U-Inf appears to be a weakening of the ω-rule, which itself is of course of formidable strength. Moreover, all known non-conservative truth theories contain some form of U-Inf. At any rate, at this point the question arises what our warrant for the truth principles that Fischer uses consists in.

7.4.3. Axiomatising implicit commitment. We have seen how Fischer's aim was to show how from a warrant for believing a mathematical theory T, a warrant for believing $T + RFN(T)$ can be obtained. In their article [**LN22**], Łełyk and Nicolai embark on a different project. Their aim is also to describe the implicit

[27] Such positions will be discussed in more detail later: see Chapter 7.

[28] Fischer himself admits that this is the hardest case.

commitment of arbitrary mathematical theories T. And on their view, it will also turn out to be the case that $RFN(T)$ is part of the implicit commitment of T. But they do not explain exactly what one's warrant for $RFN(T)$ consists in, when one has warranted belief in T. Rather, they aim to give a (partial) characterisation of the implicit commitment of T by isolating operations under which the implicit commitment of T is closed. They consider the operator $\mathfrak{I}$, that for a given mathematical theory T, yields the implicit commitment $\mathfrak{I}(T)$ of that theory. That is, when one is warranted to believe T, the sentences in $\mathfrak{I}(T)$ constitute what one is thereby implicitly committed also to believe.

A mathematical theory is taken to be given by a *decidable* predicate that isolates the axioms of the theory [**LN22**, section 3.1]. (So also the theory PA will be taken to be given by a Δ_0 predicate.)

Łełyk and Nicolai contend that for every mathematical theory T, its implicit commitment $\mathfrak{I}(T)$ is closed under at least two operations [**LN22**, section 3.2]:

(1) **Invariance** For every mathematical theory T', if it can be simply and uniformly recognised that for every φ, a proof of φ in T' can be transformed into a proof of φ in T, then $\mathfrak{I}(T') \subseteq \mathfrak{I}(T)$.
(2) **Reflection** If it can be simply and uniformly recognised that for all n, the formula $\varphi(\overline{n})$ is an axiom of T,[29] then $\forall x \varphi(x) \in \mathfrak{I}(T)$.

Here a 'simple and uniform recognition' is identified with a proof in EA, where we recall that Elementary Arithmetic is one of the weakest subsystems of PA.[30]

Invariance is so-called because it expresses that fine details about how a theory is axiomatised do not matter for implicit commitment. *Reflection* expresses the idea that you are implicitly committed to believing the universal generalisation of a predicate each instance of which you uniformly recognise to be one of your axioms.

By means of a variation on Feferman's clever proof that the uniform reflection rule is equivalent to the uniform reflection principle,[31] Łełyk and Nicolai show [**LN22**, Section 3.3, Proposition 1]:

THEOREM 7.1. *For every theory T extending EA:*

$$RFN(T) \subseteq \mathfrak{I}(T).$$

PROOF. Take an arbitrary theory T, represented by a decidable predicate τ. Given an arbitrary $\varphi(v)$, we first show that $EA \vdash \forall x \mathrm{Bew}_\tau(\theta(x))$, where $\theta(x)$ (the so-called "small reflection principle" for τ and φ) is defined as

$$\theta(x) =: \forall y_1, \forall y_2 (y_1 = (x)_1 \wedge y_2 = (x)_2 \wedge \mathrm{Proof}_\tau(y_1, \varphi(y_2) \to \varphi(y_2)),$$

where we assume a simple pairing function, and $(x)_1$ denotes the first element of the ordered pair coded by x (similarly for $(x)_2$).

Working in EA, we fix an arbitrary x and let $y_1 = (x)_1$ and $y_2 = (x)_2$. If $\mathrm{Proof}_\tau(y_1, \varphi(y_2)$, then $Bew_\tau(\varphi(y_2))$ and the claim follows by logical reasoning inside the provability predicate for τ. Similarly, if $\neg\mathrm{Proof}_\tau(y_1, \varphi(y_2)$, then provable Σ_1-completeness entails that

$$\mathrm{Bew}_\tau(\exists y_1, y_2 (y_1 = (x)_1 \wedge y_2 = (x)_2 \wedge \neg\mathrm{Proof}_\tau(y_1, \varphi(y_2))).$$

Therefore, in either case the claim follows.

[29] So if T is given by the Δ_0 predicate τ, this means: "$\tau(\varphi(\overline{n}))$ holds".
[30] See Section 4.1.1.
[31] See Theorem 6.6.

Now, for θ as above, we define

$$\tau'(x) =: EA(x) \vee \exists y \leq x(x = \theta(y)).$$

Then the previous argument shows that τ' is elementarily reducible to τ. So by Invariance, $\mathfrak{I}(\tau') \subseteq \mathfrak{I}(\tau)$. However, by the definition of τ', we obtain $EA \vdash \forall x \tau'(\theta(x))$. Then, by Reflection, we have $\forall x \theta(x) \in \mathfrak{I}(\tau')$. Hence, by the definition of θ and the fact that $EA \subseteq \tau' \subseteq \mathfrak{I}(\tau')$, we can conclude that

$$\mathfrak{I}(\tau') \vdash \forall x \forall y(\mathrm{Proof}_\tau(x, \varphi(y) \to \varphi(y)),$$

which entails the uniform reflection axiom for φ. Therefore we obtain by our earlier observation that $\mathfrak{I}(\tau') \subseteq \mathfrak{I}(\tau)$, that $RFN(T) \subseteq \mathfrak{I}(T)$.

□

Łełyk and Nicolai emphasise that on their view, implicit commitment is a one-shot affair. Unlike Feferman's notion of implicit commitment, it does not involve an iteration. This is because implicit commitment is an 'operation' on an explicitly believed theory T, the axioms of which are given by a decidable predicate: the collection $\mathfrak{I}(T)$ is not given in this way [**LN22**, section 4].

Individually, Invariance and Reflection are in some sense conservative. That Invariance on its own is conservative, is trivial. But also Reflection over T is conservative *in some sense*: it can be conservatively interpreted in T [**LN22**, section 4]. So it is the interplay of Invariance and Reflection that results in 'real' non-conservativeness.

So, in this way, Łełyk and Nicolai split Uniform Reflection into two components: Invariance and Reflection. They then argue that "Invariance and Reflection preserve justified belief" [**LN22**, section 4]. If this is indeed so, then the proof of theorem 7.1 gives a "deductive route to the justification of Uniform Reflection that is based on more fundamental principles" [**LN22**, section 4].

The structure of this deductive justification is then as follows [**LN22**, section 4]:

> [...] in our reflection principle we start with the justified belief in T and in $\forall x \tau(\varphi(x))$ [where τ is a Δ_0 predicate defining T]. Specifically, the justification of $\forall x \tau(\varphi(x))$ is given by one's justified belief in our formal syntax theory EA, and it is a basic assumption of our framework that such justification is compatible with any of the particular justifications one might have for different choices of τ. The deductive lightness of Reflection [...] enables us to justifiedly believe $\forall x \varphi$.
>
> Consequently, given one's justification for τ, all reasoning steps in [the proof of theorem 7.1] can be seen to preserve such a justification. Closure of justified belief under logical context in our abstract and mathematical context then entails that Uniform Reflection is also justified on the basis of τ.

That Invariance preserves justified belief is trivial. That Reflection also preserves justified belief is less clear, as Łełyk and Nicolai realise. So what exactly is meant by the "deductive lightness" of Reflection?

The reflection operation on T can be seen as a weak version of a uniform reflection rule for T, as a uniform reflection rule for T for *decidable* predicates. The result of extending T with this reflection rule, is interpretable in T. But this does

not mean that this reflection rule is an admissible rule of T. Indeed, if it were, then the proof of Theorem 7.1 would yield a proof of $RFN(T)$ in T.

So the *warrant* for the Reflection step in the proof of theorem 7.1 is not underwritten by a derivation in T. I contend that Łełyk and Nicolai have not made it completely clear wherein this warrant consists. Moreover, since it is an instance of a (weak) uniform reflection rule, the inference is not self-evident. Thus I conclude that Łełyk and Nicolai have not succeeded in making the apparent epistemological gap between T and $RFN(T)$ disappear.

7.5. The leaching problem reconsidered

In a way, things could not be otherwise. $RFN(T)$ does not logically follow from T. So if we *deductively* want to extend a minimal warrant for T to a warrant for $RFN(T)$, as both Fischer and Łełyk and Nicolai aim to do, then non-self-evident extra principles are needed. The warrants for these extra principles will go beyond the minimal warrant for belief in the original theory T. One may attempt to spell out these new warrants in the form of a further argument, as Fischer does. But then the premises of this further argument will not all be self-evident, and require further justification. Alternatively, one may refrain from spelling out these new warrants in detail, as Łełyk and Nicolai do. But then the justificatory obligations have not been fully discharged.

In sum, a regress problem threatens, and we seem to be back to the leaching problem that was described in Section 1.8. It was argued in that Section that mathematicians are warranted in their belief in their basic principles (*Axioms*) by their practical responsiveness to the epistemic virtues of these principles (unifying power, deductive strength, ...). In other words, on the view that I suggest, mathematical axioms are warranted by *entitlements* rather than by justifications. I suggest that they are entitlements of cognitive project in something like Wright's sense of the word.[32] Indeed, inquiry into the mathematical world is just as fundamental a cognitive project as inquiry into the external material world, and any attempt to justify basic mathematical axioms would involve premises that are of no better prior standing.

This seems to me an attractive view. For one thing, it makes our warrant for basic mathematical principles fully mathematical, whereas justificatory accounts of our warrant for basic mathematical principles make our warrant for basic mathematical axioms at least in part a philosophical affair. For another, the view that I suggest avoids talk about mathematical intuition (such as the iterative conception of sets, for instance), which is perhaps ultimately hopelessly vague and irreducibly metaphorical.

Now this may (or may not) be all right as far as mathematical axioms go. But the epistemological status of proof theoretic reflection principles is unlike that of mathematical axioms. Proof theoretical reflection principles are not axioms in the true sense of the word. They are metatheoretic in nature: they are about provability in particular theories. They therefore even do not express basic properties even of the class of natural numbers. More importantly, despite their independence, they do not play a fundamental role in the organisation of our mathematical knowledge. We are therefore not warranted in our belief in them by our practical responsiveness to their epistemic virtues.

[32] See Section 2.5.

Direct non-justificatory warrant for proof theoretic reflection principles is therefore not on the cards. This seems to drive us back to the view that proof theoretic reflection principles must be derived from more basic principles, which must themselves be mathematically warranted in a different way. If that is true, then it is hard to see how we can be warranted in believing proof theoretic reflection principles for our most encompassing mathematical theory. In other words, at least for our most encompassing mathematical theories, it is then hard to see what remains of Feferman's idea of implicit commitment.

I do not have a general solution to this problem. But I will in the next Chapter suggest a way out at least for the weakest of the proof theoretic reflection principles: consistency and local reflection. In particular, I will attempt to show how we can be warranted to believe in the consistency of our most encompassing mathematical theory without deriving the consistency statement from more fundamental principles. So, at least, for the warrant of our belief in consistency, the leaching problem can be avoided. Or so I will presently argue.

CHAPTER 8

Reflecting on Consistency

> Much justified mathematical belief is underwritten by non-demonstrative reasoning [...] Our belief in the consistency of arithmetic seems thoroughly warranted; in fact I think it constitutes knowledge. But no proof of it adds significantly to the ground for our belief.
>
> [**Bur98a**, p. 315]

In this Chapter, an account is given of our warrant for the simplest kind of proof theoretic reflection principles, namely consistency. On the proposed view, our warrant for believing in the consistency of a mathematical theory S that we are already warranted to believe, *need not* be acquired by deriving the consistency of S from a stronger mathematical theory. Instead, the warrant can be obtained by a *process of type **6** reflection* that involves both entitlement and justificatory elements. The main aim of the present Chapter is to describe this reflection process in some detail. I will also explain how the proposed account can perhaps be generalised to certain stronger proof theoretic reflection principles, such as the local reflection principle.

8.1. The phenomenology of mathematical reflection

As mentioned earlier,[1] the idea of implicit commitment seems to trace back to Kreisel's article [**Kre60**]. In later articles, Kreisel was more specific about what he had in mind than in his earlier work. He argued that our warrant for proof theoretic reflection principles derives from a process of reflection [**Kre70**, p. 489]:

> The *particular* kinds of reasoning considered in the present lecture can be roughly described as follows:
> *What principles of proof do we recognize as valid once we have understood (or, as one sometimes says, 'accepted') certain given concepts?*
> The process of recognizing the validity of such principles [...] is here conceived as a *process of reflection*; reflecting on the given concepts, reflecting on this process of reflection, and so forth. It is not assumed that every significant area of mathematics is properly analyzed in this manner; not even all those areas which may be described as: what is implicit in given concepts. For instance, if the basic concepts involve a very high degree of *self reflection*.

[1] See Section 7.1.

He emphasised that in order to make progress on the kind of epistemological questions with which we are concerned, a *phenomenological description* of the reflection process is needed [**Kre68**, p. 362], even though, like Feferman, he himself did not provide one.[2] In this Chapter, I will give a detailed phenomenological description of the reflection process that can provide a warrant for belief in the consistency of a theory that one already is warranted to believe. But it will be a phenomenological description in a looser sense than the technical way in which Husserl and his followers use the term. This process can yield a warrant for belief in *consistency*, and somewhat stronger principles such as local reflection—or so I will argue.

Reflecting on consistency is a *process* in time, consisting of three stages. The reflective process contains two *reflective acts*, but it also contains argumentative (i.e., justificatory) components.

In the process of reflection on consistency, to which we will turn shortly, the reflective mathematician will not make use of non-mathematical concepts (such as the concept of truth, for instance). Moreover, in the reflection process, the mathematician will not derive the consistency of the target theory from some stronger mathematical theory that she comes to believe. For this reason, there will be no leaching problem for the account that will be presented.

8.2. Innocence

I will presently tell a fictional tale about how you can acquire warrant for belief in the consistency of a mathematical theory that you already warrantedly believe without *deriving* the consistency statement. In my story, I draw upon concepts and distinctions that have been explicated in previous chapters.

In this section, the starting point of the reflective process is described. In the next section, we will see how the reflective process takes you from the starting point to warranted belief in the consistency of the mathematical theory you started out believing.

8.2.1. A cognitive project. I will be talking about *you* all the time in the fictional tale that I am about to tell. But nothing hinges on this. *You* might as well be a whole mathematical community for the purposes of the argument that follows.

Suppose you are a mathematician. As a mathematician, you accept and believe the axioms of Peano Arithmetic (PA). You do not accept them instrumentally or provisionally; you accept them *unconditionally*, without any reservations. Moreover, you unreservedly rely on the inference rules of classical logic when you construct *proofs* in your mathematical theory. You fully believe the *theorems* that you prove in PA.[3] This, as far as your mathematical work goes, is *all* that you unconditionally believe and accept. In this situation, you are disposed unconditionally to believe all of (the classical closure of) PA and nothing more.

In particular, the consistency of PA is not something you currently believe or are currently disposed to believing. Suppose that this disposition to believe, as far as mathematics is concerned, all of PA and nothing more, has somehow come to

[2]At least, I do not think that [**Kre68**, p. 362] can be regarded as a phenomenological description of the relevant reflection process(es).

[3]So I will from now on often identify PA with the closure of its axioms under classical logic.

be hard-wired in you. You are guided by an *algorithm* that produces all and only PA-provable statements.

As a mathematician, you have an even deeper commitment to classical logic than to PA. If you were to derive a contradiction in PA, then you would reject some mathematical principles of PA rather than principles of classical logic.

Suppose that PA is in fact *true*. Moreover, assume in addition that you as a matter of fact have epistemic *justification* for your belief in the axioms of PA.[4] You may or may not know that you have, but you have. And suppose that your justification for PA does not justify *more* than PA. In particular, suppose that it does not justify the statement expressing the consistency of PA. (Otherwise our task would be too easy.)

That such a situation is possible (for a theory such as as Primitive Recursive Arithmetic, for instance) is argued for instance in [**Dea15**]. Indeed, we have earlier seen an argument for this.[5] Suppose we had a solid argument for the thesis that for every recursively axiomatised theory T in the language of arithmetic, for you fully and justifiedly to believe T, you would in addition have to have a justified belief in the formalised consistency statement for T. Then it would follow that for *no* recursively axiomatised theory T in the language of arithmetic, you could be justified in believing T and no more than that. I.e., then as far as arithmetic is concerned, your powers would outstrip those of any Turing machine. But it is widely accepted that currently no such argument exists that carries conviction.

Insisting on restrictions on the *kind* of justification for the mathematical axioms of PA would limit the scope of the philosophical account of reflection that I am developing here, for there is no agreement about what justifies mathematical axioms that we think we know. So I impose no restrictions on the *kind* of justification that you have for the axioms of PA. Nonetheless, here is one example of a scenario that has been entertained (and criticised!) in the philosophical literature. The natural number structure is somehow given to you in intuition. You have justified the axioms of PA, and only them, by verifying that they hold in this "standard model"; you believe the axioms of PA on this basis.

The mathematical theory PA is then a fairly large scale *cognitive project* in Wright's sense of the word.[6] It can be seen as an ordered pair:

$$\langle \text{questions expressible in the language of PA}, \text{proofs and refutations in PA} \rangle.$$

Nothing hinges on the maximal mathematical theory that you unreservedly accept and believe being PA; focussing on PA is mainly done for definiteness. The analysis of the process of reflection on consistency that I am about to propose is intended to have some generality: it is intended to apply to a variety of mathematical theories.

8.2.2. The state of innocence. The consistency of PA is a *cornerstone* of your cognitive project. Wright would say that your cognitive project *presupposes* the consistency of PA.[7]

---

[4]Alternatively, we might suppose that you have some form of non-justificatory epistemic warrant for your belief in PA.

[5]See p. 187.

[6]See Section 2.5.

[7]Earlier I have expressed uneasiness with Wright's use of the term 'presupposition' in contexts such as these: see Section 2.6.

You *trust* PA, in the sense that was discussed in Section 2.7. This implies that you rely on PA's consistency even if you have never posed the question of the consistency of PA to yourself. So the formalised consistency statement for PA captures an aspect of your trust in PA [**Fef62**, p. 261].

Perhaps you deeply distrust philosophy, all distinctively philosophical concepts and philosophical theories about them. In particular, you may not believe that there is a concept of *truth* that you may legitimately use in your reasoning. Nevertheless, if you were to discover that PA is inconsistent, then *as a mathematician* you would (rightly) feel compelled to revise your mathematical commitments.

According to Davidson's theory of meaning, such a state of innocence is impossible. In his view, the *meaning* of an object language sentence s is given (under certain circumstances that we need not go into) by the Tarski-biconditional for s in a metalanguage.[8] This means that you *know* the meaning of s if you know the Tarski-biconditional for s. But that means that in order for you to know the meaning of *any* sentence, you must have a truth predicate in your language, as well as a biconditional. In particular, then, in order for you to know the meaning of a sentence of $\mathcal{L}_{\mathrm{PA}}$, you must have a predicate expressing arithmetical truth, which is not a purely arithmetical notion. But this just shows that Davidson's view is implausible, for a very young child understands very simple sentences ("*Ball gone!*") without having a truth predicate or a biconditional in her language. Indeed, I see no cogent reason for rejecting your state of innocence, as I have described it, as fundamentally incoherent.

The situation you are in satisfies Wright's conditions for entitlement of cognitive project:[9] you have no reason to think that PA is inconsistent, and an attempt to justify the consistency of PA would involve presuppositions in turn of no more secure prior standing. So you are *epistemically entitled to rely* on the consistency of PA. In other words, I am granting that your reliance on the consistency of PA is warranted by an *entitlement of cognitive project* in Wright's sense. Call the situation that you are in at this point, i.e., before you start to reflect on your belief of PA, the *state of innocence*.

The next question is: how can you come to be entitled *to believe* in the consistency of PA? I will presently argue that you can come to be in this position by *reflecting* on what you are relying on in your cognitive project. Indeed, there are circumstances in which you can, by reflection, come to *know* the consistency of PA without justifying a statement that expresses the consistency of PA.

8.3. From PA to the consistency of PA

In response to Kreisel's challenge to provide a phenomenological description of mathematical reflection, I now give a depiction of one specific three-stage mathematical reflection process. This is done in the form of a fictional story, but I claim that this is one way in which you *can* come to be warranted in believing in the consistency of PA.

8.3.1. Belief de se. We have seen that in your pursuit of your cognitive project, you are guided by an algorithm e that produces all and only PA-provable

[8] See [**Dav67**].

[9] See p. 55.

statements. Against this background, **stage 1** of the reflective process goes as follows.

We may assume that presently you do not know, or even believe, that you are guided by this algorithm e. But by *reflection* on your cognitive situation, you can obtain beliefs about your cognitive predicament. You can come to believe that in your mathematical work, you are disposed to believing what is provable in PA. This reflective moment constitutes the first stage of the reflective process.

How does this happen? You consider all Peano axioms except the mathematical induction scheme, and realise that you believe them (*introspection*). Concerning the scheme of mathematical induction, you realise that your acceptance of instances of mathematical induction does not depend on the particular formula for which it is instantiated, but that you are disposed to accept all arithmetical statements that have the *form* of a mathematical induction axiom. Similarly for the logical axiom schemes, and the logical schematic rules. Then, by mathematical induction (in a language that extends the language of arithmetic), you conclude that you are disposed to believing all proofs in PA.

Observe that this does not mean that you have thus come to believe that you are the algorithm e that was mentioned at the beginning of this subsection. You have come to believe that what you are disposed to believe is a *subset* of the arithmetical statements that are produced by e. I leave the question whether, and, if so, how, you can come to believe that you are, as far as your mathematical work goes, the Turing machine e *and no more*, for later.[10]

Note also that something fundamentally new has happened in **stage 1** of the reflective process. Up until just now, self-awareness was not involved in the tale. You were as a matter of fact explicitly accepting all of PA, but you did not know this.[11] Now, however, you do, and this involves acts of self-consciousness. This shows that the kind of reflection involved is somewhat similar to the examples of reflection that the early modern philosophers—in particular the rationalists—were occupied with, namely type **6** reflection.[12]

8.3.2. Expressing your trust. In the next phase, **stage 2** of the process, you come to see that you have been, and are, relying on the consistency of your cognitive project. You make your implicit trust explicit. How? Not by 'rational intuition', presumably, but rather by *counterfactual reasoning*.

You have recognised that, as far as mathematics is concerned, you are disposed to believing what is provable in PA (stage 1 of the reflective process). You now realise that *if you were to derive a contradiction in* PA, your commitment to your cognitive project would collapse.

This counterfactual belief is acquired through a philosophical *what if*-consideration, i.e., through a thought experiment argument. We saw earlier that Burge, like Kant, regards such arguments as reflective arguments.[13] We have resisted the proposal to classify such argumentation forms as targeting the same concept of

[10] See Section 8.4 below.

[11] A structurally similar characterisation of the 'state of innocence' of the finitist when she is working inside Primitive Recursive Arithmetic, is given in [**Dea15**, p. 53].

[12] See Chapter 3.

[13] See Sections 3.10 and 3.12.

reflection that the rationalists had in mind.[14] However, we now see that thought experiment arguments can be key components in type **6** reflection arguments.

At this juncture, there are two courses rationally open to you. *Either* you revise your commitment to your cognitive project, *or* you form a *belief* in the consistency of PA.

Suppose, for a moment, that at this juncture you do not form a belief in the consistency of PA, but instead remain agnostic about it. Then you can adopt an *instrumentalist* form of belief of PA that is difficult to distinguish from what I have called the state of innocence. You can resolve simply to continue with your mathematical practice unless and until you find a contradiction in PA. In other words, your *acceptance* of PA can be an acceptance *as if* PA holds rather than an unqualified belief in PA.[15]

The reason why this kind of instrumental acceptance is hard to distinguish from the state of innocence in which you unqualifiedly believe PA, is that you never *will* find a contradiction in PA, and even if you do, you will revise your practice in pretty much the same way as you would do if you had found the contradiction while in the state of innocence.

Nonetheless, your instrumental acceptance of PA is not the same as your full belief of PA in the state of innocence. Your instrumental acceptance is coloured by what you now take to be an epistemic possibility and which would undermine your cognitive project if it came to pass.

Concerning the doxastic aspect of your belief, it is admittedly *logically possible* for you not to change your unconditional belief in each of the axioms of PA while even at the end of the reflection process remaining agnostic about the consistency of PA. But it would be *irrational* to do so. It would be irrational even on a 'liberal' conception of rationality. Recognising as an epistemic possibility a situation of which you know that it would undermine your belief in the conjunction of the axioms of PA, rationally compels you to have less than full belief in some of the axioms of PA. For those who are sympathetic to theories of degrees of belief, the problem can also be phrased in quantitative terms. When coaxed to describe your mode of belief in quantitative terms, you give *maximal credence* to each axiom of PA. Yet you recognise as an epistemic possibility a scenario in which PA would not hold. This is irrational.

There are situations in which it is perfectly rational not to form a consistency belief as a result of the reflection process, and to withdraw to less-than-full acceptance, such as the instrumental form of acceptance that was sketched earlier. Suppose that your starting theory is not PA but standard set theory (ZFC), and you come to realise by means of the reflective process that you are relying on the consistency of ZFC, whereas you had not entertained the question of the consistency of ZFC before. You may, in that situation, not be sanguine that finding a contradiction in ZFC will never happen, even though you do not at present have even the vaguest inkling about how or where in set theory it might arise. (I know mathematicians who find themselves in this state.) In this situation, you may simply revise your unconditional belief of ZFC to a somewhat lower degree of acceptance. Your acceptance of ZFC becomes more cautious (or guarded, or provisional) than full

[14] See Section 3.13.

[15] For a discussion of the relation between acceptance and 'acceptance as if' in the context of empirical knowledge, see [**vF80**, Chapter 2, Section 3].

belief, even though this change does not leave a visible trace in your mathematical practice.

Suppose, however, that you maintain your *unreserved* commitment to your cognitive project through to the end of the reflective process. Then, if you are rational, you form an unqualified new belief: a full belief in the consistency of PA. This concludes stage two of the reflective process. This belief formation process is underwritten by an epistemic entitlement. As was argued earlier,[16] in the light of a liberal conception of rationality,[17] we have an *entitlement to reflection* on what we implicitly rely on or trust in.

Observe that this does not mean that you *voluntarily decide* to believe that PA is consistent! *If* you come to believe in the consistency of PA in this way, then it is not because you decide to do so: forming a belief is not a voluntary act.[18]

8.3.3. Arithmetisation. When you have arrived at this point, you have come to believe that PA is consistent. This is a new belief. But you have not yet acquired a new *arithmetical* belief. Nevertheless, a belief in an arithmetised consistency statement for PA can be obtained by continuing your reflection process along the following lines.

Presently you come to realise that, given a simple coding scheme, provability in PA is expressed straightforwardly by an arithmetical predicate Bew_{PA}. This is also not a straightforward process—it took the mind of Gödel to think this through. Your reasoning goes roughly as follows.

You define some convenient computable coding $\ulcorner \ldots \urcorner$ of terms and formulas of the language of PA ($\mathcal{L}_{\mathrm{PA}}$). You also construct a *standard* provability predicate $\mathrm{Bew}_{\mathrm{PA}}$ for PA.

You want to convince yourself that:

THESIS 8.1. For all $\varphi \in \mathcal{L}_{\mathrm{PA}}$: $\mathrm{PA} \vdash \varphi \Leftrightarrow \mathrm{Bew}_{\mathrm{PA}}(\ulcorner \varphi \urcorner)$.

You do this by *proving* this statement by mathematical induction (on the complexity of proofs). This statement relates syntax (symbols, terms, formulas) with numbers via your coding scheme. So, formally, this is an argument by mathematical induction in a language that does not only contain the familiar arithmetical vocabulary but also contains syntactic predicates and allows quantification over syntactic entities. This process of arithmetisation of PA constitutes **stage 3** of the reflective process.

It is of course possible that you reject meta-syntactic reasoning: if so, then you cannot carry out the proof of the statement. In this case, your reflective process will have ended at stage two. But I will assume that you accept the basic meta-syntactic reasoning required to prove the statement.

The point of spelling out what is involved in this argument in some detail is seeing that philosophical or semantic notions (such as rational belief, or truth) play no role in this reasoning. Moreover, and equally importantly, the theory in which this argument is carried out, is proof theoretically *conservative* over PA: for a proof of this fact, see [**Nic13**, Section 4.3]. So, in particular, the inductive argument above does not result in *circularity*: in constructing this argument, you

[16] See Section 2.9.
[17] See p. 67.
[18] See p. 60.

are not implicitly assuming a theory that is already strong enough to prove the consistency of PA.

As a particular instance of the new belief that you have acquired, you find that the consistency of PA is equivalent to the arithmetical statement $\neg\mathrm{Bew}_{\mathrm{PA}}(\ulcorner\bot\urcorner)$.[19] You combine this belief with the outcome of stage two of the reflective process, i.e., with your belief that PA is consistent. Thus you come to believe a new *arithmetical* sentence, i.e., $\neg\mathrm{Bew}_{\mathrm{PA}}(\ulcorner\bot\urcorner)$.

The assumption made at the beginning of this subsection, that you are using a *standard* provability predicate, is crucial. For instance, suppose that you were to formalise provability in PA instead as

$$\mathrm{Bew}_{\mathrm{PA}}(x) \wedge \mathrm{Con(ZFC)},$$

where Con(ZFC) is a standard way of formalising the consistency of ZFC in arithmetic. This new predicate would be co-extensive with the standard provability predicate $\mathrm{Bew}_{\mathrm{PA}}$. But *that* this *nonstandard* arithmetical provability predicate captures provability in PA can only be proved on the assumption Con(ZFC). However, if your process of reflecting on the consistency of PA required the consistency of set theory as an assumption, then it would of course not give you, at the end of stage three of the reflective process, a *new* entitled arithmetical belief. Similarly, there are nonstandard provability predicates $\mathrm{Bew}^*_{\mathrm{PA}}$ such that already PA proves $\neg\mathrm{Bew}^*_{\mathrm{PA}}(\bot)$.[20] Such provability predicates will not satisfy the Hilbert-Bernays derivability conditions. As we have seen earlier,[21] these requirements are seen as adequacy copnditions that any *bona fide* provability must satisfy. In other words, such non-standard provability predicates fall short of really expressing provability in PA.

This concludes the description of stage three of the reflective process, which is also the end of the reflective process as a whole. If it is at least roughly accurate, then which epistemological lessons can we draw from it?

8.4. Cognitive work

If you rely on a presupposition of your cognitive project, and are entitled to do so, then you are *entitled to articulate* what you are relying on in engaging unconditionally in your cognitive project. In this situation you are *entitled to believe* the presupposition of your cognitive project. You have 'warrant for nothing' in Wright's sense, but—*pace* Wright—not just warrant for trust, but warrant for belief. In fact, it is not completely accurate to describe the upshot as 'warrant *for nothing*'. You have earned your epistemic warrant for believing in the consistency of PA by doing cognitive work that carries cognitive risk. Nonetheless, the reflection process that you have gone through is not a philosophical proof, nor an argument, nor a justification for the consistency of PA. That you have not given independent justification for the statement that expresses the consistency of PA is not an epistemic problem for you: you did not need to.

The assumption, in the fictional tale, that you are *justified* in your belief in the axioms of your starting theory PA to begin with (regardless of whether or not you are *aware* that you are so justified), is essential: your epistemic entitlement

[19] '$\bot$' stands for your favourite contradiction.

[20] For a discussion, see [**Fra04a**, Section 12.2].

[21] See Section 3.8.

to believe in the consistency of PA, and therefore also your entitlement to believe $\neg\mathrm{Bew}_{\mathrm{PA}}(\ulcorner\bot\urcorner)$, *depends* on your justification for believing the basic axioms of PA. Suppose for a moment instead that your starting theory is not the mathematical theory PA at all, but the teachings of a guru whom you have started to consult and base your beliefs on, simply because you assume that he is holy.[22] At some point you become aware of the fact that you are relying on the guru, while you continue to rely the guru in the same way as before. Then you form a belief in the reliability of the guru, but you are not *epistemically entitled* to this belief.

We have seen that we do not have, in epistemology, anything like a clean definition of Gettier cases. But we know that Gettier can strike in the domain of mathematical beliefs as it can in other cognitive domains.[23] You may derive, for instance, a true statement from a false mathematical axiom that you are justified in believing; then you are justified in believing the true statement, but you do not know it. So having true justified belief in PA does not entail that you *know* PA, and your true epistemic entitlement to believe $\neg\mathrm{Bew}_{\mathrm{PA}}(\ulcorner\bot\urcorner)$ does not entail that you know this proposition.

But *if you are not in a Gettier situation*, then you have more than justified true belief in the axioms of PA (and in theorems of PA): you *know* them. And then, after your process of reflection, you have acquired more than an epistemically entitled true belief in $\neg\mathrm{Bew}_{\mathrm{PA}}(\ulcorner\bot\urcorner)$: you *know* this proposition. You have acquired knowledge of a cornerstone proposition of your cognitive project.

If your justification for the axioms of PA was *a priori* to start with, then the epistemic entitlement of your belief in the consistency of PA will likewise be *a priori*. The counterfactual reasoning that led you to believe that, in your cognitive project, you presuppose the consistency of your practice, is *a priori*. In the first reflective stage, you used an argument by mathematical induction. But if your justification for the arithmetical instances of mathematical induction was *a priori*, then so will, presumably, be your justification for applying mathematical induction to a formula involving the predicate 'I am disposed to believing x' (where x is an arithmetical sentence). Something similar can be said about the third reflective step. If your justification for the arithmetical instance of mathematical induction was *a priori*, then so will be, presumably, your justification for applying mathematical induction for a formula involving syntactic predicates.

In the course of your reflective process, you have done justificatory work. You might wonder if the theories in which these justificatory arguments are implicitly carried out, do not already entail the consistency of PA. If so, then your reflective process were circular.

But these worries are unfounded. The inductive argument in stage one is clearly carried out in a theory that is conservative over PA.[24] For the inductive argument in stage three this is slightly less straightforward, but we have seen that it is also carried out in conservative extension of PA.[25]

You can, of course, object to the mathematical induction argument in stage three, perhaps because you are loath to use predicates that are not fully arithmetical

---

[22]This example was suggested to me by Cezary Cieśliński.

[23]See Section 1.1.

[24]More precisely, it is carried out in PA formulated in an extended language, with the induction axiom applying to the extended language.

[25]See Section 8.3.3.

in the induction axiom. If this is your view, then your reflective process ends at the end of stage two at the latest, and you do not acquire a fundamentally new *arithmetical* belief. Nonetheless, if your reflective process carried you to the end of stage two, you have still acquired a fundamentally new entitled belief: the belief that PA is consistent. You can also object to the mathematical induction argument in stage one. In that case, you do not bring yourself to a belief in the consistency of PA by an reflective argument along the lines that I have sketched.

At the end of your reflective process, you have not *derived* $\neg\mathrm{Bew}_{\mathrm{PA}}(\ulcorner\bot\urcorner)$ from more fundamental mathematical and/or philosophical principles. An essential moment of *entitled* belief formation occurs at the moment in stage two where you *form* a belief in the consistency of PA *while* maintaining your unreserved acceptance of PA. This is a cognitive act for which you provide no justification. But, in the circumstances you are in, you are epistemically entitled to proceed in this way (entitlement to reflection).

I take all this to be a vindication of the implicit commitment thesis. But my epistemological analysis does not constitute evidence for Feferman's strong interpretation of the implicit commitment thesis, according to which one *ought* to accept proof theoretic reflection principles for any mathematical theory that one unconditionally accepts. Throughout the reflection process, I have assumed van Fraassen's 'liberal' conception of rationality. In accordance with this view of rationality, I maintain that it would not be irrational for you to refrain from following the reflective process of Section 8.3 through to the glorious end. I am *not* claiming that objecting to the inductive argument in stage three, or even to the more elementary inductive argument in stage one, would be irrational. At the point when you realise that you have been relying on the consistency of PA, you are also *rationally permitted* to choose not unconditionally to rely on PA in the future and to revise your initial beliefs instead.

Moreover, going through the reflection process described in Section 8.3 is of course not the *only* way of coming to know $\neg\mathrm{Bew}_{\mathrm{PA}}(\bot)$. To illustrate this, let us briefly go back to the (admittedly somewhat naive) scenario that was briefly sketched in Section 8.2.1, where you have verified that the axioms of PA hold in the 'standard model'. Instead of going through the reflection process described above, you might go on to argue by mathematical induction, using a Tarskian compositional notion of truth, that all theorems of PA are *true*, and that therefore, since $\bot$ is not true, PA must be consistent.[26] In this way, you might obtain *justified* (and not merely entitled) belief in $\neg\mathrm{Bew}_{\mathrm{PA}}(\bot)$. Or you might verify that the axioms of *second-order* number theory hold in the 'intended model', and *derive* $\neg\mathrm{Bew}_{\mathrm{PA}}(\bot)$ from them. Indeed, it is well-known that in this way the scope of mathematical knowledge can be extended in much more dramatic ways than by iterated consistency extensions.

It is also not part of the thesis defended in this Chapter that the implicit commitment of PA is exhausted by the reflection process discussed in Section 8.3. All I claim is that going through this reflection process is *one* way of coming to know $\neg\mathrm{Bew}_{\mathrm{PA}}(\bot)$.

The interest of this particular reflection process is epistemological. It constitutes a hitherto unexplored way of acquiring epistemic warrant for mathematical statements that are independent of your beliefs. It is fundamentally different in

[26] Cfr infra: Section 9.3.

nature, for instance, from your epistemic warrant for adopting a new mathematical axiom.

When you reach the end of your reflective process, you have come to know a fundamentally new arithmetical statement: $\neg \mathrm{Bew}_{\mathrm{PA}}(\bot)$. But you have not come to know *that* it is fundamentally new, i.e., that it was not accessible to you in your state of innocence. This is because, in the first stage of the reflective process, you come to believe that what you are disposed to believe (as far as your mathematical work goes), includes PA, not that it *coincides* with PA.

In the state of innocence, you essentially "are" a Turing machine e that enumerates PA. You can acquire the first person knowledge that you are, insofar as mathematics is concerned, the machine e.[27] *How* can you come to know that as far as arithmetic is concerned, you are e? It does not happen by *intuition* or *direct introspection* (unless the meaning of those terms is stretched). All you have to go on is a finite set of *examples* of mathematical axioms that you believe, and theorems that you have come to believe by deriving them from the axioms using classical logic. Extrapolating from this finite collection of examples, you form the *hypothesis* that you are guided by e, and you come to believe this hypothesis. You are using some form of ampliative reasoning: we may call it *abduction*.

Your abductive argument is clearly fallible. We may suppose, however, that in this instance, you not only arrive at a true conclusion, but that in addition you are *justified* in believing this conclusion on the basis of your abductive considerations. That abductive arguments sometimes lead to justified beliefs is fairly widely accepted. But there is no consensus among epistemologists on *how* abductive arguments can generate justified beliefs. I have nothing to contribute to this large epistemological debate except to say that some reliabilist account is probably called for. This should not, however, be taken to imply that this reliabilist story must then account for *all* forms of knowledge. Indeed, it is perhaps doubtful that the very same epistemological story that accounts for abductive reasoning will also account for your knowledge of the axioms of PA.[28]

You might worry that it might not be possible for you to know that, as far as your mathematical work goes, you are a Turing machine. Lucas and Penrose have famously argued that it is not even possible for you to *be*, as far as your mathematical work goes, a Turing machine. But, as mentioned earlier, it is widely held that their arguments are unpersuasive. Reinhardt has argued that, as far as your mathematical work *and your knowledge of your own work goes*, for every Turing machine e, you cannot know that e enumerates what you know [**Rei85**, Theorem 5].[29] Reinhardt's purported counterexample is a sentence that is produced by a simple diagonal argument; it contains the predicate 'knowledge' (or, to be more precise: 'absolute provability'), which expresses a non-mathematical concept. But here we are concerned only with your (true and justified) *arithmetical* beliefs. In this context, therefore, Reinhardt's considerations do not apply.

Despite all this, you may nonetheless be sceptical about the abductive argument given above. This would (again) not make you irrational; it would just mean that you have not come to know that the new consistency beliefs that you have

[27] A brief discussion of this reflective can be found in [**Fra04a**, p. 216].

[28] See Section 1.1.

[29] Carlson later showed that the Benacerrafian hypothesis ([**Ben67**] that you are a Turing machine but you don't know which one, *is* consistent [**Car00**].

acquired in the reflective process described in section 8.3 are *fundamentally* new. The abductive argument that I have described in this section is *empirical* in nature. For this reason, it is not clear if along the lines that I have sketched you can come to have *a priori* knowledge that you have acquired a fundamentally new arithmetical belief.

8.5. Ramifications

In Section 8.2, I took PA as the starting point of the reflection process described in the Section thereafter. But the account that I have given in the previous Sections aims at being fairly generally applicable. Since inductive arguments play a role in the reflection process, the starting theory must contain a modicum of mathematical induction. So my account of reflection on consistency does not apply to very weak starting points, such as Robinson's Q, or Primitive Recursive Arithmetic. But it does apply to situations where certain other fragments of PA are started from, and it applies to all supertheories of PA, such as second-order PA or ZFC. In particular, the account given in the previous sections applies to the situation where the starting theory is our most encompassing mathematical theory. The account of reflection intends to show that even for that theory, we can come to know its consistency without deriving it from stronger principles.

In Section 7.3, I did not object to Dean's claim that finitism is an epistemically stable position in the foundations of mathematics. Moreover, as we have seen, it has been argued that accepting all of PA and no mathematics that goes beyond it is likewise a stable position [**Isa87**], and so is Fefermanian predicativism. How are such claims compatible with the argument that was developed in Section 8.3? After all, in that Section a reflection process is described by means of which someone who accepts all and only the principles of PA can come to know that PA is consistent. In addressing this question, I will now concentrate on finitism because this position has received a fair amount of attention in recent literature.[30]

Tait has argued—convincingly, in many scholars' view— that the extension of finitistically acceptable mathematics is captured by the system of Primitive Recursive Arithmetic [**Tai81**]. He also pointed out that the outer limits of finitism can only be seen from a vantage point that is external to finitism proper [**Tai81**, Section IV].[31] This does not make finitism internally unstable. Locally, the finitist can see of every proof principle of Primitive Recursive Arithmetic that it is justified. But she has no way of verifying that *only* the proof principles that are included in Primitive Recursive Arithmetic are legitimate. Indeed, such a claim involves a *general* concept of function, which the finitist does not have [**Dea15**, p. 53].

This means that the finitist does not accept all the steps of the reflection process that are described in Section 8.3. For instance, she will not accept the inductive argument in the first stage of the reflective process. After all, such an inductive argument is not a finitist proof![32] (The same holds, of course, for the inductive argument in stage three of the process.) In accordance with the 'British' conception

[30]See for example [**Par07**, Chapter 7], [**Dea15**].

[31]See also [**Dea15**, Section 4.1]. Something similar can be said about Feferman-style predicativism.

[32]She could, however, come to believe that she is a "Primitive Recursive Arithmetic-machine" by abductive means. In that case, she could come to have an entitled belief in the consistency of Primitive Recursive Arithmetic. But, again, refusing to engage in the relevant abductive reasoning would not make her irrational.

of rationality,[33] I maintain that thus refraining to accept some of the steps in the reflective process of Section 8.3 does not make the finitist irrational. It is in this sense that I wholeheartedly concur with Dean's thesis that, at least for all that what is said in this Chapter, finitism is an epistemologically stable position.

Nonetheless, *excessive* caution is a sin against reason. Perhaps the Hilbertian finitist is guilty of exactly this. After all, one will be hard pressed to find more than a handful of respected mathematicians or philosophers of mathematics living today who are finitists in Hilbert's sense of the word. But to argue for this thesis would take us too far afield, so I will not pursue this further here.

In the light of the foregoing, what becomes of Feferman's view of implicit commitment that was discussed in Section 7.1?

Recall that Feferman describes what one is implicitly committed to when one believes a mathematical theory T1 as "what one ought to accept, on the same fundamental grounds, when one accepts L1" [**Fef88**, p. 131]. Moreover, he holds that uniform reflection for T1, and hence a fortiori also the consistency of T1, belongs to the implicit commitment of T1.

First of all, it should by now be clear that in my view, the phrase "ought to accept" is too strong in this context; it should be replaced by "*can* rationally come to believe".[34] Secondly, the expression "on the same fundamental grounds" is, strictly speaking, an overstatement. A mathematical reflection process of the kind that we have discussed is complicated, and it constitutes an intellectual achievement. It is true that in the course of your reflective process, you have not *deduced* the consistency of PA from more fundamental principles: the crucial moment of belief formation is one of epistemic entitlement rather than of justification. Yet we have seen in Section 8.4 that *grounds* are involved in the process that exceed what is given by PA: mathematical induction *in an extended language*, and a thought experiment ("what if I were to derive a contradiction in PA?"). So the implicit commitment of a mathematical theory T1 should in my view rather be described as what one *can* rationally come to believe when one reflects on one's belief in T1.

8.6. Strengthenings?

Let us turn to the question whether *something like* the reflection process in section 8.3 can result in warranted belief in proof theoretic reflection principles stronger than consistency. Clearly, the crucial question here is whether our description of the relevant counterfactual reasoning process[35] in stage 2 would then, *mutatis mutandis*, go through.

We have seen that the consistency statement, which is our weakest reflection principle, takes the form:

$$\text{If } 0 = 1 \text{ is provable in my practice (PA), then } 0{=}1.$$

Local reflection, which is the next strongest reflection principle that we have considered, is the *scheme* that is obtained when in this consistency statement, we allow $0 = 1$ to be replaced by any arbitrary closed arithmetical formula φ.

Let us first consider the case where φ is some concrete Δ_0 arithmetical sentence, so that either φ or $\neg\varphi$ is provable in your mathematical practice. I will argue that

[33] See p. 67.

[34] The "liberal" interpretation of implicit commitment that I am adopting here is also suggested in [**Fuj11**, p. 915].

[35] See Section 8.3.2.

in this situation, the process of reflection goes through pretty much as in Section 8.3.

Consider first the case where φ is true. Then φ is provable in your mathematical practice, from which you derive by logic:

If φ is provable in PA, then φ.

So this case is unproblematic. (Observe that since φ is a concrete arithmetical statement, you do not need, in your reasoning process, to use the concept of truth.)

Now consider the case where φ is not true. Then in your arithmetical practice, you prove $\neg\varphi$. In this situation, you ask yourself:

What if also φ is provable in your practice?

Then, again, as a mathematician, you find yourself at the epistemic crossroads that we are by now familiar with. There are only two rational ways to proceed: either you revise your acceptance of PA, or you accept the statement that φ is not provable in your practice. In the latter case, you have accepted the instance for φ of Rfn(PA). (Again, you do not use the concept of truth in your reflection process.) Since the above argument goes through for *any* concrete Δ_0 arithmetical sentence, we may conclude that the reflection process can result in explicit acceptance of any instance of Δ_0-Rfn(PA). In this sense, Δ_0-Rfn(PA) is "implicit" in PA.

Next, let us look at more complicated instances of Rfn(PA). Suppose in particular that φ is *independent* of your pre-reflection mathematical practice. For definiteness, but without loss of generality, take φ to be the Paris-Harrington sentence.[36] In this case, I will argue, the situation is more complicated.[37]

Of course we again focus on the relevant bit of counterfactual reasoning. In the case of consistency, you asked yourself a straightforward, simple counterfactual question. But now, in the case of local reflection, you have to ask yourself the following slightly more complicated, *conjunctive* question:

Might I be able to prove φ in my mathematical practice,
while at the same time $\neg\varphi$?

The second conjunct is now really needed in this question. Unlike in the case of consistency or Δ_0-Rfn(PA), since $\neg\varphi$ is independent, it *cannot* be proved in your pre-reflection theory. Thus be taken for granted, but must be explicitly mentioned in the counterfactual question.

We observe that, as before, the concept of truth is not used in this *what if*-question. Nonetheless, arguably the counterfactual question at issue here can only be properly understood by a mathematician who has at least *the beginnings of a concept of truth*. It can only properly be understood by you if you are at least open to the possibility that an arithmetical sentence—the sentence $\neg\varphi$, in the question at hand—is not always equivalent to its provability in your practice. If not, then your question collapses into the counterfactual question that is at the heart in reflecting on consistency. This stands in contrast to the earlier counterfactual question concerning the provability of contradictions. *That* question is properly understandable by any mathematician, for you cannot be a mathematician without having a passable concept of mathematical proof.

---

[36] See [**PH77**].

[37] I am indebted to Hannes Leitgeb for insightful suggestions concerning this matter.

This being said, *if* you are able to pose the relevant counterfactual question to yourself, then it seems to me that you find yourself, as before, at a rational crossroads: answering the counterfactual question in the affirmative *whilst* maintaining your full acceptance of PA, would be irrational. In this situation, it seems to me that the reflection process of Section 8.3 can lead to rational belief in any instance of Rfn(PA).

The situation for uniform reflection is still more complicated. It seems hard to see how someone who does not even have a proto-concept of truth can even ask himself the relevant counterfactual question, which is (for φ an arithmetical predicate):

Might there be a natural number n such that I can prove $\varphi(\overline{n})$,
while at the same time $\neg\varphi(n)$?

But in the counterfactual question, the concept of truth does not *explicitly* appear. And if the question makes proper sense to you, then it seems pragmatically incoherent to answer it in the affirmative whilst leaving the acceptance of your pre-reflection theory unchanged.

The case of global reflection is completely different. In stage 2 of the relevant reflection process, the pertinent counterfactual question is:

Might there be an arithmetical sentence that I am able to prove (in PA),
but that is at the same time not *true*?

Here you need to quantify over an infinite collection of sentences in a way that requires a concept of truth. So to go through this reflection process, you must have a truth predicate in your vocabulary. And having a concept of truth goes beyond your purely mathematical commitments.

Now it could be that the required truth concept can somehow be obtained by reflecting on your mathematical practice. Indeed, it might be that also new *concepts* can be implicit in theories that a mathematician already accepts.[38] For instance, even though the general concept of well-ordering is Π^1_1-complete, small transfinite wellorderings can be defined in arithmetic. Hence perhaps it can be argued that at least a fairly weak *concept* of transfinite ordinal is "implicit" in arithmetic. Similarly, perhaps, even though the concept of arithmetical truth is arithmetically undefinable, nonetheless a form of arithmetical truth is "implicit" in arithmetic.

This line of reasoning seems to be at least hinted at by Halbach in his textbook on axiomatic truth [**Hal11**, p. 324–235]:

> [T]he explicit endorsement of Peano arithmetic seems to bring an implicit commitment to principles (such as consistency) that cannot be proved in Peano arithmetic, but it also brings commitment to further conceptual resources, namely soundness and truth, that cannot be formulated in the language of arithmetic. At least, this will be the case if one agrees with the usual motivation of proof-theoretic reflection principles, according to which commitment to the soundness of Peano arithmetic is implicit in the acceptance of this theory. For the global reflection principle is the source of all the reflection principles that can be formulated in the language of arithmetic.

[38] Cfr infra, Section 9.2.

This line of thought is intriguing. But it needs to be developed in more philosophical detail, and this is a non-trivial task. In particular, we presently lack a compelling philosophical argument for the thesis that it is irrational to accept Peano arithmetic while remaining sceptical or agnostic about the acceptability of a concept of truth and basic principles (such as typed disquotational axioms) that govern its behaviour.

One main question here is how strong the concept of truth is that the reflector carries into the reflection process. If she accepts the compositional truth axioms, and if, as required by the reflection process, her acceptance of mathematical induction is open-ended, then she can *already* justify global reflection for PA.[39] If that is so, then there is no need for her to reflect on her acceptance of PA. In order to avoid this form of circularity, it is preferable that the reflecting mathematician starts out not with a compositional, but with a *disquotational* conception of truth. This suggests that we look into the philosophical ramifications of the connections between reflection principles and disquotational truth. This theme will be explored in the next Chapter.

Summing up, we may conclude that the processes of reflecting on consistency and reflecting on stronger proof theoretic reflection principles are, from an epistemic point of view, significantly different. Any mathematician is capable to go through the former reflection process. But as the reflection principles get stronger, more conceptual resources are needed to carry out the relevant reflection process.

8.7. Burge revisited

In Section 3.12, we saw how Tyler Burge argues that the rationalist philosophers from the early modern period attribute three *cardinal properties* to reflection [**Bur13d**, p. 535–537]:

(1) In reflection an individual brings to articulated consciousness steps or conclusions that are implicitly present, subliminally or unconsciously, in the individual's mind before reflection.
(2) Reflection can yield *a priori* knowledge of objective subject matters, beyond thoughts that the reflector is engaging in.
(3) Successful reflection requires skilful reasoning and is difficult: it is not a matter of one-off introspection or intuition.

We also saw that Burge agrees with Theses 2 and 3, but disagrees with Thesis 1.

Burge was, however, not only concerned with what we have called type **6** reflection, which is the type of reflection that early modern philosophers were primarily concerned with. As we saw in Section 3.13, Burge's conception of philosophical reflection also includes certain kinds of conceptual analysis, as well as philosophical thought experiment arguments. Although Burge's arguments also provide confirmation for Thesis 3 and disconfirmation for Thesis 1 if only type **6** reflection is at issue, his arguments for Thesis 2 rest on examples that we do not recognise as type **6** reflection. For this reason, Burge's arguments for Thesis 2 do not carry conviction.

The reflection process that was discussed in the present chapter is a clear case of type **6** reflection. After all, in stage 1 of this process[40] you obtain beliefs about

[39] See [**Hor11**, Theorem 30, p. 76].
[40] See Section 8.3.1.

your own cognitive state by considering your own doxastic commitments. Moreover, it is not a case of simple introspection,[41] since discursive elements play a crucial role in all three stages of the process.

In the present Chapter, then, Theses 1, 2, and 3 can be taken to have been subjected to a test by applying them to a concrete example of type **6** reflection in the foundations of mathematics, viz., reflection on implicit commitments associated with the acceptance of mathematical theories. I claim that the process involved in proof theoretic reflection is in accordance with what Burge regarded as the cardinal properties of philosophical reflection.

First, the reflective movements do not consist in drawing to the level of your consciousness representations that were already vaguely and subconsciously present. The beliefs that you form in reflection were not indistinctly, subliminally, or subconsciously, present in your mind in any way before the acts of reflection. This constitutes additional evidence against Thesis 1 for type **6** reflection.

Secondly, reflection is a complicated process indeed. I have concentrated on the simplest form of reflection, and the story already has a significant degree of epistemological complexity. It is even more complicated when we focus on stronger reflection principles. This constitutes additional evidence for Thesis 3 for type **6** reflection.

Thirdly, reflection on what you explicitly accept can yield not only be new justified beliefs, but even new knowledge. Moreover, this new knowledge can be a priori.

Lastly, and most importantly, in the reflection process that we have described, you have acquired knowledge not just about your mind or your commitment. You have acquired new knowledge about the world outside your mind, viz., the world of numbers. The statement $\neg \mathrm{Bew}_{PA}(\bot)$ is, after all, a *purely arithmetical proposition.* Of course this assumes that the reflection process includes stage 3.[42]

The considerations of the present Chapter, which go somewhat beyond the arguments that Burge has given for his stance, therefore indicate that Burge's stance concerning the three cardinal Theses on philosophical reflection are correct *even* if only type **6** reflection is intended.

We have seen that beside type **6** reflection, counterfactual reasoning plays a central role in the reflection process that we have considered. Such instances of counterfactual reasoning are closer to the Putnamian thought experiments that Burge lists as instances of reflection. We have earlier labeled such reasoning processes as "type **8** reflection",[43] even though it has to be recognised that most philosophers do not see such reasoning processes as instances of reflection in a technical philosophical sense at all. At any rate, this shows that type **6** reflection mostly shows its strength only when it operates in combination with other distinctly philosophical ways of reasoning. It makes little sense to speak of the force of type **6** reflection *on its own.*

[41] See Section 3.8.

[42] See Section 8.3.3.

[43] See p. 95.

CHAPTER 9

Truth, Justification, Reflection

The previous Chapter discussed the force of reflection in purely mathematical contexts. In this Chapter, we investigate how reflection principles can be justified on the basis of non-mathematical concepts and arguments. We consider philosophical arguments for reflection principles that use the concept of truth; to a lesser extent, we consider how reflection principles can be argued for by using epistemic notions. Attempts to justify proof theoretic reflection principles using the concept of truth arose naturally in this context, because Global Reflection, which is the Ur-reflection principle (in the proof theoretic sense), contains the concept of truth. Towards the end of this Chapter, we also discuss strategies for justifying set theoretic reflection principles.

9.1. Types and principles of infinity

In the previous Chapter, we explored a particular form of epistemic reflection as one possible way of reaching warranted belief in a stronger mathematical theory from a starting point of warranted belief in a weaker mathematical theory. The "engine" in this reflection process is implicit commitment, and we have seen that this reflective process is iterable. Nonetheless, we have also seen that the strength of such a reflection process is probably relatively modest: it consists at best in iterating local reflection on a starting theory S a number of times. But there are stronger *systematic engines* for boosting the strength of the mathematical theories that one has warranted belief in, as we will now see.

Suppose that our starting theory is a standard arithmetical theory, such as PA. This theory is formulated in the framework of classical first-order logic, where the quantifiers range over the domain of discourse (the natural numbers). In *second-order logic*, we also have quantifiers ranging over *properties* of natural numbers. These properties are interpreted extensionally, so in effect these new quantifiers range over *sets of* natural numbers. The new quantifiers are governed by natural analogues of the logical rules for the first-order quantifiers, plus a *comprehension scheme* for properties of numbers. When we move from first-order logic to second-order logic as our framework for arithmetic, we also replace the first-order mathematical induction scheme by the second-order mathematical induction axiom, and reach the theory PA^2, which is a strong and natural second-order arithmetical theory.[1] We can extend our logical framework further, by adding quantifiers that range over *properties of properties* (or: sets of sets) of natural numbers, and reach the standard theory of third-order arithmetic PA^3. Then we can go on, iterating this process further, possibly into the transfinite—but we will restrict ourselves to finite levels of the type hierarchy in what follows.

[1] See Section 4.1.4.

Expanding the framework from n-th order logic to $n+1$-th order logic is a larger step than extending the theory by proof theoretic reflection principles or by strengthening induction. Indeed, we have seen that PA^2 proves long iterations of uniform reflection applied to PA.[2] In a similar way, the higher-order theory PA^{n+1} will prove long iterations of uniform reflection applied to PA^n. In the early 1930s, Gödel describes this process of climbing up Russell's type theoretic hierarchy as follows [**Göd33b**, p. 48]:[3]

> For any formal system you can construct a proposition—in fact a proposition of arithmetic integers—which is certainly true if the system is free from contradiction but cannot be proved in the given system. Now if the system under consideration (call it S) is based on the theory of types, it turns out that exactly the next higher type not contained in S is needed in order to prove this arithmetic proposition, i.e., this proposition becomes a provable theorem if you add to the system S the next higher type and the axioms concerning it.

The question whether proceeding from PA^n to PA^{n+1} might be *warrant-preserving* has received much attention in the twentieth century. For definiteness, let us concentrate on the transition from PA to PA^2. Then the questions to consider are:

(1) Is the second-order induction axiom warranted?
(2) Is the second-order comprehension scheme warranted?

Kreisel has argued that the first question should be answered in the affirmative. He proposed the general principle that our warrant for being disposed to believe in each instance of an axiom scheme derives from our warrant for believing the next-order *single-sentence* uniformisation of the scheme. In particular, concerning the relation between the first order mathematical induction scheme and the second order mathematical induction axiom, Kreisel writes [**Kre67**, p. 148]:

> A moment's reflection shows that the evidence of the first order axiom schema derives from the second order axiom [...]

This idea is implemented in subsystems of second-order arithmetic.[4] It will play a role in discussions later in this chapter.[5]

The second question asks which formulas define properties of natural numbers. It has been discussed extensively in the twentieth century, and we will not go deeply into this discussion here. Some have argued that because the full second-order comprehension principle[6] allows second-order quantifiers in the formula $\varphi(x)$, it is somehow *viciously circular*. These foundational thinkers—Weyl, for instance—argue that only *predicative* second-order theories, such as ACA, are warranted. Others—Gödel, for instance—have argued that *every* formula of the extended language determines a property of the natural numbers. Indeed, the full

[2] See Section 6.2.

[3] Gödel clearly saw this scenario already when he proved his incompleteness theorems: see [**Göd31**, footnote 48a].

[4] See Section 4.1.4.

[5] See Section 9.4.

[6] See p. 110.

second-order comprehension scheme has, at least so far, not led to seemingly untoward consequences—certainly not to contradictions!—and predicative restrictions on properties of sets are not respected in mathematical practice.

How far does the type hierarchy continue? In the 1920s, closure principles for the levels of the hierarchy came to be seen as *mathematical* questions that should follow from mathematical axioms. Such closure principles are contained in the *first-order* axiomatisation of set theory that we are all familiar with. ZFC contains an *internalisation* of the process of going to ever higher types. It postulates very large infinite *ranks* that correspond to levels of the type hierarchy.

We started the type theoretic hierarchy from the first-order theory PA. But now we we have reached a much stronger first-order theory, ZFC, which can be taken as a new starting point. As in the case of PA, we can now move to second-order ZFC (ZFC^2), third-order ZFC, and so on. Formally, it is completely unproblematic to generate a new type theory in this way. In particular, as before, ZFC^{n+1} will be proof-theoretically stronger than ZFC^n. Indeed, ZFC^{n+1} will prove long iterations of uniform reflection over ZFC^n.

But from an interpretational point of view, matters are much less straightforward. Earlier, we took the second-order quantifiers to range over *sets of* elements of the first-order domain. But the domain of discourse of ZFC is expected to contain all the sets there are. So how should the second-order quantifiers of ZFC^2 be interpreted? This is again a question that has received much attention, and we cannot do justice to the discussion here. We will return to it later in this Chapter.[7]

Assume, then, at least for the sake of argument, that we are warranted in believing the axioms of ZFC (and perhaps even of ZFC^2). Then, using the reflection process that was discussed in the previous chapter, we might furthermore come to have warranted belief even in modest iterations of local reflection applied to ZFC. Might warranted belief in even stronger mathematical theories somehow be obtained?

There are indeed other ways of mathematically strengthening mathematical theories. Peano Arithmetic is obtained from Robinson's system Q by adding the mathematical induction scheme. Zermelo's set theory was massively strengthened in the 1920s by adding the replacement scheme. Ways of strengthening ZFC have also been proposed.

Gödel suggested very early on that set theory can be strengthened by adding *strong axioms of infinity*, also known as *large cardinal axioms*. In set theory ,adopting axioms of infinity takes the place of movings to higher types [**Göd32**, p. 237]:

> In case we adopt a type-free construction of mathematics, as is done in the axiom system of set theory, axioms of cardinality (that is, axiom postulating the existence of sets of ever higher cardinality) take the place of type extensions, and it follows that certain arithmetic propositions that are undecidable in Z become decidable by axioms of cardinality [...]

For most (first-order) large cardinal axioms I, the theory $ZFC+I$ proves (first-order) statements that ZFC^2 cannot prove. On the other hand, ZFC^2 proves statements that $ZFC + I$ cannot prove for the simple reason that $ZFC + I$ does not prove any second-order statements. So, strictly speaking $ZFC + I$ and ZFC^2

[7] See Section 9.8.

are incomparable in strength. But in a weaker sense, $ZFC + I$ is the stronger one, for it will usually prove the consistency of ZFC[2].

We have seen that the hierarchy of large cardinal principles is linear. But it is *less systematic* than the type hierarchy. On the one hand, most of the large cardinal principles have a characterisation in terms of elementary embeddings. On the other hand, the motivation behind different large cardinal axioms differs. And there is never a guarantee that significantly stronger consistent and warranted large cardinal principles can always be dreamt up. *Forcing axioms* are a very different way of strengthening our set theoretic axioms. But here, too, we have at the moment no hierarchy of warranted forcing axioms of ever-increasing strength.

From the axiom that all sets are constructible ($V = L$), *most* interesting set theoretic questions can be decided. Unfortunately, the axiom $V = L$ is incompatible with the existence of a measurable cardinal. For this reason, most set theorists believe that it is false. Woodin's *Ultimate-L* programme[8] is an attempt to identify V with an *L-like* structure ("Ultimate-L") containing all large cardinals that are thought to be consistent with ZFC, and which is such that a natural theory describing it still decides most interesting set theoretic problems. Unfortunately, Woodin's programme has been stuck for some time: "no L-like" models for supercompact cardinals have yet been found.

Wang reports that Gödel believed that *all* large cardinal axioms can be derived from reflection principles [**Wan96**, p. 285]:

> Generally I believe that, in the last analysis, every axiom of infinity should be derivable from the (extremely plausible) principle that V is indefinable, where definability is to be taken in [a] more and more generalized and idealized sense.

Moreover, Gödel held out the hope that also the most recalcitrant problems lower down in the hierarchy can be decided using large cardinal axioms.[9]

Both of these beliefs are debatablel, however. Firstly, we know since the 1960s that large cardinal axioms do not check the ability of the forcing technique to change the cardinality of power sets almost at will.[10] So it seems that not *all* interesting questions concerning lower ranks can be decided by extending the rank hierarchy ever further. Secondly, we have seen earlier that the theoretic reflection principles that currently are most widely supported at most reach to the realm of 1-extendible cardinals.

In addition, we must face the question: *what reasons do we have that ontological reflection principles are true?* This is a question which set aside for now: we will occupy ourselves with it later in this Chapter.

In sum, at present the possibility cannot be excluded that our attempts to reduce set theoretic incompleteness by new set theoretic axioms reaches an end point S. Using the proof-theoretic reflection processes described in the previous Chapter we may then get "an epsilon" further and reach a warranted theory $S + \varepsilon$ "in the limit".

---

[8] See [**Woo17**].
[9] See [**Göd47**].
[10] See [**LS67**].

9.2. Implicit commitment of concepts

In the philosophical literature on implicit commitment, authors often speak about implicit commitment of *theories*. But we have seen earlier that authors sometimes also speak of implicit commitment of *concepts*.[11] Thus the question arises whether theoretic implicit commitment should be distinguished from conceptual implicit commitment. It is notable that the pioneers of proof theory do not seem to distinguish clearly between the two.

Roughly, the notion of implicit commitment of concept has its roots in the work of Gödel, who, I suspect, influenced Kreisel's views about these matters. The notion of theoretical implicit commitment is more prominent in the work of Feferman, except perhaps in his later work on implicit commitment, which seems to be more connected to a form of conceptual implicit commitment.

At first blush, commitment to proof theoretic reflection principles appears to be a form of theoretical implicit commitment, for it springs from accepting what is proved by a formal theory. Commitment to predicative fragments of arithmetic, and to truth theories appear to be instances of conceptual implicit commitment, for they spring from the acceptance of concepts (definability, truth).

Moreover, there is an important difference between commitment to fragments of second-order arithmetic on the one hand, and commitment to truth principles on the other hand. The former is *ontologically non-conservative*, whereas the latter is only *ideologically non-conservative*. Observe, incidentally, that the transition from accepting certain predicates to accepting certain collections can be resisted. Quine, for instance, always objected to reification of meaningful predicates to abstract denotata (classes, properties) that are allowed in the range of quantifiers and thus acquire ontological rights. Indeed, there is probably no uniform attitude that one should have towards questions of reification or hypostasis. Agreeing with Weyl that the denotations of definable predicates can freely be quantified over does not commit one to allowing fictional objects in one's ontology, for instance, or to taking possible worlds to exist in the same way that the actual world exists.[12]

I have argued in chapter 8 that at least some forms of theoretical commitment are connected to a product of *reflection*. But also the champions of conceptual commitment stress the relation with a process of reflection. Lorenzen calls commitment to fragments of predicative analysis the product of what he calls *logical reflection*, which he explains as follows ([**Lor58**, p. 244]):

> [...] functions and relations are not the objects of arithmetic. They are the concepts used in speaking about numbers as the proper objects. Now the transition from arithmetic to analysis is achieved by taking as the objects of a new theory just these concepts of the old theory. Psychologically expressed, the focus of attention has to pass from the old objects, the numbers, to the functions and relations as new objects. Let me call this transition a "logical reflection", because the reflection to be performed is on the concepts occurring in the theorems of the old theory. Or more briefly, the object of the reflection is the language used so

[11] See for instance the *title* of [**Kre70**].

[12] The philosophical question *what we do when we extend our ontology* is, in my view, underexplored.

> far. In view of this one may be justified in calling it a "logical" reflection.

Kreisel agrees with him in the beginning of an article that we have briefly discussed earlier,[13] but of which we quote a slightly larger passage here [**Kre70**, p. 489]:

> *What principles of proof do we recognise as valid once we have understood (or, as one sometimes says, 'accepted') certain concepts?*
> The process of recognising the validity of such principles (including principles for *defining* new concepts, that is, formally, of extending a given language) is here conceived as a *process of reflection*, reflecting on the given concept; reflecting on this process of reflection, and soforth.

One might try to *reduce* commitment to proof theoretic reflection principles to conceptual commitment. The basic claim would then be that commitment to proof theoretical reflection principles springs from reflection *on the concept of absolute provability*.[14] The starting point would then be that a mathematician accepts a formal theory S as proving absolute proofs, i.e., that the derivations that S produces count as proofs in the informal, absolute sense of the word.[15] Moreover, reflection on the concept of (informal) proof reveals that it follows from the content of the concept—i.e., it is *analytic* of the concept—that proof entails truth. Thus provability in S entails truth, i.e., $GRF(S)$ holds.

This line of reasoning assumes that our mathematician starts by accepting S *as provable*. The account in chapter 8 of implicit commitments to reflection principles, however, is supposed to be compatible with scepticism, on the part of our mathematician, about extra-mathematical concepts such as informal provability. Thus it is not clear to me that *all* forms of reflection on theory acceptance can be reduced to reflection on concepts.

However this may be, the idea of implicit commitment of concept probably has considerable unrealised philosophical and proof theoretic potential. We find programmatic thoughts about conceptual commitment in Gödel's *Philosophical Notebooks* from roughly 1939 until 1941.

Gödel takes elements of Kant's epistemology as a point of departure.[16] Famously, Kant argues that objects can only ever be cognitively given to us as subsumed under categories of the understanding. Our knowledge of concepts and ideas, on the other hand, is in his view obtained not by subsuming an entity under categories but in a fundamentally different way. We will not go into the details of Kant's complicated account of knowledge of concepts here. But I want to draw the reader's attention to the fact that, in Kant's philosophy, the three *transcendental ideas* (the subject, the world as a whole, God) occupy a very special place among the concepts. There is a sense in which the transcendental ideas, like objects in themselves, cannot adequately be grasped. The *antinomies of reason* show that when we reason with them, we inevitably become entangled in contradictions.

---

[13] See p. 185.

[14] Some such claim appears to be implicit in [**Myh60**].

[15] The informal, absolute sense of mathematical proof was discussed in Section 1.5.2.

[16] Strangely, Gödel rarely discusses Kant's views explicitly, or mentions Kant by name, in his *Philosophical Notebooks*.

According to Gödel there is a fairly tight analogy between knowledge of *objects* on the one hand, and knowledge of *ideas* on the one hand [**Göd19**, p. 212]:

> Remark: The Kantian view that cognition consists of conceptualizing under the sensory data according to a scheme of ideas that is given a priori should probably be extended to all ideas [...]

Gödel believes that just as objects in themselves are unknowable, not only Kant's transcendental ideas, but all *fundamental ideas* (including the idea of truth, the idea of collection, and the idea of existence) lead us to contradictions when we reason with them. To this, Gödel added the remarkably prescient thought that these unknowable ideas give rise to schematic ways of generating hierarchies of perfectly coherent *concepts* that approximate the unknowable ideas, without ever being able fully to capture them [**Göd21**, p. 234]:

> Remark {Foundations}: What we grasp immediately are not concepts but "concept schemes" (set, $\exists$, etc.), which we can use to define a non-surveyable number of concepts (by "self-application").

Let us try to make the ideas behind this Remark more concrete by focussing on typefree truth. Then a first thought is that a full and coherent grasp of the idea of typefree truth cannot be had. The reason is that reasoning with it in straightforward ways leads us to incoherence. The reasoning of the liar paradox is of course a stark example of this. But more subtle support of this thought is constituted by the way in which the natural system KF is self-undermining. Nonetheless, the incoherent typefree idea of truth can be "unwound" in hierarchy of perfectly coherent and ever stronger typed (Tarskian) truth concepts. This is shown by the way in which every extension $\mathcal{E}$ of the typefree truth predicate at the α-th stage (for $\alpha < \omega_1^{CK}$) can be systematically translated into a collection $\mathcal{S}$ of sentences of a typed Tarskian language $\mathcal{L}_\alpha$ containing a sequence of length α of ever stronger typed truth predicates such that every sentence of $\mathcal{S}$ is true on its intended interpretation.[17] This is in line with the following remark that Gödel made in the context of a discussion of the paradoxes: "type theory (or the true type-free logic) would [...] have to be regarded as a successive approximation of 'truth' " [**Göd21**, p. 281].

We have seen earlier how Feferman defined, beside the theory KF, the *schematic* theory KF(P), which includes the schematic substitution rule Sub.[18] We also saw that the mathematical strength of KF(P) is greater than the mathematical strength of KF (Theorem 5.28, Theorem 5.31). We will now informally sketch how the addition of the rule Sub boosts the mathematical strength of KF, and how this relates to Gödel's views on the unfolding of fundamental ideas through self-application, as applied to truth.

We have seen how typically, for a theory containing a truth predicate T, the amount of transfinite induction that it can prove for the underlying language $\mathcal{L}_{PA}$ is greater than the amount of transfinite induction that it can prove for the whole language $\mathcal{L}_T$.[19] This phenomenon then of course also holds if the background

[17] For the details, see [**Hal97**, Section 5].

[18] See p. 141.

[19] See Theorem 5.20 and Theorem 5.29.

language is taken to be $\mathcal{L}_P$. Recall the *schematic* version KF(P) of the Kripke-Feferman theory KF.[20] Now suppose we have the truth theory $KF(P)^-$, which is just like KF, except that its background language is $\mathcal{L}_P$. So $KF(P)^-$ does not contain the new rule Sub. Then $KF(P)^-$ can prove a certain amount of transfinite induction TI(P) for the underlying language $\mathcal{L}_P$ and no more, which is more than the amount of transfinite induction TI^- that it can prove for the entire language $\mathcal{L}_{P,T}$. But now suppose we add the rule Sub to $KF(P)^-$, yielding the schematic theory KF(P). The substitution rule allows us in KF(P) to derive from TI(P) the formula TI(B) for any $B \in \mathcal{L}_{P,T}$. In other words, we have boosted the amount of provable transfinite induction for the whole language $\mathcal{L}_{P,T}$ from TI^- to TI *through self-application.* From TI for the whole language $\mathcal{L}_{P,T}$, a larger amount of transfinite induction TI^* can then be proved for the underlying language $\mathcal{L}_P$. By the substitution rule, this amount of transfinite induction can then again be lifted to the whole language, and so on, until a closure point is reached at the ordinal Γ_0.[21]

This can be seen as a way in which stronger theories can be reached by "schematically unwinding" a typefree truth concept. Clearly, the above barely scratches the surface of an important development in the mathematics and philosophy of implicit commitment, which has led to the work on the 'unfolding' of mathematical concepts,[22] which I regrettably cannot pursue further here.

9.3. From truth to reflection

Kreisel speculated about another way of reducing set theoretic incompleteness. We might add to the principles of set theory axioms concerning a new primitive concept [**Kre69**, p. 100]:

> Let us try to expand the language of set theory, that is add symbols for new primitive notions, and look for axioms in the wider language which are evident (for the notions given). They may imply set theoretic popositions, i.e., assertions in the language [...] of set theory, which are not.

In particular, Kreisel thought that *randomness* might be a concept that one might try this with.

Kreisel's suggestion has not (yet) led to axiomatic extensions of the standard axioms of set theory that are as strong as or stronger than extensions by powerful new axioms that only contain set-theoretic concepts. Nonetheless, it was clear already to (again) Gödel that reasonable theories of *truth* generate non-conservativeness for the language of set theory, when they are added to the standard axioms of set theory.

The concept of truth for a given domain of discourse—*arithmetical* truth, for instance—can be seen as a *limited* use of a higher type, since it can be used to *simulate*, in a restricted sense, quantification over sets of natural numbers.[23] For instance, we can express the proposition that every *definable* set of numbers has

[20] See p. 141.
[21] See Theorem 5.31.
[22] For a good introduction into this topic, see [**Str18**].
[23] See [**Hor11**, Section 10.3.1].

some given property Φ in a first order manner, as follows:

$$\forall x(x \text{ is a formula of } \mathcal{L}_{PA} \text{ with one free variable} \rightarrow T(\Phi(x)).$$

Moreover, an evidential claim could be suggested that is analogous to Kreisel's evidential claim for higher types.[24] For instance, starting from standard first-order Peano arithmetic, one might claim that our warrant for each instance of an axiom scheme derives from our evidence for the proposition that all instances of the scheme are *true*. So, in particular, our warrant for any instance of mathematical induction derives from our warrant for believing the statement that all instances of mathematical induction are *true*.[25]

In particular, this would mean that if you believe a mathematical theory S, your *evidence* for Con(S), Rfn(S) and RFN(S), derives from your evidence for GRF(S). For Con(PA), this would amount to the following [**Myh60**, p. 462]:

> The proof [of the consistency of PA which is not formalisable in PA] is as follows: The axioms of elementary arithmetic are true, and the rules of inference are truth-preserving. Therefore every theorem of elementary arithmetic is true. Therefore '0 = 1' is not a theorem of elementary arithmetic. Therefore a certain statement p (the arithmetization of the statement that '0 = 1' is not a theorem) is true.

In an article from around the same time, Dummett sketches the argument in slightly more detail [**Dum63**, p. 195]:

> By hypothesis the axioms of [PA] are intuitively recognized as being true, and the rules of inference of [PA] as being correct in the sense of leading from true premisses to true conclusions. Hence we may establish by an inductive argument on the lengths of formal proofs that each proof in [PA] has a true conclusion, and by another inductive argument on the number of logical constants in a statement that no statement is both true and false; concluding from this that [PA] is consistent.

This argument, and variations on it, are known as the *semantic argument* for reflection principles.

The strategy of the semantic argument is *generally* applicable: it works not only for arithmetical theories of finite order, but also for our most encompassing mathematical theories. The strategy presupposes, of course, that you accept the restricted Tarski-biconditionals for your concept of mathematical truth. But this is reasonable, since proving restricted Tarski-biconditionals is an adequacy condition for theories of truth. In other words, the idea is that the derivation of Con(S), Rfn(S) and RFN(S) from GRF(S) tracks the epistemic grounds of your belief in Con(S), Rfn(S) and RFN(S).

Some philosophers of mathematics of the post-Myhill-Kreisel-Dummett generation are sympathetic to the idea that proof theoretic reflection principles can be systematically *and generally* proved using the concept of truth. Shapiro, for instance, argues that the concept of truth plays an important role in a good explanations of *why* the Gödel sentence for PA is true [**Sha98**, p. 505]:

---

[24] See p. 220.

[25] This claim will be discussed below in section 9.4.

> Go back to our theory of arithmetic S and its Gödel sentence G_S. Suppose that a logic teacher asserts that G_S is true, and a puzzled student asks for an explanation. The student accepts the teacher's word that GS is true, but she wants to be shown why it is true. The student wants something like a convincing proof or an explanatory proof. The natural reply is to point out that all of the axioms of S are true and the rules of inference preserve truth. Thus every theorem of S is true. It follows that $0 = 1$ is not a theorem, and so S is consistent. The Gödel sentence is equivalent to the consistency of S. It seems to me that this informal version of the derivation of Con(S) and G_S is as good an explanation as there is.

Again, this explanation works not only for the Gödel sentence for arithmetical theories such as PA, but also, for instance, for the Gödel sentence for ZFC. Observe also that this strategy requires taking truth to be a *primitive notion*. If arithmetical truth would be *defined* using second-order quantification, for instance, then the strategy would collapse into Gödel's appeal to higher types, which was discussed in section 9.1, and would therefore be faced with the same limitation as the higher types approach.

Gödel believed that extensions of ZFC based on strong principles of infinity are stronger than the result of extending ZFC by truth axioms [**G̈46**, p. 151]:

> Any proof of a set-theoretic theorem in the next higher system above set theory (i.e. any proof involving the concept of truth [...]) is replaceable by a proof from [...] an axiom of infinity.

At first sight, this passage is a bit puzzling. Earlier we took "moving to a higher type" to consist in adding the natural proof principles for the next higher level of quantification. But here Gödel seems to identify moving to a higher type with adding natural axioms for the notion of truth.

However, the two conceptions of "higher type" are compatible: an extension of a first-order theory S with axioms governing a truth predicate can be seen as an extension of S with a somewhat *restricted* form of second-order quantification. We have seen in the beginning of this section how the truth predicate can be used to quantify over *definable classes*. For arithmetic, this means that, against the background PA, the typed compositional truth theory CT is intertranslatable with the predicative fragment ACA of second-order arithmetic. And against the background of ZFC, the compositional truth theory CT corresponds to ECA.[26] The latter theory is first-order non-conservative over ZFC, and therefore strictly stronger than NBG; but it is much weaker than full ZFC^2.

Tarski took proving typed disquotational principles to be an adequacy condition for theories of truth. Some authors go further and claim that proving proof theoretic reflection principles is an *adequacy condition* for truth theories [**Ket99**, p. 90]:

> Parts of the basic (not necessarily deflationist) idea about truth is that a particular statement φ and its "truth" $Tr(<\varphi>)$ are somehow "equivalent". I think this is correct (indexicals aside), and if a truth theory satisfies *Convention T* then it proves the equivalence. But we must go further. Any *adequate* theory of

[26] See p. 135.

> truth should be able to prove the "equivalence" of a (possibly infinitely axiomatised) *theory* T and its "truth" $True(T)$ (that is, the metalinguistic formula $\forall x(Prov(x) \to True(x))$).

We have seen that natural typed disquotational truth theories are conservative.[27] So they fail to prove global reflection principles for their background theories, and hence fail to meet Ketland's adequacy condition. The most natural axiomatic truth theories that do prove reflection principles for their background theory are *compositional* truth theories such as CT or KF.[28]

Given Theorem 5.17,[29] this presupposes that mathematical induction is treated in an *open-ended* way, so that the truth predicate is allowed to occur in the induction scheme. Indeed, some philosophers argue that when we *learn* the principle of mathematical induction, we directly acquire it as an open-ended principle. McGee, for instance, writes [**McG97**, p. 58]:

> Our understanding of the language of arithmetic is such that we anticipate that the Induction Axiom Schema, like the laws of logic, will persist through [changes in language]. There is no single set of first-order axioms that fully express what we learn about the meaning of arithmetical notation when we learn the Induction Axiom Schema, since we are always capable of generating new Induction Axioms by expanding the language.

This does not imply that if you treat mathematical induction in an open-ended way, you must accept *every* predicate in the induction scheme. For instance, you might still refuse to allow the predicate "large natural number" in the induction scheme, for fear that then a sorites argument will allow you to prove that all natural numbers are small. But if you encounter a predicate that is not in any way semantically deficient—let us call such predicates *determinate*—then you will allow reasoning by mathematical induction for formulas that contain that predicate.

If one accepts Ketland's adequacy condition, then one can conclude that Horwich's minimal (and disquotational) truth theory is inadequate. Moreover, it has been argued that truth theoretic deflationism *in general* is committed to proof theoretic conservativity.[30] If that is right, then the vaguer thesis that truth is an "insubstantial" concept is also flawed. However, truth theoretic deflationism is a somewhat nebulous doctrine, and it is not at all immediately obvious that it should be wedded to the thesis that truth theories should be conservative over their background theory. Indeed, attempts have been made to argue for conceptions of truth according to which truth is an "insubstantial" notion, but that they are nonetheless well captured by non-conservative axiomatic truth theories.[31]

9.4. Scepticism again

Myhill, Shapiro, and Ketland are aiming for a *general* way of *proving* proof theoretic principles for accepted mathematical theories. Ketland, for instance, explicitly writes [**Ket05**, p. 85]:

---

[27] See Section 5.1.2, Theorem 5.2.

[28] See Section 5.2.

[29] See p. 135.

[30] This line of argument was developed in [**Hor95**].

[31] See for instance [**Hor09**].

> Part of the point of the articles by Feferman, Shapiro and myself was to show how to *prove* reflection principles [...]

As an answer to this challenge, they settle on the proof of theorem 5.14.[32] Let us call this *the semantic argument* (for S).[33] According to Myhill, Shapiro, and Ketland, the semantic argument for PA constitutes a perfectly good *justification* for a belief in, for instance, the consistency of PA.

Dummett, Wright, Girard, and Dean, in contrast, deem the semantic argument unfit for justifying belief in the consistency of PA.[34] Let us take a brief look at their arguments.

Dummett starts by considering a statement that is conceptually and historically closely connected to the consistency statement for PA, namely, the *Gödel sentence* G_{PA} for PA.[35] The argument goes along the following lines:

> *If, for a contradiction, $\neg G_{PA}$, then there would be a PA-proof that G_{PA}, and therefore also that G_{PA} is PA-provable. On the other hand, if there is a PA-proof that G_{PA}, then there is also a PA-proof that G_{PA} is not provable, for G_{PA} says of itself that it is not PA-provable. So PA would be inconsistent, which of course it isn't. Therefore we must reject the assumption.*

But then, Dummett continues, in order for this argument to have justifying force, we need warrant for the assertion Con(PA) [**Dum63**, p. 193–194]. Here Myhill and consorts would appeal to the semantic argument for Consis(PA), which, as we have seen on p. 227, Dummett indeed also cites at this juncture. But, in contrast with Myhill and friends, Dummett expresses a vague epistemic dissatisfaction with the semantic argument: [**Dum63**, p. 194]:

> Such a general form of consistency proof cannot, of course, be expected to be genuinely informative; it can only be the trivial kind of proof by induction on the length of formal proofs with respect to the property of having a true conclusion.

It is facile to be *blasé* about theorems that are old and well-known and relatively simple. I dispute that the semantic argument is *trivial*. Already the reasonably detailed version of the proof that CT entails GRF(PA) takes up almost four pages in Halbach's standard textbook of axiomatic truth theory [**Hal11**, p. 102–106]. Suppose you were to teach Gödel's first incompleteness proof in detail in a graduate logic course, without talking about arithmetical truth. Then, on the exam, you give the students the typed compositional truth axioms, and you ask them to prove in reasonable detail that when these truth axioms are added to PA and the induction scheme is extended to allow for predicates that contain the notion of arithmetical truth, the consistency of PA logically follows. Then there is a chance that the head of your department calls you into her office the next day and rebukes you for setting an exam that is too difficult. Indeed, I submit that, were someone to have submitted a detailed proof of the fact that CT implies the consistency of PA for publication to the Journal of Symbolic Logic in 1937, the article might have been accepted for publication.

---

[32] See Section 5.2.1.

[33] It is called *the inductive argument* by Dean: see [**Dea15**, p. 54].

[34] See [**Dum63**], [**Wri94**], [**Gir87**], and [**Dea15**, Section 5].

[35] Indeed, it is not hard to see that, provably in PA, the statements G_{PA} and Con(PA) are equivalent with each other.

More important, however, is the question whether the semantic argument has any epistemic merit. Dummett does not elaborate on his dim view of the epistemic merits of the semantic argument. Wright aims to do better by elaborating on Dummett's cryptic sceptical remark. He agrees with it, and provides an account of *why* the semantic argument is uninformative. In his account, Wright makes use of the notion of *suasiveness* [**Wri94**, p. 177–178]:

> [...] an intellectual routine counts as a demonstration of P just in case an agnostic about P could nevertheless perfectly reasonably place confidence in the methods and principles deployed in the routine, and could arrive, on the basis of following it through, at considerations which would rationally oblige him a priori to assent to P. Say that a proof is *suasive* if it meets those conditions.

Using this notion, Wright then claims that it is obvious that the semantic argument fails to be suasive [**Wri94**, p. 178]:

> [...] since it takes the truth of the axioms and the soundness of the underlying logic as a premise, the sort of consistency proof Dummett has in mind can hardly be suasive — (except perhaps for a rather dim thinker to whom it has not occurred that you cannot get contradictions in a system with true axioms and truth-preserving rules.) So it furnishes no demonstration of $[G_{PA}]$ [...]

I do not think that it is obvious at all that the semantic argument cannot be suasive. Consider the following scenario. Emma is a mathematician who (dispositionally) believes every axiom of PA. She has considered the question whether PA is consistent, and is agnostic about it. On the one hand, she feels that she cannot rule out the possibility that there exists a derivation in PA of an inconsistency. On the other hand, she is convinced that there is no such "feasible" proof, i.e., that every such proof would have to be far longer than is derivable by any human or by any future machine.[36] After overcoming her prejudice against philosophical notions, Emma comes to accept and master the concept of arithmetical truth. She comes to believe the compositional truth axioms (for arithmetical statements), and she extends her mathematical induction scheme to cover also predicates in which the concept of arithmetical truth occurs. Then she carries out the proof of the consistency of PA in CT. This fundamentally changes Emmas doxastic state concerning arithmetic: it *persuades* her that PA is consistent.

According to Wright, this scenario would have to be obviously incoherent. Indeed, as we have seen in section 9.1, he claims that it is "hardly practical" as a mathematician to stay on the fence about the consistency of PA [**Wri94**, p. 191]. But I don't see what is problematic from a practical point of view about Emma's doxastic situation before she is down with the notion of arithmetical truth. She has no practical worries, for she is convinced that any proof of an inconsistency must be astronomically long and will therefore never be obtained. She proves arithmetical theorems like nobody's business, and that is, at least as far as her colleagues are concerned, all that matters for her as a mathematician.

[36] On the concept of "long inconsistency proofs", see [**Woo98**].

Girard gives another argument for the thesis that the semantic argument cannot "suade" [**Gir87**, p. 64]:

> [A]ll theorems of PA are true, and justification lies in the stupid [*sic*] remark that the axioms of PA are true, and the rules of the predicate calculus preserve truth. Of course, the epistemological value of this result is limited, because we have chosen the axioms of PA precisely because we believe in their truth [...]

So the idea is that we believe some of the premises of the semantic argument *only because* we *already* believe in the consistency of PA. Thus the epistemic defectiveness of the semantic argument is, according to Girard, due to the fact that it is implicitly *viciously circular*.

But an antecedent belief in the truth of the axioms of PA is not why we chose them as basic principles for our arithmetical reasoning. Rather, we selected them as axioms in direct response to our mathematical experience[37]—which is not to say that this "choosing" was in any way a straightforward or simple process! Indeed, as argued in Section 8.2.2, as mathematicians, some of us might harbour deep suspicions towards philosophical concepts such as arithmetical truth. For these reasons, I find Girard's argument without merit. Moreover, for the same reasons, I am not convinced of the Kreiselian hypotheses, discussed earlier,[38] that we only accept a schema *because* we believe the statement that all instances of the scheme are true, or because we believe a higher-order universal statement from which all instances of the scheme follow.

Like Girard, Dean also expresses scepticism about "whether any useful epistemic work is achieved by [the semantic argument] beyond the promissory character of its informal counterpart" [**Dea15**, p. 61]. The main epistemic weakness of the semantic argument, in Dean's view, derives from the fact that mathematical induction in the extended language containing the concept of truth is used in the proof of GRF(PA) [**Dea15**, p. 59]:

> [...] although most of us believe that all instances of first-order induction are true, our rationale for doing so presumably cannot rest on the derivability of a statement like
>
> $$[\forall y(y \text{ is an arithmetical formula} \rightarrow T(Ind(y))]^{39}$$
>
> in a formal theory of truth such as CT. For although this formula can be interpreted as expressing that all instances of [the mathematical induction scheme for the language of PA] are true, its proof relies on an application of induction for a formula in the richer language $\mathcal{L}_T$ which presumably is no more evident than [the mathematical induction scheme for the language of PA] itself.

Here again the objection appears to be Kreiselian in the sense discussed above: we believe the few instances of mathematical induction for $\mathcal{L}_T$ that we need in our proof of GRF(PA) *only because* we *already* believe GRF(PA).[40] Again it is unclear

[37] See the discussion in Section 7.5.

[38] See Section 9.1.

[39] $Ind(y)$ denotes the instance of mathematical induction for the formula y.

[40] Indeed, Dean surmises that the point that he is making is similar to the point that is made in the passage by Girard quoted above [**Dea15**, p. 59].

wherein the evidence for this contention consists. It is perfectly conceivable that some of us form, in *direct* response to our mathematical experience, a dispositional acceptance of mathematical induction for *any* predicate that we regard as determinate. They then can go on to prove GRF(PA), relying on open-ended mathematical induction *and the compositional truth axioms*. In so doing, they obtain justification for believing in GRF(PA).

In the above we assumed that *before* going through a truth-theoretic argument for Con(PA), the reasoner fully believes PA, while being neutral about Con(PA)—perhaps because she never even *considered* the statement Con(PA). In such circumstances, we have argued, a semantic argument can convince the reasoner of Con(PA). If, before going through the semantic argument, the reasoner does *not* fully believe PA, then she may not obtain full confidence in Con(PA) by going through a standard semantic argument.

In general, a consistency argument for a theory T carried out in an extension of a *fragment* of T can be persuasive even for a reasoner who does not *fully* trust T. Consider Eva, who at the outset is worried about the principle of excluded third, but has full confidence not only in the other principles of classical logic, but also in the arithmetical axioms of PA. In this situation, she does not have full confidence in PA. Presently she convinces herself also of transfinite induction up to small countable ordinals including ε_0. She then carries out Gentzen's proof of the consistency of PA in Heyting Arithmetic plus transfinite induction up to ε_0, and thereby convinces herself of at least the consistency of classical PA.[41]

The mere finiteness of formal proofs entails that in the ordinary semantic argument for the consistency of PA, beside compositional truth axioms only a fragment of PA (in the extended language) is involved. A closer inspection of standard semantic arguments for reflection principles for PA shows that they can straightforwardly be carried out in the compositional theory over Elementary Arithmetic rather than full PA. A mathematician may therefore at the outset be sceptical about mathematical induction for complicated formulas and only fully trust a weak theory of arithmetic that contains no more than Σ_0 induction. In sum, the reasoner need not, it seems, before embarking on a semantic argument, trust all of PA.

But in the case under consideration, this appearance is illusory. The aforementioned close inspection of the semantic argument works against the background of Elementary Arithmetic *in the extended language* only because, given the Tarski-biconditionals for arithmetical formulas,[42] every complicated arithmetical formula $\varphi(x)$ is equivalent to a Δ_0 formula $T\varphi(x)$. In other words, the Tarski-biconditionals allow us to 'hide' the complexity of arithmetical formulas. So if one distrusts mathematical induction for complicated purely arithmetical formulas, then one should likewise distrust mathematical induction for quantifier-free formulas in the extended language $\mathcal{L}_T$.

9.5. New conceptual resources

According to the semantic argument, reflection principles for a theory S that we are warranted to believe, are justified by deriving them from principles that go

[41]This does not exclude that she *could* also have obtained her confidence by other means, for instance by the double negation interpretation of PA into Heyting arithmetic, if she at the outset explicitly believed that Heyting Arithmetic is consistent.

[42]See Proposition 5.13.

beyond what S can prove. This raises the question: wherein consists our warrant for believing these new principles?

Several answers to these questions are possible. One might argue that we are somehow *entitled* to accept these new principles. Or one might argue that we have *justification* for accepting them. Here we focus on the claim that we are *implicitly committed* to accepting the new principles from which reflection principles are derived.

In Chapter 7, several proposals for systematically deriving reflection principles were already discussed.[43] The important difference with the semantic argument lies in the fact that the latter involves a *new concept*: truth. As we have seen, in this respect, it rather resembles Gödel's proposal of proving reflection principles by "moving to higher types".[44] For instance, one can prove the consistency of PA by appealing to *second-order* induction and a *second-order* comprehension principle. So beside the concept of natural number that is described by PA, we are now helping ourselves also to a new concept: the concept of *set of* natural numbers.

In a recent article, Nicolai and Piazza argue that the implicit commitment of mathematical theories is weaker than it is commonly thought to be. They focus on *arithmetical* theories. We will do likewise in the remainder of this section. But, as always, it is worth bearing in mind that much of the discussion below also applies to stronger starting theories.

Nicolai and Piazza find that recent philosophical work on systematically deriving reflection principles shows "how hard it is to eradicate the intuition that reflection principles are conceptually dependent on the notion of truth" [**NP19**, p. 923]. They believe that when we accept an arithmetical theory, we are *implicitly committed to the concept of truth* [**NP19**, p. 931–932]:

> The notion of truth is integral to any reasonable articulation of what we are implicitly committed to when accepting a given arithmetical theory.

With this, they step into Feferman's footsteps [**Fef91**, p. 2]:

> Which statements in the base language L of S [...] ought to be accepted if one has accepted the basic axioms and rules of S? The answer is given as an ordinary theory $Ref(S)$ formulated in a language $L(T, F)$ [...] where T and F are partial truth and falsity predicates which are self-applicable in the sense that they apply to (codes of) statements of $L(T, F)$ [...] Thus, for example, we may reason in $[KF]$ by induction about the truth of statements which contain the notion of truth, and so arrive at statements of the form: $\forall x[\mathrm{Bew}_{PA}(x) \to T(x)]$, and by repeating this kind of argument derive iterated reflection principles for arithmetic.

Earlier, in Section 7.3, we discussed Dean's claim that theories of implicit commitment must respect the epistemic stability of certain foundationally significant mathematical theories, such as S^1_2, PRA, and PA. Nicolai and Piazza accept Dean's claim [**NP19**, p. 929]:

---

[43] See Sections 7.4.2 and 7.4.3.

[44] See Section 9.1.

> In other words, we do not claim that, say, Con(S) is not a natural principle to endorse once one has endorsed S, but what we share with Dean is the view that *if* the justification of Con(S) is equivalent to principles that are incompatible with the alleged epistemic stability of S, [...] then such a justification cannot be implicit in the mere acceptance of S but should stem from more general considerations.

In Section 7.3 it was argued that, for instance, a Hilbertian finitist is indeed not committed to uniform reflection over PRA, since this requires an awareness the boundaries of her own *explicit* commitments, which she does not have. But this time we cannot satisfy Dean's requirement so easily: adding compositional truth axioms and extending the induction scheme does not require an awareness of the boundaries of the explicit commitments of the position.

In the face of this challenge, Nicolai and Piazza propose to *weaken* the set of new principles from which reflection principles can be derived. They aim to make them truth theoretically as strong as possible without resulting in non-conservativeness. Specifically, they argue that what they call the *semantic core* of the implicit commitments of a theory S—which we shall abbreviate as SC(S)—consists of $CT^{-}(S)$ plus the claim "All axioms of S are true" [**NP19**, p. 928]. It is not claimed that SC(S) exhausts the implicit commitment of S in all cases: "[SC(S)] counts only as a class of necessary conditions that our notion of truth has to satisfy" [**NP19**, p. 933]. Thus Nicolai and Piazza allow that when the foundational position that S intends to capture does not preclude mathematical principles exceeding S being warranted, the implicit commitments of S may exceed SC(S).

At least for mathematical theories S that exceed Elementary Arithmetic, the semantic core of S has the following properties:

(1) SC(S) is arithmetically conservative over S [**Lei15**, Theorem 2];
(2) SC(S) proves "All propositional tautologies are true" [**NP19**, p. 932];
(3) SC(S) proves that the rules of inference are truth-preserving [**NP19**, p. 932].

But then of course SC(S) will not be able to prove that all *theorems* of S are true.

Shapiro and Ketland would probably not be satisfied with this proposal. As we have seen, they take it a task of truth theory to provide justification for *strong* reflection principles, such as GRP, and not only for weak reflection principles such as (2). Nicolai and Piazza's view is not in conflict with Shapiro and Ketland's *dictum*. They would concede that the minimal implicit commitment of a theory—its semantic core—does not provide this. But they leave open the possibility that more robust truth theories do provide such justification. It is just that to some foundational points of view, such stronger truth theories are unacceptable.

Nicolai and Piazza only consider *arithmetical* starting theories. But suppose that your starting theory is PA with *open-ended* induction. Moreover, suppose that you also regard truth as a determinate predicate, suitable for appearing in the induction scheme. Suppose you are a disquotationalist, and take this truth predicate to be governed by the restricted Tarski-biconditionals. Call your theory PA^*. Then

$$SC(PA^*) = CT(PA),$$

which is not arithmetically conservative over PA. (So for *some* theories S, the semantic core of S is not mathematically conservative over S.) So PA^* must be regarded

as an unstable theory. But it is hard to see what is epistemically unstable about it. So I regard this as an unwelcome consequence of Nicolai and Piazza's theory. The reason why they are committed to it, is that they buy into the strong "ought" of Feferman's conception of implicit commitment: they work with a "Prussian" concept of rationality.

There is also, I believe, a *lacuna* in Nicolai and Piazza's proposal. They hold that when you accept a mathematical theory, you implicitly commit yourself to certain properties of a new concept: truth. But *how* is this new concept implicit in the acceptance of a mathematical theory?—the epistemological story is on this point wholly missing in Nicolai and Piazza's theory. In Section 8.2.2, I described a mathematician who is sceptical about the notion of truth. In Nicolai and Piazza's view, such a person would be implicitly incoherent. But exactly how is that so?

9.6. From reflection to truth

In the previous Sections, we have discussed how truth principles can contribute to providing a warrant for believing proof theoretic reflection principles. In this Section, the epistemic direction is reversed: we explore whether and how reflection principles can contribute to providing warrant for truth theories.

Clearly, if you start out with a purely mathematical theory, and add proof theoretic reflection principles to it, you will not thereby be able to derive (non-tautological) principles concerning a new concept such as truth. So in this Section I will assume that you are already at the outset warranted in believing certain weak truth principles: a class of restricted Tarski-biconditionals, perhaps. In Chapter 5 we saw that by repeatedly adding reflection principles to such a starting theory, a stronger truth theory is obtained. Moreover, we saw that this is so both in a typed and in a typefree context, and both in the context of classical and of partial logic. In this Section, we examine the epistemological significance of these phenomena.

9.6.1. Horwich. In Section 5.3.2, we discussed Horwich's views on truth. We have seen how he argues that some *disquotational theory* is our best theory of (typed) truth. Let us call this theory MT. Moreover, he argues that all we ever need to know about the concept of truth follows from the disquotational axioms of MT. We call this combined logico-philosophical package the *minimalist conception* of truth.

We have raised the question whether the minimalist conception can deal with the *truth generalisation problem*. I.e., we have asked how truth generalisations such as

$$\forall \varphi \in \mathcal{L}_{PA} : T(\varphi \to \varphi)$$

follow from the minimalist conception.

The vagueness of Horwich's conception of truth makes it not easy to see how this question should be answered. Horwich's conception of truth is indeed rather vague because he does not describe his *theory* of truth MT with mathematical precision. Several aspects of MT are unclear. First, we have seen that Horwich takes *propositions* as truth bearers. So in order to obtain a precise truth theory, the disquotational axioms of MT have to be supplemented with a precise theory of propositions. The history of philosophy of language has taught us that this is a daunting task. Secondly, Horwich does not describe the collection of disquotational axioms of MT with mathematical precision.

In response to these difficulties, we will make MT more precise. We stick to the approach that we have taken throughout this book, namely to sidestep the philosophical problems connected to propositions by taking sentences to be the bearers of truth and falsehood. Moreover, for definiteness, we assume that MT is the disquotational theory TB. And we know, of course by Theorem 5.2, that TB does not prove interesting truth-generalisations.

We will now discuss how Horwich gradually came to see a need for strengthening his disquotational conception with a reflection rule in response to the generalisation problem.[45]

In the *Postscript* of the revised edition of his book *Truth*, Horwich formulates his **first response** to the truth generalisation problem. There he writes [**Hor98**, p. 137–138]:

> However, it seems to me that in the present case, where the topic is *propositions*, we can find a solution to this problem. For it is plausible to suppose that there is a truth-preserving rule of inference that will take us from a set of premises attributing to each proposition some property, F, to the conclusion that all propositions have F. No doubt this rule is not *logically* valid, for its reliability hinges not merely on the meanings of the logical constants, but also on the nature of propositions. But it is a principle we do find plausible. We commit ourselves to it, implicitly, in moving from the disposition to accept any proposition of the form 'x is F' (where x is a proposition) to the conclusion 'All propositions are F'. So we can suppose that this rule is what sustains the explanations of the generalizations about truth with which we are concerned. Thus we can, after all, defend the thesis that the basic theory of truth consists in some subset of the instances of the equivalence schema.

This truth-preserving rule amounts to a form of Hilbert's ω-rule [**Raa05**, p.175], which we have encountered before.[46]

Theorem 6.4 tells us that over PA, the ω-rule is strong: it causes *all* true arithmetical sentences to become provable. When added (for all formulas of $\mathcal{L}_T$) to TB, it is even stronger: it then causes all acceptable truth-generallisations for $\mathcal{L}_{PA}$ to become provable.

However, certain features of the ω-rule render this proposal problematic, and in particular unacceptable to the minimalist truth theory. As finite human beings, we cannot take infinitely many premises into consideration simultaneously [**Wan61**, p. 349]:

> There is a temptation to cut through the foundational problems by using the non-constructive rule of induction (the omega-rule) [...] [But] we can never go through infinitely many steps in a calculation or use infinitely many premises in a proof unless we have somehow succeeded in summarizing the infinitely many with a finite schema in an informative way. Both mathematical induction and transfinite induction are principles by which we make

[45] For a fuller discussion of Horwich's reaction to the generalisation problem, see [**HZng**].

[46] See p. 153.

> inferences after we have found by mental experimentations two suitable premises which summarize together the infinitely many premises needed. A very essential purpose of the mathematical activity is to devise methods by which infinity can be handled by a finite intellect. The postulation of an infinite intellect has little positive content except perhaps that it would make the whole mathematical activity unnecessary.

Therefore, even if theory MT plus the ω-rule is capable of proving acceptable truth generalisations, those generalisations are beyond the reach of ordinary human beings [**Raa05**, p. 176].

It is not clear whether Horwich has accepted this objection to his first proposal. In a more recent publication Horwich still seems to propose using the ω-rule as a solution to the truth generalisation problem [**Hor05b**, p.84]:

> For it is plausible to suppose that there is a truth-preserving rule of inference that will take us from a set of premises attributing to each proposition of a certain form some property, G, to the conclusion that the *all* proposition have property G. And this rule – not *logically* valid, but nonetheless necessarily truth-preserving given the nature of proposition – enables the general facts about truth to be explained by their instances.

Yet in *most* of his recent writings, Horwich advocates an alternative resolution, based on an introspective process. To this proposal we now turn.

Over the years, Horwich's formulation of his **second response** has varied, and it is not easy to select a preferred formulation from these variants. Nonetheless, we will see that all variants of Horwich's second proposal need emendation in order to solve the truth generalisation problem.

A first fomulation of Horwich's second attempt emerges in *A Defense of Minimalism* (2001) [**Hor01b**, p.157]:

> Whenever someone can establish, for any *F*, that it is *G*, *and recognizes that he can do this*, then he will conclude that every *F* is *G*.

Call this *Solution 2.0* (with the earlier proposal of adding the ω-rule being *Solution 1)*. The new solution also consists in adding an additional rule of inference to MT, but the additional rule of inference of Solution 2.0 is different from the ω-rule.

In a revised version (2010) of the same paper, Horwich formulates a variant on this new proposal [**Hor10**, p. 45]:

> Whenever someone is disposed to accept, for any proposition of structural type F, that it is G (and to do so for uniform reasons) then he will be disposed to accept that every F-proposition is G.

To the above statement, he adds the following *proviso* [**Hor10**, p. 44–45]:

> We cannot conceive of there being additional Fs – beyond those Fs we are disposed to believe are G – which we would not have the same sort of reason to believe are Gs.

Call this *Solution 2.1.*[47]

Armour-Garb argues that Solution 2.1 is unsatisfactory because [**AG10**, p.699]:

[47]Horwich endorses this same solution in 2005 [**Hor05b**, p. 84].

> [O]ne will not be disposed to accept (the proposition) that all F-propositions are G, from the fact that, for any F-proposition, she is disposed to accept that it is G (NB, even for uniform reasons), unless she is *aware* of the fact that, for any F-proposition, she is disposed to accept that it is G.

The *proviso* that Horwich added to Solution 2.1 does not provide such an awareness component. It merely adds a *negative* condition ("not being able to conceive of there being F's that are not G"), while Armour-Garb's awareness-requirement is a positive condition. In contrast with this, Solution 2.0 incorporates exactly the awareness condition that Armour-Garb insists on ("and recognises that he can do this"). Therefore Horwich's Solution 2.0 must be regarded as superior to his Solution 2.1.

Nevertheless, Armour-Garb would not be satisfied with Solution 2.0, either. He argues that the switch, in the move from the premise to the conclusion of the rule of inference in Solution 2.1, of 'for any F-proposition' from outside the 'disposed to accept'-context to inside the 'disposed to accept'-context, is "viciously circular". He is certainly right that this *quantifier shift*, which is also present in Solution 2.0, is not derivable in classical logic. Nonetheless, I take issue with this aspect of Armour-Garb's critique of Horwich's second proposal. I agree with Cieśliński that Armour-Garb's dismissal of Horwich's second solution on the ground of its being viciously circular is "hasty" [**Cie18**, p.1082]: I will come back to this later.

It is time to spell out the content of Horwich's Solution 2.0 in more precise terms.[48] This is done by formalising Horwich's informally expressed and somewhat vague rule of inference in first-order logic. The aim is to be charitable. I do not claim that Horwich would agree with the proposed formalisation, but I will argue that there are good reasons for him to do so. Firstly, Solution 2.0 contains the phrase '*will* conclude', making it seem like a psychological prediction.[49] If it is taken in this way, then whether it is true or not, is an empirical matter. But this is presumably not what Horwich intends. Rather, what he means, is that the agent will be disposed to drawing this conclusion *if she is rational*. In other words, Horwich purports to propose a *rational rule of inference* here. So it might be better to replace, in Solution 2.0, "will conclude" by "may (rationally) conclude", or perhaps even "should (rationally) conclude". Secondly, since we are concerned with *establishing* truth generalisations, let us identify the concepts 'being disposed to accept' and 'recognising' with being *provable* (Bew). Thirdly, let us identify provability with provability in the background theory, which we have taken to be TB. If we were to identify provability with provability in the system *including the rule*, then the proposed rule would indeed be viciously circular, confirming Armour-Garb's suspicions. But if we identify provability with provability in TB then there is no circularity. Fourthly, let us *omit* the concept of provability ("being disposed to accept") from the conclusion of the rule. With these qualifications in place—I leave it open whether they are in accordance with what Horwich intended—we obtain the following schematic rule:[50]

---

[48] See [**HZng**, Section 4].

[49] Cieśliński criticises this aspect of Horwich's account: see [**Cie18**, p. 1085].

[50] Here H stands for *Horwich*.

$$\text{(H)} \qquad \frac{\text{Bew}_{TB}(\forall x : F(x) \to \text{Bew}_{TB}(G(x))}{\forall x : F(x) \to G(x)}.$$

Observe that, unlike the ω-rule, H is an effective rule: adding it to TB (=MT) yields an axiomatic system. The rule H can be seen as an effective version of the ω-rule.

Worries based on the lottery paradox might cause one to doubt the rationality of rule H. For any ticket (in a large, fair lottery), I believe that it is not the winning ticket (and I believe this for "uniform reasons"). But from this, I am not prepared to infer that every ticket is a losing ticket [**Kyb70**, p.56]. But such a worry would be ill-founded, for the situation under consideration is different in one key respect. The irrationality of the lottery paradox inference stems from the fact that many small but non-zero probabilities (of being the winning ticket) can add up to a large probability (of one of a large collection of tickets being the winning one). But what is provable, has probability 1 rather than $1 - \varepsilon$ for some small ε, since provability in a sound system from necessary premises, is itself necessary, and necessary truths by a Kolmogorov axiom for probability receive probability 1. So the fair lottery phenomenon is irrelevant to the evaluation of rule H.

Addressing a worry is one thing, a positive argument is another. Horwich does not tell us how our acceptance of rule H is justified. Nonetheless, let us shelve this question for now.

Let us denote MT+H as MT_1. Now that we have made Horwich's Solution 2.0 precise, we address the question whether the rule H is sound, and the question whether MT_1 can prove all intuitively acceptable truth generalisations.

The first of these two questions is easy. It is clear that given a *sound* theory S, adding H (with Bew_{MT} replaced by Bew_S) to S, results in a sound system. So, in particular, MT+H is a sound system.

Next, we make the crucial observation that H is equivalent to a reflection rule that has intensively been investigated in proof theory, and that we have already discussed at several places in this book. From Section 6.1, we recall the distinction between reflection principles and reflection rules. However, we also know that they are closely related. By Feferman's little reflection theorem (Theorem 6.6) the uniform reflection principle RFN(S) is equivalent to the uniform reflection rule $RFR(S)$. In the light of this, it is easy to see that [**HZng**, Section 5.1]:

PROPOSITION 9.1. *H is equivalent to $RFR(MT)$, and therefore also to RFN(MT).*

At this point, a connection with Horwich's *first* solution also becomes apparent. Indeed, the uniform reflection rule is widely seen as an effective version (a "tamed" version) of the ω-rule. Horwich's appeal to the ω-rule was rightly rejected by Wang, Raatikainen, and others on account of its non-effectiveness. Uniform reflection rules, on the other hand, cannot be rejected on the same grounds.

It is fairly generally accepted that from the compositional axioms for truth, all intuitively acceptable truth generalisations logically follow.[51] So if Horwich can derive the truth axioms of CT, then he has solved the truth generalisation problem.

[51] See [**Hor11**, chapter 6]. An apparent counterexample is a proposition such as 'there are as many truths as there are untruths' [**Gup93**, p.363]. But this is a second-order statement, involving not just sentences but also *sets* of sentences. So it falls outside the scope of $MT(= TB)$, which cannot even express *claims* involving sets of sentences.

With only one exception, the compositional truth axioms can indeed be derived in MT_1. As an example, let us consider the compositional axiom for negation:

$$\forall x \in \mathcal{L}_{PA} : T(\neg x) \leftrightarrow \neg Tx.$$

We know from Proposition 5.12 that every *instance* of this axiom can be proved in TB (using Tarski-biconditionals). Moreover, PA can formalise the proof that this is the case: PA recognises, as a combinatorial fact, that every instance of the compositionality of negation can be proved in TB. In other words, we have

$$PA \vdash \forall x \in \mathcal{L}_{PA} : Bew_{MT}(T(\neg x) \leftrightarrow \neg Tx).$$

Then by the rule H we indeed obtain $\forall x \in \mathcal{L}_{PA} : T(\neg x) \leftrightarrow \neg Tx$.

The other compositional axioms can be derived in a similar way in MT_1, with the sole exception of the quantifier axiom:

$$\forall \varphi(x) \in \mathcal{L}_{PA} : T(\forall x \varphi(x)) \leftrightarrow \forall x T\varphi(x).$$

We cannot prove in MT, for every $\varphi(x) \in \mathcal{L}_{PA}$, that $T(\forall x \varphi(x)) \leftrightarrow \forall x T\varphi(x)$. The reason is that TB $(=MT)$ only contains Tarski-biconditionals for *sentences*, i.e., for *closed* formulas. In order to prove, for each $\varphi(x) \in \mathcal{L}_{PA}$, that $T(\forall x \varphi(x)) \leftrightarrow \forall x T\varphi(x)$, we need a slight strengthening of the Tarski-biconditionals of TB, namely the *uniform* arithmetical Tarski-biconditionals of the theory UTB.[52]

In sum, by adopting the rule H, Horwich *almost* achieves his aim of deriving all acceptable truth generalisations, but not quite. But Horwich could completely achieve his goal by *strengthening* his minimal theory of truth from a collection of Tarski-biconditionals (TB) to a collection of *uniform* Tarski-biconditionals (UTB). This would mean a transition from a theory of truth to a theory of *true of* (*satisfaction*). Horwich might well be open to this proposal.

9.6.2. From disquotational to compositional truth. Beside the argumentative strategy I have just sketched, there is also another way of justifying the compositionality of truth on the basis of disquotationalism and reflection principles. In Chapter 6, we saw that it does not matter whether our background theory is TB or TB^-,[53] since by uniform reflection we can prove mathematical induction for the extended language (Theorem 6.23). In other words, there is no need to take the *open-endedness* of mathematical induction as an *assumption* in our argumentation. Moreover, we also saw (Lemma 6.24) that the *uniform* Tarski-biconditionals can be derived from the Tarski-biconditionals using one application of uniform reflection.

In other words, there is no need to take the notion of satisfaction to be *basic*: the more general notion of satisfaction is *implicitly contained* in the concept of truth, and is made explicit by one application of uniform reflection. Taking all this together, we see that the full typed compositional truth theory CT is implicit in the weak disquotational theory TB^-, and becomes explicit by applying uniform reflection to TB^- *twice* (Theorem 6.26).

When we move to a classical *type-free* setting, the situation is similar. When we start from the weak typefree disquotational theory TBF,[54] we obtain the typefree compositional theory Pos(KF) by two applications of uniform reflection (Theorem 6.27). And this is one way of addressing the truth generalisation problem.

---

[52]The uniform Tarski-biconditionals were introduced on p. 129.

[53]The theory TB^- was introduced on p. 129.

[54]This disquotational theory, consisting of all positive truth-and-falsity biconditionals, was introduced on p.131.

At this point, it is worth taking a step back from the *minutiae* of the story. Horwich is a truth theorist with a philosophical but not a proof theoretic background. From a *purely philosophical perspective*, he came to propose the thesis that our belief in the compositionality of truth is grounded in our acceptance of proof theoretic reflection rules. This shows that Feferman's idea of implicit epistemic commitment to reflection principles is a quite natural thought. But the preceding pages also show *the importance of proof theory* for truth theory. It takes careful (if elementary) proof theoretic analysis to see that one round of uniform reflection does not suffice for obtaining compositionality from disquotational truth axioms. In any event, in the end a pleasing picture emerges about the relation between disquotationalism and the compositionality of truth.

Thus we end up in a curious situation. On the one hand, strong proof-theoretic reflection principles can be derived from compositional truth axioms. On the other hand, compositional truth axioms can be obtained from disquotational truth axioms by twofold uniform reflection. So from a purely logical point of view, disquotational truth together with uniform reflection is just as basic as compositional truth. Nonetheless a philosophical question stays with us: which is *epistemologically* more basic: disquotational truth together with uniform reflection, or compositional truth? The situation here is similar to the relation between proof theoretic reflection principles and principles of transfinite induction.[55] Here, too, even though there are many mathematical equivalences (as is emphasised in [**Dea15**]), they do not in and of themselves answer the corresponding epistemological questions.

Let us first consider the question of our warrant for disquotational truth axioms and for reflection principles.

Horwich argues that disquotational truth axioms are warranted by a *success argument*. On the one hand, the Tarski-biconditionals form a *simple and basic* theory [**Hor98**, p. 51]:

> [The Tarski-biconditionals] could be explained only by principles that are simpler and more unified than they are—principles concerning propositional elements and the conditions in which truth emerges from combining them. But the single respect in which the body of minimal axioms is not already perfectly simple is that there are so many of them—infinitely many; and no alleged explication could improve on this feature. For there are infinitely many constituents to take into account: so any characterization of them will also need infinitely many axioms.

On the other hand, the Tarski-biconditionals play a central role in implementing the *explanatory function* that the concept of truth plays in our theoretical and practical lives [**Hor98**, Chapter 3].

Halbach pointed out that the disquotational principles cannot be taken to be *analytical* truths, for they are not conservative over *logic* [**Hal01b**].[56] The simple argument goes as follows. Consider the Tarski-biconditionals

$$T(1 = 1) \leftrightarrow 1 = 1,$$
$$T(1 \neq 1) \leftrightarrow 1 \neq 1.$$

[55] See Section 6.2.1.

[56] In [**HH17**] it is shown that even in free logic, simple Tarski-biconditionals are non-conservative over logic.

The right-hand sides of these statements have (by a law of identity) opposite truth-values. So by Leibniz's law that identical objects have the same properties, the sentences '1=1' and '1≠1' must be different objects. Thus, *by logic alone*, simple Tarski-biconditionals allow us to conclude that there are at least two objects.

So if analyticity entails ontological neutrality, then the Tarski-biconditionals are not analytical. Nonetheless, there might be a case for saying that they are analytical in the sense that, just as the introduction and elimination rules of conjunction may be taken to give the meaning of the concept of conjunction,[57] disquotational truth rules or principles *give the meaning* of the concept of truth.

This line of reasoning has been questioned. Williamson argues that there is always a possibility for a logical expert to come to doubt some of these principles and rules on the basis of a complicated and ingenious, but ultimately confused argument. In such a situation, Williamson contends, we should not say that the experts fails to grasp the *meaning* of some of the logical connectives.[58]

Concerning uniform reflection, we are even more in the dark. Feferman claims that we are *implicitly committed* to uniform reflection principles that we accept without reservation. This claim invites questions about how and why we are so committed, and how we can proceed from such implicit commitments to corresponding explicit warranted beliefs. In the previous Chapter we have attempted to address such questions about the *microstructure* of epistemic commitments that accompany the explicit acceptance of theories. For consistency and local reflection we believe that we were relatively successful. For uniform reflection, however, we did not find satisfactory answers.[59] Consider in this context also that due to the "transcendence" of uniform reflection over local reflection, nothing less than a direct philosophical account of our epistemic warrant for uniform reflection will do here. In sum, whereas plausible epistemic accounts of our warrant for TB might be available, the same cannot be said for uniform reflection—not yet, anyway.

Let us now briefly turn to the question how an account of our epistemic warrant for compositional truth principles might look like. Here we seek an account that does not derive that warrant from a prior warrant in believing disquotational truth principles and (iterated) uniform reflection along the lines that were discussed earlier in this Section.

The claim that compositional truth principles directly give the meaning of the concept of truth does not carry conviction. Davidson sought to give a success argument of sorts for the thesis of the compositionality of truth. He held that this thesis must plays a central role in any theory that gives us a grip on the *meaning of natural language expressions*. However, the Davidsonian research programme of developing truth functional theories of natural language meaning has not been a resounding success story. So it seems that, at least so far, our warrant for compositional truth principles cannot rest on the success of the theories of meaning in which they figure prominently.

So the situation can be summed up as follows. The epistemic account of the compositionality of truth in terms of disquotational truth principles and iterated uniform reflection principles certainly has an appeal. However, the big unresolved question in this story is how we come to have warranted belief in uniform reflection

[57] See for instance [**Pra83**].

[58] See [**Wil07**, Chapter 4].

[59] See Section 8.6.

statements for theories that we already believe. This does not at this point come as a surprise, since we saw earlier that it is very difficult convincingly to justify uniform reflection without drawing on new conceptual resources.

9.6.3. From classical logic to partial logic. Beside the question how belief in uniform reflection is systematically warranted, there is, unfortunately, another problem with Horwich's strategy. Earlier, we saw that KF is in a sense self-undermining. In the same breath, it asserts the liar sentence (λ) and that the liar sentence is not true ($\neg T(\lambda)$).[60] But even PA, *formulated in the extended language* $\mathcal{L}_T$, and with the truth predicate allowed in the induction scheme, is arguably not sound for its intended interpretation. Call this theory PA^T. Using the diagonal lemma, we obtain in PA^T the sentence

$$(\lambda \wedge \neg T\lambda) \vee (\neg\lambda \wedge T\lambda), \tag{L}$$

where λ is the liar sentence. But (L) seems clearly unacceptable on its intended interpretation: *both* disjuncts of (L) are unacceptable, so (L) as a whole is unacceptable. If someone were to assert "this sentence is not true", and immediately afterwards state that what she has just said is untrue, we would not know what to make of her assertions. Similarly, something would be deeply wrong with an assertion of the second disjunct of (L). But if one of the two disjuncts is the case, then what would be wrong with asserting it?[61]

Since $RFN^2[TB] \supseteq PA^T$, the theory $RFN^2[TB]$ seems then likewise unsound. This would imply that the derivation of CT from $RFN^2[TB]$ does not have warranting force, and the truth generalisation problem remains unsolved.

But perhaps it would be too harsh to say that all truth theories that are formulated in classical logic are unsound. Following [**Fie94**], we may say that there are (at least) *two* concepts of truth. The first is a concept of truth that plays some role in scientific explanations—e.g. explaining communication by specifying truth-conditions for some natural language expressions; we can call this theoretical notion *scientific truth*. The second is a notion of truth that is governed by rules of semantic ascent and descent, and we can call it *disquotational truth*.

The scientific concept of truth is a *theoretical concept*, like the concept of force in classical mechanics, for instance. It is related to our pre-theoretic concept of truth. But there is no reason to think that it should coincide with it, just as there is no reason to expect that the scientific concept of force coincides with our pre-theoretic concept of force. The scientific concept of truth is governed by classical logic, since *all* scientific concepts are governed by classical logic.

The disquotational concept of truth, on the other hand, coincides with our pre-theoretic truth concept, which unrestrictedly serves as a a device of full quotation (*semantic ascent*) and disquotation (*semantic descent*).

Indeed, a core part of the meaning of the pre-theoretic truth predicate is given by principles that allow for a substitution of a sentence ϕ by the statement of its truth $T\phi$ and *vice versa*, at least in all extensional contexts. The principles of ascent and descent are the following rules of inference:

$$\frac{\phi}{T\phi} \qquad \frac{T\phi}{\phi}.$$

[60] See Proposition 5.32.

[61] This argument is discussed, among other places, in [**Fie08**, Chapter 6] and in [**Hor11**, Section 9.6.3].

The liar paradox then teaches us that *if* there is a coherent concept of type-free disquotational truth, then it is governed by non-classical logic.

The principles governing the scientific concept of truth are not warranted by their intrinsic plausibility, but by the extent to which they organise our empirical and non-empirical scientific knowledge effectively. In this way, the scientific conception of truth fits into a holistic, Quinean picture of scientific knowledge.

Davidson may still be proven right in his expectation that truth theories in the end play a pivotal role in successful linguistic theories, and more specifically, in linguistic theories of natural language semantics. Because linguistics is a scientific discipline, it will be governed by classical logic. And this means that the relevant truth theories will then also be governed by classical logic. In Quinean fashion, these truth theories will then gain support from their scientific consequences, despite the fact that they do not fully agree with our pre-theoretical intuitions about truth.

Suppose, just for the sake of argument, that KF is such a theory. The interesting question then arises to which extent KF, in such a situation, can play a justificatory role *in mathematics*. Suppose that you have mathematically warranted belief in PA *and in nothing more*. Presently you come to realise that KF plays a central role in certain highly successful theories of natural language semantics. Does that then warrant you to believe in the arithmetical consequences of predicative analysis? *If so*, then the nature of your warrant for believing in the arithmetical fragment of KF that goes beyond PA will be different from the nature of your warrant for believing in PA. The former is then at least in part empirical; the latter is *a priori*.

For disquotational truth, the situation is different. The full typefree disquotational rules *in partial logic*, which result in the weak disquotational theory TS_0,[62] can stake a claim to intrinsic plausibility. Also, repeated application of the uniform reflection rule to this theory does not yield any counterintuitive consequences. Theorem 6.36 teaches us that two applications of a natural uniform reflection rule to this disquotational theory results in a natural theory of compositional truth in partial logic, namely, PKF. Moreover, it follows from Theorem 6.35 that if we continue uniformly reflecting in this way along an autonomous path (in the sense of Feferman), then we are able to prove the theorems of a fragment of classical predicative analysis.[63]

Halbach and Nicolai have formulated a critique of the *disquotational* concept of truth that was discussed earlier,[64] and the partial theories of typed and untyped compositional truth that it gives rise to.[65] They argue for instance that PKF is inferior to KF on *foundational* grounds. Whereas KF proves and therefore justifies (the arithmetical part of) classical predicative analysis up to level ε_0 (Theorem 5.29), PKF falls short of this (Theorem 5.37). Moreover, we have seen that iterated applications of a uniform reflection rule applied to KF or even TFB lead one to warranted acceptance of *full* predicative analysis in an autonomous manner (Theorem 6.28).

Given that KF is to some extent self-undermining,[66] the question whether KF can *justify* mathematical theories at all is not easy to adjudicate. If one judges

[62]The theory TS_0 was introduced on p. 132.
[63]See [**FHN21**, Section 5].
[64]See Section 9.6.3.
[65]See [**HN18**].
[66]See Proposition 5.32 and the discussion following it.

that they cannot, then it is not clear that the fact that PKF cannot justify (the arithmetical part of) classical predicative analysis is objectionable. Also, we must in this context remember that Theorem 6.37 shows that this discrepancy in mathematical strength between KF and PKF disappears when we instead consider the *schematic* versions of these theories.

9.7. From rationality and justification to truth

Instead of taking the idea that proof-theoretic reflection principles express trust or acceptance for granted, 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.

9.7.1. Galinon. An attempt at doing this was made by Galinon in [**Gal14**], where he 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* [**Gal14**, p. 328]:

> If a rational agent accepts a collection T of propositions, then she must accept "T is acceptable".

Second, he endorses the following principle [**Gal14**, p. 325]:

> A rational agent must accept that if a collection of propositions is acceptable, then that collection is coherent.

Galinon argues for the Principle of Responsibility on the basis of norms of rationality [**Gal14**, section 7], and he argues for the second principle on the basis of a "Gödelian dutch book argument" [**Gal14**, section 5]. Using these two principles, Galinon develops the following argument for the acceptance of consistency statements [**Gal14**, p. 329]. 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.

The Principle of Responsibility seems a demanding requirement, however. In the light of our discussion in Section 8.3.2, 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.[67]

9.7.2. Believability theory. Galinon's account of the relation between norms of rationality and reflection is rather sketchy. In the final chapter of [**Cie17**], Cieśliński aims to give an epistemic account that explains why it is rational to accept reflection principles for a theory that one has good reasons to believe. (In what follows, PA will play the role of such a reasonable background theory.) Cieśliński's account of the relation between principles of rational belief on the one hand, and

[67] This stance is taken in [**FHN21**].

reflection principles on the other hand, is more detailed than Galinon's account. So let us discuss Cieśliński's view in some detail.

Mathematicians are typically *cautious*. As we have seen, they may be suspicious of the notion of truth: they may find it philosophical and speculative.[68] For this reason, Cieśliński develops an epistemic account of reflection in which the notion of truth does not play a role [**Cie17**, p. 252].

Mathematicians have reasons for their mathematical beliefs. We may reasonably expect of mathematicians that they reflect on these reasons, and on the concept of reason more generally. Cieśliński argues that this leads the mathematician to come to consider the concept of *believability*, and to accept certain principles that govern it. This seems a promising way to develop the connection between rationality and reflection principles. Rationality, after all, is a matter of having good reasons for one's beliefs. Moreover, in spelling out in some detail how, reflecting on reasons, a mathematician can come rationally to believe reflection principles for PA, Cieśliński's account goes substantially beyond that of Galinon.

Cieśliński explains the content of the concept of believability as follows [**Cie17**, p. 251]:

> The expression 'φ is believable' means that there is a good reason to accept φ. (To be more exact, the intuitive intended interpretation of 'φ is believable' is that there is a reason to accept φ which is *normally* good enough, with 'normally' meaning 'in the absence of strong reasons to accept the negation of φ').

Here I take 'there is a good reason' to mean that the mathematician in question *has* a good reason, i.e., that we a subjective rather than an objective notion of reason is at play here.

Believability in Cieśliński's sense is a term of art that is unfamiliar to most epistemologists. Believability is clearly closely related to the familiar notion of *justified belief*: it is common to say that a person is justified in her belief that φ if she has good reasons for believing φ. Cieśliński avoids the term 'justified belief' because he does not want to be burdened with the associations that stem from the formidable body of epistemological literature that is devoted to this notion. Intuitively, it indeed appears that the believability concept and justified true belief are not completely co-extensional. Consider again the situation where our mathematician has reasons that speak in favour of φ, as well as reasons that speak against φ. In those circumstances, φ would be believable, but if the mathematician were to believe φ, she would not be taken to be justified in doing so. Nonetheless, it is worth bearing the close connection between believability and justified belief in mind. Indeed, if the predicate B of believability is translated into the predicate J, then a super-theory of the base theory of Schuster and Horsten's basic theory of justified belief *minus* the consistency axiom $\neg J(0 = 1)$ is obtained.[69]

Cieśliński argues that from a rational acceptance of PA, our mathematician should rationally progress to a rational acceptance of Bel(PA). As we have seen in section 6.4, from the believability theory *over* PA, the believability of iterated uniform reflection principles for PA logically follows (Theorem 6.45).

Believability is a notion that cannot always rationally be discharged: "there is no automatic transition from $B(\varphi)$ to the rational acceptance of φ" [**Cie17**, p. 269].

[68] See Section 8.2.2.

[69] See Section 4.3.

The Co-Necessitation rule

$$\frac{\vdash B(\varphi)}{\vdash \varphi}$$

is indeed not valid. After all, a mathematician might *at the same time* have good reasons for φ and good reasons against φ. In that case, φ is still believable in the sense above, but it would be irrational for the mathematician to infer φ. In particular, this means that even though from Bel(PA), the believability of reflection principles for PA can be proved, there indeed is no "automatic transition" to an acceptance of these reflection principles themselves. It is only when she has no reasons that speak against them, that the mathematician can rationally infer to the reflection principles themselves. But of course, the mathematician can find herself in exactly this position, and come through the reflection process that Cieśliński describes come rationally to believe reflection principles for PA.

Typically, the theory Bel(S) will be proof-theoretically conservative over the background arithmetical theory S for $\mathcal{L}_{PA}$. The *internal theory* of Bel(S), i.e., the collection of statements that Bel(S) proves to be believable,[70] in contrast, will not be arithmetically conservative over S (Theorem 6.45). This internal theory of Bel(S) is then identified by Cieśliński with the *implicit commitment* of S. So the implicit commitment of S consists of what a mathematician rationally should judge to be believable if she accepts S. But since co-necessitation does not hold for B, she need *not* rationally come to believe all statements that she is implicitly committed to.

9.7.3. Believing the believability theory. The basic principles of believability theory sound like *norms of rationality* in the sense of Galinon. But the former are more precise than Galinon's two basic principles of rationality. In addition, Cieśliński gives a more detailed account of the structure of a process of reflection than anyone before him. Nonetheless, Cieśliński's account is still incomplete.

A first sense in which Cieśliński's account is incomplete concerns the epistemic warrant for the principles of believability theory themselves. In [**Cie17**], a fully detailed account of the mathematician's warrant for the believability axioms and rules is not given.[71] Let us consider this question in some detail.

We start by making two preliminary points. First, although the notion of believability may somehow be less speculative or philosophically charged than the notion of believability, it is a *new*, non-mathematical concept, which needs to be acquired in order for the reflective process to go through. Indeed, one can act from and believe on the basis of reasons without having a general concept of reason, whereas a general concept of reason is required for having the concept of believability. So the question arises whether the general (epistemic) concept of believability is kosher. I do not doubt that it is, but it is something of which the mathematician has to convince herself. In this respect, the reflection process that Cieśliński describes differs from the reflection process described in Chapter 8, which involves no new, non-mathematical concepts. Secondly, some idealisation is needed in Cieśliński's account. The steps in Cieśliński's reflection principles are non-trivial cognitive actions. So, just as the discharging of the believability predicate is not "automatic", neither is the progression from a rational acceptance of PA to a rational acceptance of Bel(PA).

[70] See Definition 6.42.

[71] Cieśliński's relevant considerations can be found in [**Cie17**, p. 252–256].

The starting point of a reflection process in Cieśliński's sense consists in the mathematician accepting PA on the basis of having good grounds for doing so. This entails her being in a situation where PA is believable for her. So, by introspectively becoming aware of the grounds of her acceptance of PA, and knowledge of the concept of believability, she can come to have good grounds for accepting Axiom B2 of Bel(PA).[72] By her understanding of the concept of believability, and her knowledge that having good reasons is closed under *Modus Ponens*, our mathematician can also come to have good grounds for accepting Axiom B3.

Rule B4 is somehow in the ballpark of uniform reflection: like RFN, it has the flavour of an effective version of Hilbert's ω-rule. Cieśliński is of course fully aware of the strength of this principle, and of the central role that it plays in his believability theory [**Cie17**, p. 255]:

> [Rule B4] is absolutely crucial. It is exactly this [Rule] which permits us to derive (in the scope of 'B') strong consequences, possibly unprovable in [PA] itself.

It is immediate that Rule B4 implies the much weaker Converse Introspection Rule

$$\text{(CIR)} \qquad \frac{\vdash BB(\varphi)}{\vdash B(\varphi)}$$

We already know from section 4.3 that in full generality, *most* introspection principles for justified belief (even if not CIR itself!) are inconsistent over a weak and seemingly unproblematic base theory. This gives us reason to believe that the justification of introspection axioms is no trivial matter. Indeed, even the principle CIR appears to be doubtful in full generality. In [**SH22**, Section 6], the following situation is considered:[73]

> Suppose that Catrin finds an apparent proof, which is complicated and long, for a mathematical statement ϕ. She takes her argument to be a valid mathematical proof, and checks it several times. She lets some of her colleagues check her proof, too: no one finds a mistake. Yet there *is* a subtle mistake in Catrin's mathematical argument. In this situation, it seems that Catrin is not justified in ϕ, since her argument contains a mistake. Nonetheless she is justified in believing that she is justified in believing ϕ. She fulfilled her epistemic obligations in that regard: she did all she could do to secure her belief that she has justified the conclusion of her mathematical argument.

There are variations on this scenario. For instance, instead of consulting her colleagues, she could run her putative proof through a proof checker, which, because of some hardware glitch or software problem would give the argument a clean bill of health. In situations such as these, it seems that Catrin has good reasons for believing $B(\phi)$; yet she does not have good reasons to believe ϕ. In any case, there is a non-trivial question here: *what are the mathematician's reasons for accepting Rule B4?*

---

[72] The Axioms of Cieśliński's believability theory were discussed in Section 6.4.

[73] In [**Hor18**], another worry about Rule B4 is discussed.

Rule B5 seems to be inductively justified, conditional on the rational acceptability of the other rules and axioms. *If* the principles B1–B4 are justified then it can easily be seen inductively that B5 must be fine, too.

Aside from the worry concerning Rule B4, it seems to me that in [**Cie17**] the *defeasible nature* of the inference rule

$$\frac{\vdash B(\varphi)}{\vdash \varphi}$$

has not been described in sufficient detail. In order to be a *complete* account of reflection, the conditions under which the believability operator may rationally be discharged from a proposition of the form $B(\varphi)$ must be described more fully. We also need an account of the way in which such a discharge can later again be withdrawn when reasons against φ are found. In other words, believability theory must be integrated with a theory of *belief revision*. I see no reason why this cannot be done, so I take this to be a task for future research on believability theory.

9.7.4. Discussion. The main worry appears to be the justification of Rule B4. But *if* epistemic warrant for that Rule can be found, then rational belief and even knowledge of (iterated) uniform reflection principles can be obtained through reflecting on believability.

I will now compare the believability reflection process with the reflection process that was described in Chapter 8, assuming that *both* processes can lead to rational belief in or even knowledge of reflection principles. Let *Belinda* be a mathematician who reflects on believability (over PA) and who agrees with Cieśliński about the philosophical interpretation and rationale of this process, and let *Conny* be a mathematician who reflects on consistency (of PA) in the manner described in Chapter 8.

The following two differences between the two reflection processes are obvious. Through her process of reflection, Belinda arrives at stronger reflection principles (uniform reflection principles) than Conny does through her process of reflection (consistency statements and perhaps local reflection principles). Belinda's reflection process essentially involves a new non-mathematical concept (believability), whereas Conny's reflection process does not. Therefore Belinda's eventual knowledge of reflection principles will not be purely mathematical in nature, whereas Conny's knowledge of consistency statements will be purely mathematical.

Belinda's notion of acceptance of the background theory (PA) is guarded, whereas we have seen Conny's notion of acceptance to be unconditional. Conny believes PA without any reservations. In particular, she does not doubt its consistency—she has not even considered the question. Belinda is more cautious in the sense that she does not *unqualifiedly* accept PA. She regards PA merely as very plausible *for the time being*, pending further evidence.

Belinda believes that Conny is rash. After stage 1 of her reflection process, Conny is able to carry out the following quick consistency proof.[74] She believes that she is disposed to accept any consequence of PA. In particular, she can see (by universal instantiation) that if she were to obtain a PA-proof of 0=1, then she would on the basis of that come to believe 0=1. But she can also easily come to know that she is not disposed to accept 0=1. Therefore, by *Tollendo Tollens*, she concludes that PA does not prove that 0=1. Indeed, this Myhill's consistency

[74]Thanks to Cezary Cieśliński for pressing these concerns in private conversation.

argument that we have considered earlier,[75] and it is a bare bones version of the reflection process that was described in chapter 8. Belinda's concern is that this is too easy, that epistemic warrants for the consistency of PA do not come this cheaply. This would have been avoided, Belinda says, if Conny's initial belief in PA would have been more guarded and conditional.

Conny believes, on the other hand, that Belinda is epistemically intolerant. Belinda maintains that the mathematician *should* believe the consistency of PA, based on a believability reflection argument [**Cie17**, p. 274]:

> [Accepting the believability axioms for the theory Th that he already accepts] is something which [the epistemic agents who reflects on his practice] *should* do, as rejecting (or even suspending judgement on) the statement 'All theorems of Th are believable' would make his practice irrational.

Conny believes that this is going too far. She urges Belinda to opt instead for a more liberal conception of rationality, on which it is not irrational to refrain from accepting Bel(PA), possibly because of doubt about some specific principle (such as Rule B4, perhaps), or because of scepticism about the viability of the general notion of justified belief (and thus also of believability). Nonetheless, Conny does not have to reject Belinda's rejection process in its entirety. She may admit that reflecting on believability may be *one* way to come to know proof theoretic reflection principles of PA. (Indeed, Conny sees that her own reflection process is not powerful enough to lead her to accept iterated uniform reflection principles for PA.)

9.7.5. Instrumental acceptance and believability. In the foregoing, we have investigated in some detail several forms of *acceptance of a theory*: full acceptance, acceptance as true, and believability. In this section, we take a closer look at *instrumental* acceptance of a theory.

In recent times, instrumentalism is probably best known from the philosophy of science. According to van Fraassen's anti-realist view, we should accept our best scientific theories in an instrumentalist sense [**vF80**]. Here instrumentalist acceptance of a scientific theory S is interpreted as entailing unconditional belief in the *observational consequences* of S, but no more. We *may,* as rational beings, also fully believe unobservational parts of S, but the canons of rationality do not oblige us to do so.

In the foundations of mathematics, Hilbert articulated and defended an instrumentalist position [**Hil26**]. According to this view, ZFC is an *ideal theory*, which we should not unconditionally believe. Instead, we should accept it as a useful engine for proving *real statements*, which we should fully believe. Hilbert was somewhat vague about the question which theorems of set theory count as real statements in this context. Perhaps the real statements of ZFC are those ZFC-theorems that are purely about the hereditarily finite sets, or, equivalently, its arithmetical statements. But there are also indications that Hilbert interpreted the class of real statements more restrictively as the ZFC-theorems that can be interpreted as statements in the quantifier-free language of PRA.

Hilbert's programme went further than his instrumentallist stance towards ideal mathematics. The aim of his program was to prove the consistency of set theory

[75] See p. 227.

(or perhaps even the uniform reflection principle for ZFC restricted to arithmetical statements) in a "real" mathematical theory such as PRA (or, perhaps, less restrictively: in PA). We know that this program failed in spectacular fashion. Nonetheless, one can adopt Hilbert's instrumental stance towards infinitary mathematics without signing up to his program.[76] Observe that if one would take such an instrumentalist stance towards ideal mathematics, then it seems that one would be implicitly committed to certain statements that go far beyond PA. If one had reasons to believe $\neg Con(PA)$, for instance, then it could not play the instrumentalist role that Hilbert wants it to play.

We have seen that the most popular theory of type-free truth, KF, has self-undermining features (Proposition 5.32). So it is difficult to fully believe everything that KF proves. For this reason, instrumentalism is particularly appealing in the area of theories of self-referential truth.

Even though it is in some sense self-undermining, KF is *arithmetically sound* (Theorem 5.26). So one option is to see KF as an engine that reliably produces arithmetical truths. Perhaps, in some of his moods, Feferman saw things along these lines, for he writes in his classical article on KF [**Fef91**, p. 3]:

> The schematic notion of reflective closure meets among other things the aim to give a more perspicuous generation procedure for predicativity without use of progressions of theories or *prima facie* impredicative notions such as those of ordinals or well-orderings. As already mentioned, this had previously been accomplished [...] in a quite different way tied essentially to the basic ideas of predicativity. The characterization there may still be considered more persuasive for that purpose, but I hope the reader will find the use of the notions here to be of independent interest as a general means of expressing closure under the reflective process.

Reinhardt has also advocated an instrumentalist stance towards KF. But he believed that instead of only its first order arithmetical consequences, we should fully believe its entire internal logic IKF [**Rei86**].[77] The reason is that IKF belongs to the extension of the truth predicate at the least fixed point model, which is an intended interpretation of KF. It is not clear that such a semantic justification for IKF is acceptable: it appears to be an appeal to a notion of truth in the metatheory that goes against the spirit of the axiomatic approach to truth. But let us set that worry aside here.

At this point one may wonder whether the *instrumental* acceptance of KF carries with it implicit commitments that go beyond KF. The investigation of this question, using tools of believability theory, is the subject matter of [**CG23**].

A first observation is that we cannot identify the implicit commitments of an instrumental acceptance with Bel(KF), where Bel is Cieśliński's believability theory that we have discussed earlier: otherwise someone who instrumentally accepts would be implicitly committed to the whole of KF! So the believability theory would have to be modified for instrumental acceptance.

---

[76] Similarly, one could take an instrumentalist stance towards class theory, and see a class theory such as MK merely as an 'ideal' instrument for proving first-order 'real' (i.e., first order) statements about sets.)

[77] The theory IKF was defined on page p. 142.

Reinhardt argued that someone who *instrumentally* accepts KF (in his sense) is not implicitly committed to $RFN[KF]$, but instead to the following "internal" uniform reflection principle [**Rei86**, p. 232]:

DEFINITION 9.2. $RFN_I[KF] \equiv \forall \varphi \in \mathcal{L}_T : \mathrm{Bew}_{KF}(T(\varphi)) \to T(\varphi)$.

Importantly, Reinhardt observes that $RFN_I[KF]$ is in the extension of the truth predicate in the least fixed point model. So it can be coherently added to IKF to obtain a stronger but equally unobjectionable truth theory.

All this is taken by Castaldo and Głowacki as a clue for how Cieśliński's believability theory should be modified for the appropriate epistemic attitude towards KF. The resulting believability theory for instrumental acceptance of KF, which we call $Bel^*(KF)$ is formulated in the language $\mathcal{L}_{T,B}$, and consists of the following principles:[78]

$B1^*$
$$KF;$$

$B2^*$
$$\forall \varphi \in \mathcal{L}_T : \mathrm{Bew}_{KF}(T(\varphi)) \to B(\varphi);$$

$B3^*$
$$\frac{\vdash \forall x B(\varphi(x))}{\vdash B \forall x \varphi(x)};$$

$B4^*$
$$\frac{\vdash T(\varphi) \to T(\psi)}{\vdash B(\varphi) \to B(\psi)}.$$

Clearly Axiom $B2^*$ is inspired by Reinhardt's reflection principle $RFN_I[KF]$. Rule $B3^*$ is again a version of a formalised ω-rule. Rule $B4^*$ expresses (roughly) that having a proof that a statement is *true* is a good reason for judging that statement to be believable.

Castaldo and Głowacki go on to show that the new believability theory $Bel^*(KF)$ has the following properties:

THEOREM 9.3. *$Bel^*(KF)$ is conservative over KF for sentences not containing the believability predicate B.*

THEOREM 9.4. $Bel^*(KF) \vdash B(\neg \mathrm{Bew}_{KF}(0 = 1))$.

THEOREM 9.5. $Bel^*(KF) \vdash B(RFN(IKF))$.

Theorem 9.4 says that even instrumental acceptance of KF *à la* Reinhardt entails some sort of commitment to the consistency of KF (namely, commitment to the believability of its consistency), in the same way that Hilbertian instrumentalism entails a commitment to the consistency of ideal mathematics.

Theorem 9.5 says that Reinhardt's reflection principle $RFN_I[KF] (= RFN[IKF])$ also belongs to the implicit commitments of instrumental acceptance of KF. This makes one wonder how tight the connection between the "normal" believability theory over IKF on the one hand, and the modified believability theory over KF is. After all, much if not all of the content of Reinhardt's instrumentalism seems to be that we should fully believe IKF. So, in particular, one might wonder:

[78] Actually, this is only one out of four believability theories modified for instrumental acceptance that are discussed in [**CG23**]. The authors do not express a preference for one of these four believability theories.

QUESTION 9.6. How does $Bel^*(KF)$ relate to $Bel(IKF)$?

In some sense, one would expect them to coincide. But literally speaking, this can hardly be correct since IKF is a non-classical theory, so presumably Bel(IKF) is non-classical, too. Indeed, in order to make progress with this question, one would first have to spell out in precise terms what the believability theory over a non-classical theory looks like.

In our discussion of Cieśliński's believability theory for unconditional acceptance,[79] we expressed some reservations about a variant of Rule $B3^*$. It is not hard to see that these worries also apply to Rule $B3^*$ itself. Against the background of this, it is worth remembering that the detour through believability theory is perhaps not necessary. One might take the implicit commitment of instrumentalist acceptance of KF to be *directly* given by $RFN_I[KF]$ (and iterations of it). But then more would need to be said. Just as there is the question how and why we are implicitly committed to uniform reflection for theories that we unconditionally accept, we have the question why and how we are committed to the uniform reflection principle $RFN_I[KF]$ if we instrumentally accept KF.

9.8. Warrant for set theoretic reflection

In the last Section of this Chapter, we turn to questions about warrant for the set theoretic reflection principles that were discussed in Section 6.5. Many philosophers of mathematics believe that we have a good grip on questions of warrant for set theoretic reflection principles. Against this, I will argue that the question why (and indeed whether) set theoretic reflection principles are true, is in fact not an easy one.

9.8.1. From theology to richness. We have seen earlier how Cantor adopts Augustine's view that the mathematical objects (such as sets) are ideas in the mind of God.[80] Moreover, Cantor argues that not only God's mind as a whole, but also the mathematical part of it, cannot even approximately be known [**Can32**, Abhandlungen zur Mengenlehre III, Endnote to section 4, p. 205]:

> The Absolute can only be acknowledged, but never known, nor even approximately known.

Here Cantor takes us to already have 'approximate knowledge' of God if we can 'take the measure' of a dimension or compartment of God's mind. In particular, we would have approximate knowledge of God's mind if we could 'measure' the extent of it by means of the natural numbers, since the concept of the natural numbers is, for Cantor, perfectly clear. This leads him to posit a *reflection argument* in mathematics [**Can32**, Abhandlungen zur Mengenlehre III, Endnote to section 4, p. 205]:

> Whereas hereto, the infinity of the first number class (I) [i.e., the class of finite cardinal numbers] alone has served as such a symbol [of the Absolute], for me, precisely because I regarded that infinity as a tangible or comprehensible idea, it appeared as an utterly vanishing nothing in comparison with the absolutely infinite sequence of numbers.

[79] See Section 9.7.4.

[80] See p. 76.

In other words, Cantor concludes from the unknowability of the mathematical universe as a whole to the existence of ω as a completed infinity.

His reasoning here goes along the following lines.[81] Suppose there is a one-to-one correspondence between the natural numbers and the mathematical world as a whole. Then 'the measure has been taken' of the mind of God using a perfectly clear measuring stick (the natural numbers). So by elementary knowledge of the natural numbers we have knowledge of the mathematical part of the mind of God. But this is incompatible with the epistemic transcendence of God. Therefore the collection of the natural numbers must be of bounded size in comparison to the immeasurability of the mind of God. Bounds are given by numbers. Therefore there must be a number that measures the size of the natural numbers. This will then have to be a transfinite number: a bounded completed infinity.

There seems no obstacle to the *human knowability* of the number that measures the size of the natural numbers, since this knowing this number would give us no knowledge of the mathematical compartment of the mind of God, which immeasurably transcends the collection of the natural numbers. All this also holds for other 'clear' collections of numbers, such as the rational numbers and the real numbers. So might as well *try* to come to know the cardinal number of the natural numbers, as well as the cardinal numbers of other infinite collections, and how to calculate with these transfinite numbers. And this is of course exactly what Cantor did. In this way, he went beyond what Augustine thought possible. The latter's remarks were tentative, and he thought that in any event calculating with transfinite numbers is beyond the intellectual capacities of humans.

Cantor's reflection argument is restricted in scope. The Burali-Forti argument shows that the plurality of all ordinal numbers does not form a set. So, for Cantor, the infinity of all ordinals cannot be a 'tangible, comprehensible idea', and is therefore not subject to a reflection principle.[82] This is somewhat puzzling, though, since the definition of the concept of ordinal seems quite perspicuous.

We have seen how Gödel thought that set theoretic reflection principles follow from the unknowability of V. Actually, it seems rather the *undefinability* of V, *together with its classes*, that is the fundamental thought here. It motivates the that *if* a definable property holds in V (with its classes), then

Modern set theoretic reflection arguments more closely follow Philo's reasoning[83] than Cantor's reasoning about the Absolute. Set theoretic reflection principles center around *indiscernibility*. They somehow express that the set theoretic universe V is indistinguishable from certain parts of V. It is not completely clear which notion of indiscernibility Philo had in mind: perceptual indiscernibility, epistemic indistinguishability in general, semantic indiscernibility. . . In modern set theory, the focus is firmly on a form of *semantic indiscernibility*. Like in Philo's reflection from God to certain angels, and in contrast to Cantor's reflection arguments, no distinction is made in modern set theoretic reflection between 'clear' and 'unclear' (or 'indefinite') infinities.

Many (but by no means all) foundational researchers today believe reflection arguments in set theory to be warranted. Maddy, for instance writes that reflection argumentation "is probably the most universally accepted rule of thumb in higher

[81] See [**Hal84**, p. 116–118].

[82] In his later work, Cantor calls the plurality of all ordinals an *inconsistent multiplicity*.

[83] See Section 3.3.

set theory" [**Mad88**, p. 503], where "rules of thumb" are "vague intuitions about the nature of sets, intuitions too vague to be expressed directly as axioms, but which can be used in plausibility arguments for more precise statements" [**Mad88**, p. 484].

Contemporary foundational researchers seek to support reflection arguments and reflection principles in set theory by *richness considerations*.[84] The idea is that "the set-theoretic universe should be sufficiently rich (in the sense that there are sufficiently many sets of varied kinds) that we are unable to distinguish the universe from one of its initial segments" [**Bar16**, p. 354].

Maddy gives an excellent example of this kind of reasoning. Let me quote her somewhat at length here [**Mad88**, p. 750–752]:

> The most general version of Vopenka's principle states that any proper class of structures for the same language will contain two members, one of which can be elementarily embedded in the other. The rule of thumb usually cited as lying behind this principle is the idea that the proper class of ordinals is extremely rich [...] Suppose, for example, that a process is repeated once for each ordinal—*Ord*-many times, we might say—and every step produces a structure. Then richness implies that no matter how closely we keep track of the structures generated, there are so many ordinals that some will be indistinguishable. A similar idea can be developed from reflection: Anything true of V is already true of some R_α, that is, there is an R_α that resembles V. This property of V should also be reflected, that is, there is an R_α with a smaller R_β that resembles it.
>
> Either way, we get a new rule of thumb, *resemblance*:
>
> ...there are R_α's that resemble each other.
>
> ...there should be stages R_α and R_β which look very much alike.
>
> [...] The trick, of course, comes in spelling out "resembles".
>
> To do this, let us go back to richness and imagine ourselves in an *Ord*-long process, generating an R_α, at each stage, one for each ordinal. Suppose we step several ranks at a time, so that by step α, we are already to R_{γ_α}, for some $\gamma_\alpha > \alpha$. We keep careful track of the structures at each stage by making copious notations on a clipboard, one scoresheet for every stage; we note down every detail of the structure we have just generated, along with every detail of the process that got us there. Richness then implies that with so many stages, our scoresheets cannot all be different. At step one, we record the complete diagram of
>
> $$(R_{\gamma_0}, \epsilon, \langle R_\beta : \beta < 0\rangle).$$
>
> At step two, we look to see if that scoresheet is satisfied by
>
> $$(R_{\gamma_1}, \epsilon, \langle R_\beta : \beta < 1\rangle).$$
>
> Of course it is not, so we write down the complete diagram of this new structure. And so on. At each step, we generate a new

[84] See for instance [**Mad88**], [**Bar16**].

structure, then check to see if any of our old scoresheets will do; if not, we prepare a new one.

Richness then guarantees that we will eventually reach a step α' where one of our old scoresheets will match up. That is, we will reach a step α' where

$$(R_{\gamma_{\alpha'}}, \epsilon, \langle R_\beta : \beta < \alpha' \rangle)$$

is a model of the complete diagram of

$$(R_{\gamma_\alpha}, \epsilon, \langle R_\beta : \beta < \alpha \rangle)$$

for some $\alpha < \alpha'$. This means that the smaller structure can be elementarily embedded in the larger; that is:

$$\exists j : (R_{\gamma_\alpha}, \epsilon, \langle R_\beta : \beta < \alpha \rangle) \longrightarrow_e (R_{\gamma_{\alpha'}}, \epsilon, \langle R_\beta : \beta < \alpha' \rangle).$$

This embedding must be nontrivial, because:

$$\begin{aligned} \alpha' &= \text{length}(\langle R_\beta : \beta < \alpha' \rangle) \\ &= \text{length}(j(\langle R_\beta : \beta < \alpha \rangle)) \\ &= j(\text{length}(\langle R_\beta : \beta < \alpha \rangle)) \\ &= j(\alpha). \end{aligned}$$

We thus have a nontrivial elementary embedding of R_α into $R_{\alpha'}$. Our conclusion is a special case of Vopenka's principle, namely, that in any proper class of R_α's, there is a nontrivial elementary embedding of one into another.

However, statements about the richness of V do not have the privileged epistemological status that bible-based statements about the nature and properties of God once had. This raises the question: how do we *know* that the set theoretic universe has the required richness to support reflection arguments and reflection principles? We have seen that Maddy at this point makes a vague appeal to mathematical intuition. But this is hardly satisfactory, and more needs to be said.

9.8.2. The nature of classes. Let us tread carefully and to go slowly at this point. First we must prepare the background by being clear about what the prerequisites are for set theoretic reflection principles to *make sense*. Then we can proceed to investigate why we might accept (some of) them.

We have seen that set theoretic reflection principles are mostly not only about sets but also about proper classes. They cannot be true if there are no proper classes. The universe V must exist, as well as many other proper classes (such as Ord), and quantification over proper classes must make good sense. Let us for the sake of argumentation grant all this.

Russell's argument shows that proper classes cannot be sets. But we have seen that a predicative conception of classes also will not do. Indeed, against the background of a predicative class theory such as NBG, already Bernays' reflection principle entails that the classes are governed by the Morse-Kelly axioms.[85] Moreover, under these assumptions there would in addition have to be a global well-ordering of the set theoretic universe.[86] There is no consensus among set theorists

[85]See Theorem 6.51.

[86]See Theorem 6.52.

about whether there is a global well-ordering of the universe; whether there exists an (in some sense) *definable* well-ordering of V is of course at least as dubitable.

One person's proof is another person's reductio. On the one hand, one may see this connection as conclusive evidence that the predicative conception of classes is untenable. On the other hand, one can see this as conclusive evidence that set theoretic reflection principles are false. The latter position is taken and defended in [**Sch94**].

There appear to be three conceptions of classes that are robust enough for set theoretic reflection principles to make sense. First, there is the conception of proper classes as very large collections. Second, there is the conception of classes as *mereological parts* of the set theoretic universe.[87] Third, there is the conception of classes as *pluralities of sets*.[88]

Until the 1980s, the first conception, which we may perhaps still call the *standard conception*, was the only rival to the predicative conception of classes. On this view, proper classes are collections, and do not differ in this respect from ordinary sets. Nonetheless, proper classes cannot be sets. According to the standard conception, this is because proper classes are *too large* to be a set. Of course Russell's argument can be applied to classes too. There can be no class of all proper classes, for such a collection would be too large to be a proper class. The collection of all classes is then a hyperclass. The collection of all hyperclasses is again too large to be a hyperclass, and so on.

According to the mereological conception of classes, the set theoretic universe is not a collection but a mereological whole, where the *parthood relation* is the sub-class relation. The laws that govern the mereological parthood relation are quite different from the laws that govern the elementhood relation. For instance, a part of a part of x is always a part of x, whereas all elements of elements of y are elements of y only if y is a transitive set.

The plural conception of classes is a *no class*-theory of classes. According to this conception, sets are in the final analysis all the mathematical entities there are. But there exist two types of quantification. First, there is *singular quantification*, which takes the following form:

There *is* at least one entity x such that the property φ holds of *it*,

where for our purposes, the entities that are quantified over are the sets. Aside from this, in ordinary discourse we also use *plural quantification*, which generally takes the following form:

There *are* some entities xx, such that the property φ holds of *them*.

Singular quantification is of course standardly formalised by the familiar tool of first-order quantification ("$\exists x \varphi(x)$"); its laws are described in first-order logic. Plural quantification is easily confused with singular quantification over entities that are of a different nature than sets, namely classes. But this interpretation can be, and should be, resisted. The reason is that plural quantification over sets does not carry with it an ontological commitment to entities that are in certain respects like sets, but that are not sets but proper classes.

These three conceptions of classes all motivate *impredicative* class theories. Indeed, arguably on each of these conceptions, class theory is governed by the

[87] See [**WH16**].
[88] See [**Boo85**].

laws of ZFC^2. So if we want to decide between them, then we have to do so on philosophical grounds. So let us now briefly discuss some philosophical differences between the three conceptions.

Hard questions can be, and have been, asked about the standard conception of classes. According to this view, our most general theory of collections is not given by set theory, but by a theory of the hierarchy of sets, classes, hyperclasses, and so on, where the question arises how far this hierarchy continues. At this point, one's instinct is to *internalise* the type structure of this hierarchy in roughly the way in which ZFC internalises the structure of the ranks of the universe, which can after all also be seen as types. But once this is done, one wonders whether anything was gained by the whole exercise. After all, one ends up with a theory of collections that is virtually indistinguishable from some strong theory of sets that includes many inaccessible ranks.

Let us now turn to the mereological conception of classes. Material objects are standardly thought to be governed by the part-whole relation. Perhaps also mental entities have parts. Thoughts and emotions may, for instance, perhaps be seen as parts of minds. But the mereological conception of classes non-trivially extends the province of the part-whole relation by positing that even certain abstract entities (such as V) have proper parts.

When set theorists engage in class talk, as they freely do in informal practice, they often *seem* to use the machinery of singular predication (*the* class of all sets, *the* class of the ordinals). Indeed, when one tries to paraphrase apparently singular quantification over more complicated classes in terms of plural quantification, it sounds increasingly awkward and hard to understand. Thus, when taken at face value, class talk commits set theorists to extending their ontology to extend their ontological commitment to include proper classes. So the plural interpretation of class talk involves a *rational reconstruction* of this piece of informal discourse.

As far as providing a suitable background for set theoretic reflection principles goes, the three conceptions discussed above are certainly more suitable than a predicativist conception of classes. Still, there are some subtleties here.

The mereological conception of classes provides a very natural natural framework of set theoretic reflection principles. The quantification involved is ordinary first-order quantification, which is well understood. So set theoretic reflection principles can be interpreted straightforwardly as saying that the set theoretic universe, as a mereological whole, with all its parts, resembles a small part of the set theoretic universe (a set), with *its* parts. This is in harmony with the history of ontological reflection, and in particular with the way in which Philo conceived of ontological reflection.[89]

If we opt for the plural interpretation of classes, then the notion of *resemblance* that is at the heart of set theoretic reflection has to be "pluralised", too. There will then be no question of a small entity *resembling a large entity*, because the large entity does not exist. Instead, on this conception, a set theoretic reflection principle expresses that some sets and their relations resemble some other sets and relations on *them*, where the latter sets are "small in number".

In the more established set theoretic reflection principles, there is mostly no talk about hyper-classes, hyper-hyperclasses, etcetera. These reflection principles do not express that the universe, with all its, classes, hyper-classes,... resemble some

[89]See Section 3.3.

set with its power set, the power set of its power set, etcetera. This may be seen as a mark against the standard conception of classes as an appropriate framework for considering set theoretic reflection principles. On the other hand, the standard conception can accommodate higher-order reflection principles, which is not easily done with the mereological conception or with the plural conception.[90] Thus the standard conception of classes might serve as a framework for considering higher-order versions of Welch Reflection,[91] which yield α-extendible cardinals for $\alpha > 1$.[92] *Perhaps* the standard conception of classes can even function as a framework that accommodates plausible set theoretic reflection principles that entail the existence of much larger infinities, such as supercompact cardinals.

9.8.3. What rests on what? Let us now take up the question with which we were left at the end of Section 9.8.1, namely the question wherein our warrant for set theoretic reflection principles consists.

We have seen that strong set theoretic reflection principles are class theoretical statements. Many set theorists are somewhat suspicious of classes in general. So these researchers will regard set theoretical reflection principles as not warranted at all.

Many of those who do take set theoretical reflection principles to be warranted (such as Gödel) take them to be *intrinsically justified*.[93] More specifically, they believe that there exists such a rich variety of sets (and in particular ranks of V) that the universe as a whole, together with its classes, has a high degree of *undefinability*.[94]

We have seen how something like this metaphysical thought used to be backed by *theology*. The theological justifications for ontological reflection that were given before the twentieth century appeal to what the bible teaches us about God. As a source of knowledge, the bible had a status similar to that of perceptual experience. Its statements were treated as *givens*.

When God disappeared, He cleared out his bank account. The old theological arguments convince no longer. Indeed, we live in an age in which all theological argumentation is met with a great deal of scepticism. So we must really face the question: *why should we believe that the set theoretic universe is so rich that it supports set theoretic reflection principles?* Unfortunately, contemporary philosophy of set theory has not really engaged with this question to any great extent. At any rate, vague appeals to mathematical intuitions at this point do not suffice. If the richness premise is not argued for in detail, then it is hard to evaluate the strength of the argument in favour of set theoretic reflection principles.

Against the background of this, Sam Roberts defends the claim that set theoretic reflection principles are not intrinsically but *extrinsically justified*: we should believe these principles because of their consequences.[95] He argues that within

---

[90]The plural conception would have to be extended in such a way that hyper-classes are conceived of as "super-pluralities".

[91]See p. 169.

[92]For considerations in this direction, see [**Rob17**].

[93]We discussed the distinction between intrinsic and extrinsic epistemic warrant in Section 1.7.

[94]For an example of such a defence of strong set theoretic reflection principles, see for instance [**McC21**].

[95]See [**Rob**].

the rather broad category of extrinsic justification, set theoretic reflection principles score particularly highly on the dimension of *unification* of our set theoretic knowledge. We have seen that even weak set theoretic reflection principles make the Axioms of Infinity and Replacement superfluous. Moreover, strong reflection principles entail most of the strong principles of infinity that play an important role in set theoretic practice. Furthermore, set theoretic reflection principles entail natural generalisations of axioms of standard first-order set theory (Replacement, Choice). Lastly, one might add that set theoretic reflection principles entail (long iterations of) proof theoretic reflection schemes for the background theory ("good" consequences) in ways that large cardinal axioms typically don't.[96]

The truth of set theoretic reflection principles is not the *only* explanation of their success. *Embedding principles* seem to equally good at unifying and organising our set theoretic knowledge, but they differ from 'standard' reflection principles in not implying the existence of a global well-ordering of V. Indeed, if there are plausible large cardinal principles that can be formulated as embedding principles but not as 'standard' reflection principles, then embedding principles are perhaps even better at unifying our set theoretic knowledge. In other words, it could be seen as an alternative explanation of the "success" of set theoretic reflection principles that their desirable consequences are entailed by embedding principles. (Of course, the success of embedding principles and of set theoretic reflection principles could still be good reasons for accepting *both* of them.)

One might wonder, however, if the explanatory direction might not be the other way round. In Section 1.7, I argued that neither the acceptance of the axioms of ZFC, nor the acceptance of large cardinal axioms *needs* justification: the acceptance can be warranted without being justified. Now suppose that acceptance of large cardinals principle *is* warranted in some such non-justificatory way. Large cardinal axioms entail reflection phenomena *within* V. They entail that there are ranks V_κ that are very similar to many other ranks, in the way that "miniature versions" of reflection principles express.[97] Moreover, reflecting ranks resemble the true state of affairs to a significant degree. (For instance, they are standard models of the basic axioms of set theory.) Now suppose that one also accepts that the universe V and its "large parts" exist as entities, namely as proper classes. The universe V, with its classes, *is* the true state of affairs, and is suitably closed in the way that reflecting ranks are. *Therefore*, one might say, one might expect ontological reflection to hold also for V and its classes. Therefore, in other words, set theoretic reflection principles are likely to be true. (A similar argument can of course be given for class embedding principles.)

Some set theorists may actually reason in this way. They are interested in what holds for all sets, for all ordinals, etcetera. In discussing the situation they are interested in, they *naturally* extend their ontology: they treat the set theoretic universe, the ordinals, and so on, as entities that are *much like* sets. Of course these mathematicians are aware of the fact that these "entities" are not sets. They have typically never heard of plural quantification, or of mereological wholes. So they are somewhat vague about the nature of these new entities, about how many of them there are, and about the laws that govern them. Moreover, they do not want to worry much about such matters, because they know that these new entities are not

[96] See Theorem 6.7.

[97] First-order versions of embedding principles were briefly discussed on p. 174.

mathematical entities in exactly the same sense as the objects of their investigations, i.e., the sets. Nonetheless, they take the similarity between V and its classes on the one hand, and reflecting ranks (with their subsets) on the other hand, to be sufficiently strong to accept precise set theoretic reflection principles for V.

Reflection *arguments* and *phenomena* play an essential role in set theoretic practice: they crop up everywhere. But if the foregoing considerations are along the right lines, *set theoretic reflection principles* do not play an important role in set theory, and set theoretic reflection principles do less justificatory work in set theory than is fairly commonly believed.

Part IV

CHAPTER 10

Outlook

We have discussed lines of research that have been pursued intensely in mathematical logic since the 1950s. These research directions were from the start connected with deep foundational ideas. Philosophers have been slow to analyse the philosophical ideas behind these research programmes, and they (we) are still lagging far behind. That is probably just in the nature of the discipline: *the owl of Minerva takes flight only at dusk.* Be that as it may: there is much work for philosophers still to do.

In this short final chapter, I do two things. First, I try to canvas what we have have learned in our investigation of reflection in the mathematical sciences. Second, I identify some main open problems, and point to possible directions of future research. I will be happy if concerning one or two of them, significant progress will be made in the coming years.

10.1. Looking back

Even though from a logico-mathematical point of view, quite a bit is known, our philosophical discussion of numerous issues has been rather exploratory in nature. In particular, we might *think* that we know quite a bit about the epistemology of reflection principles in the mathematical sciences. But this, I believe, is an illusion.

Reflection principles are thought to play a significant role in reducing incompleteness in the mathematical sciences. Proof theoretic reflection principles reduce Gödelian incompleteness, whereas set theoretic reflection principles reduce incompleteness phenomena that are revealed by forcing. (Of course there may be sources of mathematical incompleteness that we do not yet know about.)

We know fairly well, I believe what proof theoretic reflection principles are, and what set theoretic reflection principles are. But in the case of set theoretic reflection principles, even this is not completely trivial. It is not always clearly recognised, for instance, in which sense embedding principles differ from "real" set theoretic reflection principles.

Proof theory has done an excellent good job, over the past 70 or so years, in calibrating the strength of proof theoretic reflection principles, both the "purely mathematical" ones, an the proof theoretic reflection principles in which the concept of truth is involved. Even though the situation in the case of set theoretic reflection principles is not as clear, there also we know quite a bit about their strength. Surprisingly, it has emerged in the last decade that natural set theoretic reflection principles can have quite a lot of large cardinal strength. Indeed, Gödel's bold conjecture that all plausible large cardinal axioms can be obtained from set theoretic reflection principles seems to me still not to have been refuted. But as for large cardinal principles, there seem to be limits to the extent to which set theoretic reflection principles can reduce incompleteness revealed by forcing. It seems

unlikely, for instance, that reflection principles will be of much help in resolving the Continuum Problem. As far as consistency strength goes, but not necessarily as far as outright strength goes, set theoretic reflection principles easily outperform proof theoretic reflection principles.

On the question of *epistemic warrant* for reflection principles, it seems to me that we are in a better position *vis-à-vis* proof theoretic reflection than *vis-à-vis* set theoretic reflection. There is a sense, I think, in which warranted full acceptance of S puts us in an excellent epistemic position for coming to accept iterations of consistency and local reflection principles for S in a warranted manner. Concerning set theoretic reflection, the smoke has not lifted yet. On the "intrinsic" side, it is at present unclear what the epistemic force of richness arguments is. The reason for this is that these arguments have not yet been worked out in sufficient detail to evaluate them. The strength of "extrinsic" arguments for set theoretic reflection principles seems to me at present also somewhat unclear, because it is not clear how integrated the use of set theoretic reflection principles is in set theoretic practice.

10.2. Looking forward

It is needless to say that it is difficult to look into the future. Nevertheless, I will now briefly discuss what I take to be some of the main open problems.

10.2.1. Reflection in the history of philosophy. In Chapter 3, and attempt was made at tracing the concept of reflection in the history of philosophy and theology. But our journey through the history of philosophy and theology was very far from complete.

We saw in Section 3.9 that in the nineteenth century, reflection started to play a role in mathematics. But this does not mean that the concept of reflection ceased to evolve in philosophy and theology. So we may ask:

Question 10.1. Which nineteenth and twentieth century philosophical or theological theories of reflection are of relevance to the philosophical understanding of reflection in the mathematical sciences?

The traditions of *idealism* and *phenomenology* may be good places to look.

10.2.2. Warrant for uniform and global reflection. In Chapter 8 I have argued that if someone fully accepts a theory S, and is moreover warranted to do so, then a process of *reflecting on the acceptance* of S can lead her to accept iterations of consistency and local reflection statements for S in an equally warranted fashion. What about uniform reflection?

It has been known for a long time that truth might be able to come to the rescue. The proof theoretic global reflection principle $GRF(S)$ for S can be derived from compositional truth principles against S as a background theory (Theorem 5.14), and from $GRF(S)$ the uniform reflection principle for S can be easily derived. General scepticism about the epistemic import of the derivation of proof theoretic reflection principles from truth axioms seems unwarranted.[1] But we do, on this line of reasoning, face the question of our epistemic warrant of the compositional truth axioms. We have seen that compositional truth can be derived from (twofold) iteration of uniform reflection (Theorem 6.26). But this would be putting the cart before the horse—which is not to say, however, that there may not be convincing

[1] See Section 9.4.

alternative accounts of how we can acquire warranted belief in uniform reflection statements.

The question how we can attain warranted belief in uniform reflection principles without appealing to "higher-order concepts" remains. It is not clear how reflecting on the acceptance of a theory S can warrant explicit acceptance of the uniform reflection principle for S.[2] The thesis of implicit commitment to uniform reflection, which goes back all the way to Feferman's work on transfinite progressions of formal theories in the 1960s. has been spelled out in other ways. However, I have argued in Section 7.4 that known attempts at justifying uniform reflection for an accepted theory S without appealing to "higher-order concepts" are in one way or another unsatisfactory. This leaves us with the following open question:

QUESTION 10.2. Are we implicitly warranted to believe in uniform reflection principles for theories that we already accept in a warranted manner?

10.2.3. Set theoretic reflection. As a result of the work of Koellner, Roberts, and others, it has become clearer over the past decade what set theoretic reflection principles *are*. A distinction can furthermore be made between set theoretic reflection principles that are natural on the one hand, and set theoretic reflection principles that are perhaps less natural on the other hand. Among the natural reflection principles I count Bernays' reflection principle; among the perhaps less natural ones I count the reflection principles that have been proposed by Tait and those that have been proposed by Marshall.[3]

We have seen how Gödel believed that all strong principles of infinity flow from natural set theoretic reflection principles, and how Koellner, in sharp contradistinction, believes that even the Axiom of measurable cardinals does not follow from set theoretic reflection principles. I believe that it has become clear that natural set theoretic reflection principles *at least* entail the existence of 1-extendible cardinals.[4] Nonetheless, Gödel's bold conjecture is still open, i.e., we may ask:

QUESTION 10.3. How much large cardinal strength have the strongest natural set theoretic reflection principles?

Here the work of Marshall and the recent work of McCallum may be of considerable importance. From a philosophical point of view it is, at this point in time, not well enough understood.

Many are sceptical about the prospects of intrinsically justifying large large cardinal axioms and embedding principles. Compared to first-order large cardinal axioms and first-order embedding principles, set theoretic reflection principles pose additional difficulties because of their essential second-order nature (except Montague-Levy reflection, of course). Nonetheless, we have seen how it has been argued that natural set theoretic reflection principles can be intrinsically justified on the basis of large cardinal considerations. But aside the question how richness arguments should be spelled out with some degree of precision, we may pose the following elementary question:[5]

---

[2]See Section 8.6.

[3]See [**Tai05**], [**Mar89**].

[4]See Section 6.5.

[5]"*Out of the mouths of babes and children...*"

Question 10.4. Why should we believe that the set theoretic universe (with its classes) has the required richness to support natural set theoretic reflection principles?

This question has, in my view, not received nearly enough attention.

Lastly, as mentioned in the Introduction, set theoretic reflection also plays a role in potentialist and multiversist theories of sets.[6] But concerning this matter, much work remains to be done, so we may simply ask:

Question 10.5. What is the relation between set theoretic potentialism and set theoretic reflection principles? What is the relation between set theoretic multiversism and set theoretic reflection principles?

10.2.4. Between ontological and epistemic reflection. The following concerns a puzzle that I have not been able to resolve. On the one hand, we have seen in Section 6.6, that there are reflection principles that have characteristics of epistemic reflection principles and of ontological reflection principles. From a mathematical point of view, there is much that we do not yet know about this intermediate area of reflection principles. But these recent results suggest that there may be a *continuum* between epistemic and ontological reflection reflection principles. On the other hand, we have defended the thesis that our epistemic warrant at least for many proof theoretic reflection principles ("implicit commitment") is very different from our warrant for ontological reflection principles (richness or perhaps success arguments).

How can that be? Indeed, we may ask:

Question 10.6. Corresponding to the continuum between epistemic and ontological reflection principles, might there be a continuum between the kinds of warrant for reflection principles?

10.2.5. Unfolding and what is implicit in concepts. In his work on reduction of Gödelian incompleteness and on predicativism, Feferman pursued three strategies.

In early work, Feferman investigated transfinite progressions of formal theories, powered by uniform reflection principles. Later he turned to theories of self-referential truth, such as the theory KF, as a way of capturing predicativist mathematics. We have seen how, at the end of [**Fef91**], Feferman formulated and investigated a schematic version of KF.[7] This served as an important inspiration for his later work with Strahm, in which Feferman studied the operations on objects and predicates that are implicit in the full acceptance of a formal theory.[8]

In the foregoing, we have discussed the epistemological ideas behind Feferman's work on transfinite progressions on formal theories and behind his work on axiomatic self-referential truth in much detail. But I have scarcely given any attention to unearthing the epistemological ideas behind Feferman and Strahm's technically challenging work on unfolding schematic theories. This is a challenging task, which, in my view, philosophers should take up:

---

[6]For the connection between set theoretic reflection and set theoretic potentialism, [**Rei74**] is a seminal article.

[7]See p. 141.

[8]See [**Str18**].

Question 10.7. Given that a person fully accepts a schematic formal theory S, is she somehow implicitly warranted explicitly to accept the schematic unfolding of S? If so, how exactly can she come to do so in a warranted manner?

More in general, the relation between implicit commitment of mathematical theories on the one hand, and implicit commitment of concepts[9], on the other hand, is at present ill-understood.

10.2.6. Probabilistic reflection. In Section 6.7, variants of van Fraassen's probabilistic reflection principle were briefly discussed. Even though such principles are much discussed in formal epistemology, they have not been much investigated in a truly formal setting. Indeed, it has only recently become known that, against a fairly weak background theory of type-free probability, the synchronic version of van Fraassen's reflection principle cannot be consistently added as a new axiom.

Since by now we have a fairly good grip on theories of type-free truth, and the notion of type-free subjective probability is clearly related to the notion of type-free truth, it seems that a basic understanding of type-free subjective probability should be within reach. Moreover, results in type-free truth—in particular, results on theories of *positive* type-free truth—might suggest new versions of van Fraassen's synchronic reflection principles that can consistently be added but are independent of the background probability theory. Moreover, formal understanding might to some extent guide conceptual understanding, so perhaps in this way we can even arrive at *warranted* probabilistic reflection principles.

Nonetheless, at the moment, the field is completely open. Our formal and conceptual understanding of probabilistic reflection seems to me very poor, so we may simply ask:

Question 10.8. Understand probabilistic reflection better!

[9]See Section 9.2.

Bibliography

[AG10] B. Armour Garb, *Horwichian minimalism and the generalization problem*, Analysis **70** (2010), no. 4, 693–703.

[AH78] Kenneth Appel and Wolfgang Haken, *The four-color problem.*, Philosophy of Mathematics: An Anthology (Dale Jacquette, ed.), Blackwell 2002, 1978.

[Ale23] J. W. Alexander, *A lemma on systems of knotted curves*, Proceedings of the National Academy of Sciences **9** (1923), no. 3, 93–95.

[All44] Rudolf Allers, *Microcosmus: From anaximandros to paracelsus*, Traditio **2** (1944), 319 – 407.

[AM10] Marianna Antonutti Marfori, *Informal proofs and mathematical rigour*, Studia Logica **96** (2010), no. 2, 261–272.

[AMH19] Marianna Antonutti Marfori and Leon Horsten, *Human-effective computability?*, Philosophia Mathematica **27** (2019), no. 1, 61–87.

[AN62] Antoine Arnauld and Pierre Nicole, *Logic, or, the art of thinking: Containing, besides common rules, several new observations appropriate for forming judgment. edited and translated by j. v. buroker.*, Cambridge University Press, 1662.

[Ans79] Gertrude Elizabeth Margaret Anscombe, *What is it to believe someone?*, Rationality and Religious Belief (C. F. Delaney, ed.), University of Notre Dame Press, 1979.

[Ari] Aristotle, *n the soul. parva naturalia. on breath.*, Harvard University Press 1989.

[Arn83] Antoine Arnauld, *On true and false ideas. translated by s. gaukroger*, Manchester University Press, 1990 (1683).

[Art08] Sergei Artemov, *The logic of justification*, Review of Symbolic Logic **1** (2008), no. 4, 477–513.

[Aud08] Robert Audi, *Belief, faith, and acceptance*, International Journal for Philosophy of Religion **63** (2008), no. 1-3, 87–102.

[Aug] Augustine, *The city of god (*De civitate Dei*). translation by henry bettenson 1972*, Penguin Books.

[Avi03] Jeremy Avigad, *Number theory and elementary arithmetic*, Philosophia Mathematica **11** (2003), 257–284.

[Avi20] Jeremy Avigad, *Reliability of mathematical inference*, Synthese **198** (2020), no. 8, 7377–7399.

[Azz04] Jody Azzouni, *The derivation-indicator view of mathematical practice*, Philosophia Mathematica **12** (2004), no. 2, 81–106.

[Bag23] Joan Bagaria, *Large cardinals as principles of structural reflection*, Bulletin of Symbolic Logic **29** (2023), 19–70.

[Bar16] Neil Barton, *Richness and reflection*, Philosophia Mathematica **24** (2016), 330–359.

[BBPJ02] George Boolos, John Burgess, Richard P., and C. Jeffrey, *Computability and logic*, Cambridge University Press, 2002.

[Bek95] Lev Beklemishev, *Iterated local reflection versus iterated consistency*, Annals of Pure and Applied Logic **75** (1995), no. 1-2, 25–48.

[Bek05] Lev Beklemishev, *Reflection principles and provability algebras in formal arithmetic*, Russian Mathematical Surveys **60** (2005).

[Ben65] Paul Benacerraf, *What numbers could not be*, Philosophical Review **74** (1965), 47–73.

[Ben67] Paul Benacerraf, *God, the devil, and gödel*, The Monist **51** (1967), no. 1, 9–32.

[Ben73] Paul Benacerraf, *Mathematical truth*, Journal of Philosophy **70** (1973), 661–679.

[Ben16] Sebastian Bender, *Reflection and rationality in leibniz*, Subjectivity and Selfhood in Medieval and Early Modern Philosophy (Jari Kaukua and Tomas Ekenberg, eds.), Springer, 2016, pp. 263–275.

[Ber61] Paul Bernays, *Zur frage der unendlichkeitsschemata in der axiomatische mengenlehre. in y. bar-hillel et al (eds) essays on the foundations of mathematics.*, pp. 3–49, Magnus Press, 1961.

[Bla02] Stephen Blamey, *Partial logic*, Handbook of Philosophical Logic (Dov M. Gabbay and F. Guenthner, eds.), Springer Netherlands, Dordrecht, 2002, pp. 261–353.

[BMP23] Nuel Belnap, Thomas Müller, and Thomasz Placek, *Branching space-times: theory and applications*, Oxford University Press, 2023.

[Boa80] George Boas, *Macrocosm and microcosm*, Dictionary of the History of Ideas Vol. 3 (Philip P. Wiener, ed.), Macmillan, 1980, pp. 126–131.

[Bog03] Paul Boghossian, *Blind reasoning*, Proceedings of the Aristotelian Society. Supplementary Volume **LXXVII** (2003), 225–248.

[Bon85] Lawrence Bonjour, *The structure of empirical knowledge*, Harvard University Press, 1985.

[Boo71] George Boolos, *The iterative conception of set*, Journal of Philosophy **68** (1971), no. 8, 215–231.

[Boo85] ———, *Nominalist platonism*, Philosophical Review **94** (1985), no. 3, 327–344.

[BT23] Joan Bagaria and Claudio Ternullo, *Intrinsic justification for large cardinals and structural reflection*, ArXiv **2310.05841** (2023), 42p.

[Bur79] Tyler Burge, *Individualism and the mental*, Midwest Studies in Philosophy **4** (1979), no. 1, 73–122.

[Bur93] Tyler Burge, *Content preservation*, Philosophical Review **201** (1993), 457–488.

[Bur97] ———, *Interlocution, perception, and memory.*, Philosophical Studies **86** (1997), 21–47.

[Bur98a] ———, *Computer proofs, a priori knowledge, and other minds.*, Philosophical Perspectives **12** (1998), 1–37.

[Bur98b] ———, *Reason and the first person.*, Cognition Through Understanding: Self-Knowledge, interlocution, reasoning, reflection: Philosophical Essays, Volume 3, no. 2, Oxford University Press, 2013(1998), pp. 383–406.

[Bur03] ———, *Perceptual entitlement.*, Philosophy and Phenomenological Research **67** (2003), 503–548.

[Bur11a] ———, *Self and self-understanding*, Journal of Philosophy **108** (2011), 287–383.

[Bur11b] ———, *Epistemic warrant: humans and computers*, Cognition Through Understanding: Self-Knowledge, interlocution, reasoning, reflection: Philosophical Essays, Volume 3, no. 2, Oxford University Press, 2013(2011), pp. 489–507.

[Bur13a] ———, *Cognition through understanding: Self-knowledge, interlocution, reasoning, reflection: Philosophical essays, volume 3*, no. 1, Oxford University Press, 2013.

[Bur13b] ———, *Introduction*, Cognition Through Understanding: Self-Knowledge, interlocution, reasoning, reflection: Philosophical Essays, Volume 3, no. 3, Oxford University Press, 2013, pp. 534–555.

[Bur13c] ———, *Postscript: 'content preservation'*, Cognition Through Understanding: Self-Knowledge, interlocution, reasoning, reflection: Philosophical Essays, Volume 3, no. 2, Oxford University Press, 2013, pp. 534–555.

[Bur13d] ———, *Reflection.*, Cognition Through Understanding: Self-Knowledge, interlocution, reasoning, reflection: Philosophical Essays, Volume 3, no. 1, Oxford University Press, 2013, pp. 534–555.

[Bur13e] ———, *A warrant for belief in other minds*, Cognition Through Understanding: Self-Knowledge, interlocution, reasoning, reflection: Philosophical Essays, Volume 3, Oxford University Press, 2013, pp. 362–379.

[Can32] Georg Cantor, Gesammelte Abhandlungen mathematischen und philosophischen Inhalts. Mit erläuternden Anmerkungen sowie mit Ergänzungen aus dem Briefwechsel Cantor-Dedekind (Ernst Zermelo, ed.), Springer Berlin, Heidelberg, 1932.

[Can89] Andrea Cantini, *Notes on formal theories of truth*, Zeitschrift für mathematische Logik und Grundlagen der Mathematik **35** (1989), 1452–1468.

[Car00] Timothy J. Carlson, *Knowledge, machines, and the consistency of reinhardt's strong mechanistic thesis*, Annals of Pure and Applied Logic **105** (2000), no. 1, 51–82.

[CG23] Luca Castaldo and Maciej Glowacki, *Provably true sentences across axiomatizations of kripke's theory of truth*, submitted for publication (2023).

[Che73] Jingrun Chen, *On the representation of a large even integer as the sum of a prime and a product of at most two primes*, Scientia Sinica **16** (1973), 157–176.

[Chi77] Roderick Chisholm, *Theory of knowledge, second edition*, Englewood Cliffs, NJ, USA: Englewood Cliffs, N.J., Prentice-Hall, 1977.

[CHL22] Cezary Cieslinski, Leon Horsten, and Hannes Leitgeb, *Axioms for typefree subjective probability*, arXiv 10.48550/ARXIV.2203.04879 (2022).

[Cie15] Cezary Cieslinski, *The innocence of truth*, Dialectica **69** (2015), 61–85.

[Cie17] Cesary Cieslinski, *The epistemic lightness of truth*, Cambridge University Press, 2017.

[Cie18] Cezary Cieśliński, *Minimalism and the generalisation problem: On horwich?s second solution*, Synthese **195** (2018), no. 3, 1077–1101.

[Cie20] Cezary Cieśliński, *Believabilitiy theories. corrigendum to: "the epistemic lightness of truth. deflationism and its logic."*, http://cieslinski.filozofia.uw.edu.pl/Corrigendum.pdf, 2020.

[Coh89] Jonathan Cohen, *Belief and acceptance*, Mind **98** (1989), 367–389.

[Côt02] Antoine Côté, *L'infinité dans la théologie médiévale (1220–1255)*, Vrin, 2002.

[Dav67] Donald Davidson, *Truth and meaning*, Synthese **17** (1967), no. 1, 304–323.

[Dav04] Martin Davies, *Epistemic entitlement, warrant transmission and easy knowledge*, Proceedings of the Aristotelian Society. Supplementary Volume **78** (2004), 213–245.

[Dea15] Walter Dean, *Arithmetical reflection and the provability of soundness*, Philosophia Mathematica **23** (2015), no. 1, 31–64.

[Ded88] Richard Dedekind, *Was sind und was sollen die Zahlen?*, 1 ed., Verlag von Friedrich Vieweg und Sohn, Braunschweig, 1888.

[Des41] Rene Descartes, *Meditations on first philosophy/meditationes de prima philosophia: A bilingual edition*, University of Notre Dame Press, 1990 (1641).

[Dev84] Keith Devlin, *Constructibility*, Springer, 1984.

[Dro19] Adam Drozdek, *Infinity in augustine's theology*, The Infinity of God: New Perspectives in Theology and Philosophy (Benedikt Paul Gčke and Christian Tapp, eds.), University of Notre Dame Press, 2019, pp. 37–53.

[dT21a] Silvia de Toffoli, *Groundwork for a fallibilist account of mathematics*, Philosophical Quarterly **7** (2021), no. 4, 823–844.

[DT21b] Silvia De Toffoli, *Reconciling rigour and intuition*, Erkenntnis (2021), no. 86, 1783–1802.

[dTG16] Silvia de Toffoli and Valeria Giardino, *Envisioning transformations ? the practice of topology*, Mathematical Cultures: The London Meetings 2012–2014 (Brendan Larvor, ed.), Zurich, Switzerland: Birkhäuser, 2016, pp. 25–50.

[Dum63] Michael Dummett, *The philosophical significance of gödel's theorem*, Ratio (Michael Dummett, ed.), Duckworth, 1963, pp. 186–214.

[Eas09] Kenny Easwaran, *Probabilistic proofs and transferability*, Philosophia Mathematica **17** (2009), no. 3, 341–362.

[Fef62] Solomon Feferman, *Autonomous transfinite progressions of formal theories*, Journal of Symbolic Logic **27** (1962), 259–316.

[Fef64] ———, *Systems of predicative analysis*, Journal of Symbolic Logic **29** (1964), 1–30.

[Fef88] ———, *Turing in the land of o(z). in r. herken (ed) the universal turing machine: A half-century survey.*, pp. 113–147, Kammerer und Unverzagt, 1988.

[Fef91] ———, *Reflecting on incompleteness*, Journal of Symbolic Logic **56** (1991), 1–49.

[Fef96] Solomon Feferman, *Gödel's program for new axioms: why, where, how and what?*, Lecture Notes in Logic. Gödel '96. Logical foundations of mathematics, computer science and physics—Kurt Gödel's legacy (P. Hájek, ed.), Association for Symbolic Logic, 1996, pp. 3–22.

[Fef05] ———, *Predicativity*, Oxford Handbook of Philosophy of Mathematics and Logic (Stewart Shapiro, ed.), Oxford: Oxford University Press, 2005, pp. 590–624.

[FHN21] Martin Fischer, Leon Horsten, and Carlo Nicolai, *Hypatia's silence. truth, justifcation, and entitlement*, Noûs (2021), 62–85.

[Fie94] Hartry Field, *Deflationist views of meaning and content*, Mind **103** (1994), no. 411, 249–285.

[Fie99] Hartry Field, *Deflating the conservativeness argument*, Journal of Philosophy **96** (1999), 533–540.

[Fie08] Hartry Field, *Saving truth from paradox*, Oxford University Press, 2008.

[Fis23] Martin Fischer, *Another look at reflection*, Erkenntnis **88** (2023), 479–509.

[FN18] Greg Fitch and Michael Nelson, *Singular Propositions*, The Stanford Encyclopedia of Philosophy (Edward N. Zalta, ed.), Metaphysics Research Lab, Stanford University, Spring 2018 ed., 2018.

[FNH17a] Martin Fischer, Carlo Nicolai, and Leon Horsten, *Iterated reflection on full disquotational truth*, Journal of Logic and Computation **27** (2017), 2631–2651.

[FNH17b] ———, *Iterated reflection over full disquotational truth*, Journal of Logic and Computation **27** (2017), no. 8, 2631–2651.

[FNH21] ———, *Hypatia's silence*, Noûs **55** (2021), 62–85.

[Fou66] Michel Foucault, *Les mots et les choses. archéologie des sciences humaines*, Éditions Gallimard, 1966.

[Fra04a] Torkel Franzen, *Inexhaustibility: a non-exhaustive treatment*, Association for Symbolic Logic, 2004.

[Fra04b] ———, *Transfinite progressions: A second look at completeness.*, Bulletin of Symbolic Logic **10** (2004), 367–389.

[Fre84] Gottlob Frege, *Die Grundlagen der Arithmetik: eine logisch-mathematische Untersuchung über den Begriff der Zahl*, The Foundations of Arithmetic: A Logico-Mathematical Enquiry Into the Concept of Number (J. L. Austin, ed.), Blackwell 1953, 1953 (1884).

[Fri06] Elizabeth Fricker, *Martians and meetings: Against burge's neo-kantian apriorism about testimony*, Philosophica **78** (2006), 69–84.

[FS62] Solomon Feferman and Clifford Spector, *Incompleteness along paths in progressions of theories*, Journal of Symbolic Logic **27** (1962), 383–390.

[Fuj11] Kentaro Fujimoto, *Autonomous progression and transfinite iteration of self-applicable truth*, Journal of Symbolic Logic **76** (2011), 914–945.

[Fuj12] ———, *Classes and truths in set theory*, Annals of Pure and Applied Logic **163** (2012), 1484–1523.

[Fuj19] ———, *Predicativism about classes*, Journal of Philosophy (2019), no. 116, 206–229.

[Fuj23] ———, *A few more dissimilarities between second order logic and set theory*, Archive for Mathematical Logic **62** (2023), 147–206.

[G̈46] Kurt Gödel, *Remarks before the princeton bicentennial conference. in: S. feferman et al (eds) kurt gödel. collected works. volume ii, 1990*, pp. 150–153, Oxford University Press, 1946.

[Gai86] Haif Gaifman, *A theory of higher-order probabilities. in: J. halpern (ed) tark '86: Proceedings of the 1986 conference on theoretical aspects of reasoning about knowledge*, pp. 275–292, Morgan Kaufmann, 1986.

[Gal14] Henri Galinon, *Acceptation, cohérence et responsabilité **, Liber Amicorum Pascal Engel, Université de Genève, 2014.

[Gel05] Albert-Kees Geljon, *Divine infinity in gregory of nyssa and philo of alexandria*, Vigiliae Christianae **59** (2005), no. 2, 152–177.

[Get63] Edmund Gettier, *Is justified true belief knowledge?*, Analysis **23** (1963), 121–123.

[Gir87] Jean-Yves Girard, *Proof theory and logical complexity*, Bibliopolis, 1987.

[GL23] Maciej Glowacki and Mateusz Lelyk, *Reflecting on believability. on the epistemic approach to justifying epistemic commitments*, under review (2023), no. , 29p.

[GO24] Victoria Gitman and Jonathan Osinski, *Upward löwenheim-skolem-tarski numbers for abstract logics*, arXiv https://arxiv.org/abs/1501.05438 (2024).

[Göd31] Kurt Gödel, *On formally undecidable propositions of principia mathematica and related systems - i.*, Kurt Gödel. Collected Works. Volume I. (Feferman et al., ed.), Oxford University Press, 1986(1931), pp. 144–195.

[Göd32] ———, *Über vollständigkeit und widerspruchsfreiheit. ergebnisse eines mathematischen kolloquiums, vol. 3, pp. 12-13; text and translation in*, Kurt Gödel. Collected Works. Volume I. (S. Feferman et al., ed.), Oxford University Press, 1986(1932), pp. 234–237.

[Göd33a] ———, *Eine interpretation des intuitionistischen aussagenkalküls*, Kurt Gödel. Collected Works. Volume I. (Feferman et al., ed.), Oxford University Press, 1986(1933), p. 300.

[Göd33b] ———, *The present situation in the foundation of mathematics*, Kurt Gödel. Collected Works. Volume III. (S. Feferman et al, ed.), Oxford University Press, 1995(1933), pp. 45–53.

[Göd47] ———, *What is cantor's continuum problem?*, American Mathematical Monthly **54** (1947), 515–525.

[Göd51] ———, *Some basic theorems on the foundations of mathematics and their implications*, Oxford University Press, 1995(1951), pp. 304–323.

[Göd53] ———, *Is mathematics syntax of language?*, Kurt Gödel. Collected Works. Volume III. (S. Feferman et al, ed.), Oxford University Press, 1995(1953), pp. 334–364.

[Göd19] ———, *Philosophical notebooks*, Kurt Gödel, Philosophical Notebooks. Volume 1, Maxims 0. Herausgegeben von Eva-Maria Engelen (and translated from German by Merlin Carl) (Eva-Maria Engelen, ed.), De Gruyter, Berlin, Boston, 2019.

[Göd21] ———, *Philosophical notebooks*, Kurt Gödel, Philosophical Notebooks. Volume 3, Maxims III. Herausgegeben von Eva-Maria Engelen (and translated from German by Merlin Carl) Maximen III / Maxims III (Eva-Maria Engelen, ed.), De Gruyter, Berlin, Boston, 2021.

[Gol79] Alvin Goldman, *What is justified belief?*, Justification and Knowledge (George Pappas, ed.), Boston: D. Reidel, 1979, pp. 1–25.

[Gra92] Robert Graves, *The greek myths. the complete and definitive edition*, Penguin books, 1992.

[Gra20] Peter Graham, *What is epistemic entitlement? reliable competence, reasons, inference, access*, Virtue-Theoretic Epistemology: New Methods and Approaches (John Greco and Christoph Kelp, eds.), New York, USA: Cambridge University Press, 2020.

[Gup93] Anil Gupta, *Minimalism*, Philosophical Perspectives **7** (1993), 359–369.

[Guy04] Richard Guy, *Unsolved problems in number theory*, Springer, 2004.

[Hal84] Michael Hallett, *Cantorian set theory and limitation of size*, Oxford, England: Clarendon Press, 1984.

[Hal97] Volker Halbach, *Tarskian and kripkean truth*, Journal of Philosophical Logic **26** (1997), 69–80.

[Hal99] ———, *Deflationism and infinite conjunctions*, Mind **108** (1999), 1–22.

[Hal01a] Volker Halbach, *Disquotational truth and analyticity*, Journal of Symbolic Logic **66** (2001), no. 4, 1959–1973.

[Hal01b] ———, *How innocent is deflationism?*, Synthese **126** (2001), no. 1, 167–194.

[Hal05] Thomas C. Hales, *A proof of the kepler conjecture*, Annals of Mathematics **162** (2005), no. 3, 1065–1185.

[Hal09] Volker Halbach, *Reducing compositional to disquotational truth*, Review of Symbolic Logic **2** (2009), no. 4, 786–798.

[Hal11] Volker Halbach, *Axiomatic theories of truth*, Cambridge University Press, 2011.

[Hal14] ———, *Axiomatic theories of truth.*, Cambridge University Press, 2014.

[Hei88] John Heil, *Privileged access*, Mind (1988), no. 97, 238–251.

[Hel92] Glen Helman, *Proofs and epistemic structure*, Proof, Logic, and Formalization (Michael Detlefsen, ed.), Routledge, 1992, p. 24.

[Hel15] Harald Helfgott, *The ternary goldbach problem*, arXiv 10.48550/ARXIV.1501.05438 (2015).

[HH02] Volker Halbach and Leon Horsten, *Contemporary methods for investigating the concept of truth. in: V. halbach and l. horsten (eds), principles of truth.*, pp. 11–35, Dr. Hänsel-Hohenhausen, 2002.

[HH06] ———, *Axiomatizing kripke's theory of truth*, Journal of Symbolic Logic (2006), no. 71, 677–712.

[HH17] Jan Heylen and Leon Horsten, *Truth and existence*, Thought: A Journal of Philosophy **6** (2017), no. 1, 106–114.

[Hil99] David Hilbert, *Grundlagen der geometrie, 14. auflage*, Teubner, 1999(1899).

[Hil26] David Hilbert, *Über das unentliche*, Mathematische Annalen **95** (1926), 161–190.

[Hil96] ———, *The grounding of elementary number theory (1931). in: E.w. ewald (ed) from kant to hilbert. a source book in the foundations of mathematics. volume 2*, pp. 1148–1157, Oxford University Press, 1996.

[HL17] Leon Horsten and Graham Leigh, *Truth is simple*, Mind (2017), no. 126, 195–232.

[HN18] Volker Halbach and Carlo Nicolai, *On the costs of nonclassical logic*, Journal of Philosophical Logic **47** (2018), no. 2, 227–257.

[Hod78] Shadworth Hollway Hodgson, *The philosophy of reflection. two volumes*, Longmans, 1878.

[Hor90] Paul Horwich, *Truth*, Clarendon Press, 1990.

[Hor94] Leon Horsten, *Modal-epistemic variants of shapiros system of epistemic arithmetic*, Notre Dame Journal of Formal Logic **35** (1994), no. 2, 284–291.

[Hor95] ———, *The semantical paradoxes, the neutrality of truth and the neutrality of the minimalist theory of truth*, The Many Problems of Realism (Studies in the General Philosophy of Science: Volume 3) (P. Cartois, ed.), Tilberg University Press, 1995.

[Hor98] Paul Horwich, *Truth*, Oxford University Press, 1998.

[Hor01a] L. Horsten, *Platonistic formalism*, Erkenntnis **54** (2001), no. 2, 173–194.

[Hor01b] Paul Horwich, *A defense of minimalism*, Synthese **126** (2001), no. 1, 149–165.

[Hor05a] Leon Horsten, *Remarks on the content and extension of the notion of provability*, Logique Et Analyse **48** (2005), no. 189-192, 15–32.

[Hor05b] Paul Horwich, *A minimalist critique of tarski on truth*, Deflationism and Paradox (J. C. Beall and Bradley Armour-Garb, eds.), Oxford University Press, 2005.

[Hor09] Leon Horsten, *Levity*, Mind **118** (2009), no. 471, 555–581.

[Hor10] Paul Horwich, *Truth – meaning – reality*, Oxford University Press, 2010.

[Hor11] Leon Horsten, *The tarskian turn: Deflationism and axiomatic truth*, MIT Press, 2011.

[Hor18] Leon Horsten, *Book review: The epistemic lightness of truth. deflationism and its logic. by c. ciésliński*, Notre Dame Philosophical Reviews (2018), .

[HP93] Petr Hajek and Pavel Pudlàk, *Metamathematics of first-order arithmetic*, Springer, 1993.

[Hum40] David Hume, *Philosophical essays concerning human understanding. a treatise of human nature: being an attempt to introduce the experimental method of reasoning into moral subjects.*, John Noon, 1739–40.

[HZng] Leon Horsten and Li Zhang, *The minimalist theory of truth and the generalisation problem*, Dialectica (forthcoming).

[Inc16] Luca Incurvati, *Maximality principles in set theory*, Philosophia Mathematica **25** (2016), 159–193.

[Isa87] Daniel Isaacson, *Arithmetical truth and hidden higher-order concepts*, Logic Colloquium '85 (Paris Logic Group, ed.), North Holland, 1987, pp. 147–69.

[Isa92] ———, *Some considerations on arithmetical truth and the co-rule*, Proof, Logic, and Formalization (Michael Detlefsen, ed.), Routledge, 1992, p. 94.

[Jen07] Carrie Jenkins, *Entitlement and rationality*, Synthese **157** (2007), 25–45.

[Jon98] Vaughan V.R. Jones, *A credo of sorts*, Truth in Mathematics (Harold Garth Dales and Gianluigi Oliveri, eds.), Oxford University Press, 1998.

[Kan87] Immanuel Kant, *Critique of pure reason (1781-1787), trans. kemp smith 1929*, London: MacMillan, 1929 (1781/1787).

[Kan94] Akihiro Kanamori, *The higher infinite*, Springer, 1994.

[Kay91] Richard Kaye, *Models of peano arithmetic*, Oxford University Press, 1991.

[Ket99] Jeffrey Ketland, *Deflationism and tarski's paradise*, Mind **108** (1999), no. 429, 69–94.

[Ket05] ———, *Deflationism and the gödel phenomena: Reply to tennant*, Mind **114** (2005), no. 453, 75–88.

[Ket10] J. Ketland, *Truth, conservativeness, and provability: Reply to cieslinski*, Mind **119** (2010), no. 474, 423–436.

[KKL81] H. Kotlarski, S. Krajewski, and A.H. Lachlan, *Construction of satisfaction classes for nonstandard models*, Canadian Mathematical Bulletin **24** (1981), no. 3, 283–293.

[KL68] Georg Kreisel and Azriel Levy, *Reflection principles and their use for establishing the complexity of formal systems*, Zeitschrift für mathematische Logik und Grundlagen der Mathematik **14** (1968), 97–142.

[Kle38] Stephen Kleene, *On notation for ordinal numbers*, Journal of Symbolic Logic **3** (1938), 150–155.

[KM60] D. Kaplan and R. Montague, *A paradox regained*, Notre Dame Journal of Formal Logic **1** (1960), no. 3, 79–90.

[Koe09] Peter Koellner, *On reflection principles.*, Annals of Pure and Applied Logic **157** (2009), 206–219.

[Kol33] Andrei Nikolajewitsch Kolmogorov, *Grundbegriffe der wahrscheinlichkeitrechnung (ergebnisse der mathematik)*, Foundations of the theory of probability. (Nathan Morrison, ed.), Chelsea Publishing Company, 1956(1933), p. 94.

[Kor12] Hilary Kornblith, *On reflection*, Oxford University Press, 2012.

[Kre60] Georg Krerisel, *Ordinal logics and the characterization of informal concepts of proof*, Proceedings international congress of mathematicians, Edinburgh (1958) (J.A. Todd, ed.), Cambridge University Press, 1960, pp. 289–299.

[Kre67] Georg Kreisel, *Informal rigour and completeness proofs. in i. lakatos (ed) problems in the philosophy of mathematics*, pp. 138–157, North-Holland, 1967.

[Kre68] G. Kreisel, *A survey of proof theory*, Journal of Symbolic Logic **33** (1968), no. 3, 321–388.

[Kre69] Georg Kreisel, *Two notes on the foundations of set-theory*, Dialectica **106** (1969), 93–114.

[Kre70] ———, *Principles of proof and ordinals implicit in given concepts.*, Studies in Logic and the Foundations of Mathematics **60** (1970), 489–516.

[Kri75] Saul Kripke, *Outline of a theory of truth*, Journal of Philosophy **72** (1975), no. 19, 690–716.

[Kri80] Saul A. Kripke, *Naming and necessity: Lectures given to the princeton university philosophy colloquium*, Cambridge, MA: Harvard University Press, 1980.

[Kuh62] Thomas S. Kuhn, *The structure of scientific revolutions*, University of Chicago Press, 1962.

[Kul91] Mark Kulstad, *Leibniz on apperception, consciousness and reflection*, Philosophia, 1991.

[Kun71] Kenneth Kunen, *Elementary embeddings and infinitary combinatorics*, Journal of Symbolic Logic **36** (1971), no. 3, 407–413.

[Kyb70] Henry E. Kyburg, *Conjunctivitis*, Induction, Acceptance, and Rational Belief (Marshall Swain, ed.), Dordrecht: D. Reidel, 1970, pp. 55–82.

[Lak68] Imre Lakatos, *Criticism and the methodology of scientific research programmes*, Proceedings of the Aristotelian Society **69** (1968), no. 1, 149–186.

[Lak76] ———, *Proofs and refutations: The logic of mathematical discovery*, Cambridge and London: Cambridge University Press, 1976.

[Las09] Kathryn Laskey, *Axiomatic first-order probability*, Proceedings of the Fifth Workshop on Uncertainty Reasoning for the Semantic Web (URSW 2009) (2009), 51–62.

[LCF+73] William Little, Jessie Senior Coulson, H. W. Fowler, C. T. Onions, and G. W. S. Friedrichsen, *The shorter oxford english dictionary on historical principles; prepared by william little, h. w. fowler and jessie coulson; revised and edited by c. t. onions*, 3rd. ed.; completely reset with etymologies revised by g. w. s. friedrichsen and with revised addenda. ed., Clarendon Press Oxford, 1973 (English).

[Lei91] G.W. Leibniz, *G.w. leibniz's monadology. (ed. by n. rescher)*, University of Pittsburgh Press, 1991.

[Lei09] Hannes Leitgeb, *On formal and informal provability*, New Waves in Philosophy of Mathematics (Ø. Linnebo O. Bueno, ed.), Palgrave Macmillan, 2009, pp. 263–299.

[Lei15] Graham E. Leigh, *Conservativity for theories of compositional truth via cut elimination*, Journal of Symbolic Logic **80** (2015), no. 3, 845–865.

[Lei16] Graham Leigh, *Truth and reflection*, IfCoLog Journal of Logics and Their Applications **3** (2016), 557–594.

[Lel23] Mateusz Lelyk, *Model theory and proof theory of the global reflection principle*, Journal of Symbolic Logic **88** (2023), 738–779.

[Lév60a] Azriel Lévy, *Axiom schemata of strong infinity in axiomatic set theory.*, Pacific Journal of Mathematics **10** (1960), no. 1, 223 – 238.

[Lev60b] Azriel Levy, *Principles of reflection in axiomatic set theory.*, Fundamenta Mathematicae **49** (1960), 1–10.

[Lew80] David Lewis, *A subjectivist guide to objective chance. in r. carnap and r. jeffrey (eds), studies in inductive logic and probability*, pp. 263–293, University of California Press, 1980.

[LM08] Benedikt Löwe and Thomas Müller, *Mathematical knowledge is context dependent*, Grazer Philosophische Studien **76** (2008), no. 1, 91–107.

[LN13] Graham Leigh and Carlo Nicolai, *Axiomatic truth, syntax and metatheoretic reasoning*, Review of Symbolic Logic **4** (2013), 613–626.

[LN22] Mateusz Łełyk and Carlo Nicolai, *A theory of implicit commitments for mathematical theories*, Synthese **200** (2022), article number 284.

[Loc89] John Locke, *An essay concerning human understanding*, Clarendon Edition of the Works of John Locke (Peter Nidditch, ed.), Oxford University Press UK, 1975 (1689).

[Lor58] Paul Lorenzen, *Logical reflection and formalism*, Journal of Symbolic Logic **23** (1958), 241–249.

[LS67] Azriel Levy and Robert Solovay, *Measurable cardinals and the continuum hypothesis*, Israel Journal of Mathematics **5** (1967), 234–248.

[Luc61] John R. Lucas, *Minds, machines and gödel*, Philosophy **36** (1961), no. 137, 112–127.

[Mad88] Penelope Maddy, *Believing the axioms.*, Journal of Symbolic Logic **53** (1988), no. 2, 481–511 and 736–764.

[Mad07] Penelope Maddy, *Second philosophy: a naturalistic method*, Oxford University Press, 2007.

[Mag71] Menachem Magidor, *On the role of supercompact and extendible cardinals in logic*, Israel Journal of Mathematics **28** (1971), 147–157.

[Mar76] Donald Martin, *Hilbert's first problem: the continuum hypothesis, in: F. e. browder (ed). mathematical developments arising from hilbert problems. proceedings of symposia in pure mathematics, vol. 28*, pp. 81–92, American Mathematical Society, 1976.

[Mar89] Victoria Marshall, *Higher-order reflection principles*, Journal of Symbolic Logic **54** (1989), 474–489.

[McC21] Rupert McCallum, *Intrinsic justifications for large-cardinal axioms*, Philosophia Mathematica **29** (2021), 195–213.

[McG92] Vann McGee, *Maximal consistent sets of instances of tarski?s schema*, Journal of Philosophical Logic **21** (1992), no. 3, 235–241.

[McG97] ———, *How we learn mathematical language*, Philosophical Review **106** (1997), no. 1, 35–68.

[Men12] Stephen Menn, *Self-motion and reflection: Hermeias and proclus on the harmony of plato and aristotle on the soul*, Neoplatonism and the Philosophy of Nature (James Wilberding and Christoph Horn, eds.), Oxford Up, 2012, pp. 44–67.

[Mon61] Richard Montague, *Fraenkel's addition to the axioms of zermelo. in y. bar-hillel et al (eds) essays on the foundations of mathematics.*, pp. 91–114, Magnus Press, 1961.

[Moo39] George Edward Moore, *Proof of an external world*, Proceedings of the British Academy **25** (1939), 273–300.

[Mor05] Richard Moran, *Getting told and being believed*, Philosophers' Imprint **5** (2005), 1–29.

[MR20] Julien Murzi and Lorenzo Rossi, *Conservative deflationism?*, Philosophical Studies (2020), no. 177, 535–549.

[MV11] Menachem Magidor and Jouko Väänänen, *On löwenheim-skolem-tarski numbers for extensions of first order logic*, Journal of Mathematical Logic **11** (2011), 87–113.

[Myh60] John Myhill, *Some remarks on the concept of proof.*, Journal of Philosophy **57** (1960), 461–471.

[Nel86] Edward Nelson, *Predicative arithmetic*, Princeton University Press, 1986.

[Nel11] Edward Nelson, *Warning signs of a possible collapse of contemporary mathematics. in: M. heller and w.h. woodin (eds), infinity. new research frontiers*, pp. 76–85, Cambridge University Press, 2011.

[Nic13] Carlo Nicolai, *Truth, deflationism, and the ontology of expressions: An axiomatic study*, Ph.D. thesis, Oxford University, 2013.

[Nic18] Carlo Nicolai, *Provably true sentences across axiomatizations of kripke's theory of truth*, Studia Logica **106** (2018), 101–130.

[NP19] Carlo Nicolai and Mario Piazza, *The implicit commitment of arithmetical theories and its semantic core*, Erkenntnis **84** (2019), no. 4, 913–937.

[Par71] Rohit Parikh, *Existence and feasibility in arithmetic*, Journal of Symbolic Logic **36** (1971), no. 3, 494–508.

[Par90] Charles Parsons, *Introductory note to 1946.*, Kurt Gödel. Collected Works. Volume II: Publications 1938–1974 (S. Feferman et al, ed.), Oxford University Press, 1990, pp. 263–299.

[Par07] ———, *Mathematical thought and its objects*, Cambridge University Press, 2007.

[Pas07] Alexander Paseau, *Boolos on the justification of set theory*, Philosophia Mathematica **15** (2007), no. 1, 30–53.

[Pas15] ———, *Knowledge of mathematics without proof*, British Journal for the Philosophy of Science **66** (2015), no. 4, 775–799.

[Pea04] Christopher Peacocke, *The realm of reason*, Oxford University Press, 2004.

[Pea19] Kenneth L. Pearce, *Locke, arnauld, and abstract ideas*, British Journal for the History of Philosophy **27** (2019), no. 1, 75–94.

[Pen89] Roger Penrose, *The emperor's new mind*, Oxford University Press, 1989.

[Pen94] ———, *Shadows of the mind*, Oxford University Press, 1994.

[PH77] Jeff Paris and Leo Harrington, *A mathematical incompleteness in peano arithmetic*, Handbook of Mathematical Logic (J. Barwise, ed.), North-Holland, 1977, pp. 1133–1142.

[Pra83] Dag Prawitz, *Proofs and the meaning and completeness of the logical constants. in: J. hintikka (ed), essays in mathematical and philosophical logic.*, pp. 25–39, Reidel, 1983.

[Pry04] James Pryor, *What is wrong with moore's argument?*, Philosophical Issues **4** (2004), 349–378.

[Put71] Hilary Putnam, *Philosophy of logic*, London: Allen & Unwin, 1971.

[PW21] Fedor Pakhomov and James Walsh, *Reducing ω-model reflection to iterated syntactic reflection*, arXiv https://arxiv.org/abs/2103.12147 (2021).

[Qui36] W. V. Quine, *Truth by convention*, Philosophical Essays for Alfred North Whitehead, London: Longmans, Green & Co., 1936, pp. 90–124.

[Qui54] Willard V.O. Quine, *Carnap and logical truth*, The ways of paradox and other essays. Revised and enlarged edition, Harvard University Press, 1976, 1954, pp. 107–132?

[Qui69] W.V. Quine, *Epistemology naturalized. in: Ontological relativity and other essays*, pp. 69–90, Columbia University Press, 1969.

[Qui86] Willard V.O. Quine, *Philosophy of logic. second edition*, Harvard University Press, 1986.

[Raa05] Panu Raatikainen, *On horwich's way out*, Analysis **65** (2005), no. 3, 175–177.

[Rad66] Colin Radford, *Knowledge—by examples*, Analysis **27** (1966), no. 1, 1–11.

[Rav99] Yehuda Rav, *Why do we prove theorems?*, Philosophia Mathematica **7** (1999), no. 1, 5–41.

[Rei74] William Reinhardt, *Remarks on reflection principles, large cardinals, and elementary embeddings. in proceedings of symposia in pure mathematics. volume 10.*, pp. 189–205, American Mathematical Society, 1974.

[Rei85] William N. Reinhardt, *Absolute versions of incompleteness theorems*, Noûs **19** (1985), no. 3, 317–346.

[Rei86] William Reinhardt, *Some remarks on extending and interpreting theories with a partial predicate for truth*, Journal of Philosophical Logic **15** (1986), 219–251.

[Rob] Sam Roberts, *Reflection principles: a survey.*

[Rob17] ———, *A strong reflection principle*, Review of Symbolic Logic **10** (2017), no. 4, 651–662.

[Rot97] Giancarlo Rota, *The phenomenology of mathematical proof*, Synthese **111** (1997), no. 2, 183–196.

[Rou] Sherrilyn Roush, Stanford Encyclopedia of Philosophy.

[Ryl49] Gilbert Ryle, *The concept of mind*, University of Chicago Press, 1949.

[Sch64] Kurt Schütte, *Eine grenze für die beweisbarkeit der transfiniten induktion in der verzweigten typenlogik*, Archive for Mathematical Logic **7** (1964), no. 1-2, 45–60.

[Sch65] Kurt Schütte, *Predicative well-orderings. in: J. crossley and m. dummett (eds), studies in the logic and the foundations of mathematics. volume 40*, pp. 280–303, North-Holland, 1965.

[Sch79] Ulf Schmerl, *A fine structure generated by reflection formulas over primitive recursive arithmetic*, Logic Colloquium '78 (M. Boffa, d. van Dalen, and K. McAloon, eds.), North-Holland, 1979, pp. 335–350.

[Sch94] Ralph-Dieter Schindler, *A dilemma in the philosophy of set theory*, Notre Dame Journal of Formal Logic **35** (1994), 458–463.

[Sch98] Richard Schantz, *Was Tarski a deflationist?*, Logic and Logical Philosophy **6** (1998), 157–172.

[Sch08] Kevin Scharp, *Locke's theory of reflection*, British Journal for the History of Philosophy **16** (2008), no. 1, 25–63.

[Sch19] Eric Schwitzgebel, *Introspection*, Stanford Encyclopedia of Philosophy (Edward N. Zalta, ed.), Metaphysics Research Lab, Stanford University, 2019.

[Seg77] Alan Segal, *Two powers in heaven. early rabbinic reports about christianity and gnosticism*, Brill, 1977.

[SH22] Daniela Schuster and Leon Horsten, *On the pure logic of justified belief*, Synthese **200** (2022), article number 425.

[Sha85] Stewart Shapiro, *Intentional mathematics*, Elsevier, 1985.

[Sha91] ———, *Foundations without foundationalism: A case for second-order logic*, Oxford, England: Oxford University Press, 1991.

[Sha98] ———, *Proof and truth: Through thick and thin*, Journal of Philosophy **95** (1998), no. 10, 493–521.

[Sha99] ———, *Do not claim too much: Second-order logic and first-order logic*, Philosophia Mathematica **7** (1999), no. 1, 42–64.

[Sho77] Joseph R. Shoenfield, *Axioms of set theory*, Handbook of Mathematical Logic (Jon Barwise and H. Jerome Keisler, eds.), North-Holland Pub. Co., 1977, p. 90.

[Sim11] Stephen Simpson, *Subsystems of second order arithmetic*, Springer, 2011.

[Sko23] Thoralf Skolem, *The foundations of elementary arithmetic established by means of the recursive mode of thought without the use of apparent variables ranging over infinite domains*, From Frege to Gödel. A source book in mathematical logic, 1879–1931 (J. van Heijenoort, ed.), Harvard University Press, 1967(1923), pp. 302–333.

[Smo77] Craig Smorynski, *The incompleteness theorems*, Handbook of Mathematical Logic (Jon Barwise, ed.), North-Holland, 1977, pp. 821–865.

[Sor07] Richard Sorabji, *Porphyry on self-awareness, true self, and individual*, Bulletin of the Institute of Classical Studies. Supplement (2007), no. 98, 61–69.

[Spo12] Wolfgang Spohn, *The laws of belief. ranking theory and its philosophical applications*, Oxford University Press, 2012.

[Sta74] Robert Stalnaker, *Pragmatic presuppositions*, Semantics and Philosophy (Milton Karl Munitz and Peter K. Unger, eds.), New York University Press, 1974, pp. 197–214.

[Str50] Peter Strawson, *On referring*, Mind **59** (1950), no. 235, 320–344.

[Str85] ———, *Skepticism and naturalism: Some varieties, the woodbridge lectures 1983*, Columbia University Press, 1985.

[Str00] Thomas Strahm, *Autonomous fixed point progressions and fixed point transfinite recursion. in: S. buss (ed) logic '98. collected works. lecture notes in logic.*, pp. 449–464, A K Peters, 2000.

[Str18] ———, *Unfolding schematic systems. in: G. jäger and w. sieg (eds) feferman on foundations: logic, mathematics, philosophy*, pp. 187–208, Springer, 2018.

[Tai81] William Tait, *Finitism*, Journal of Philosophy **78** (1981), 524–546.

[Tai05] ———, *Constructing cardinals from below. in: W. tait, the provenance of pure reason: essays in the philosophy of mathematics and its history*, pp. 133–154, Oxford University Press, 2005.

[Tar44] Alfred Tarski, *The semantic conception of truth and the foundations of semantics*, Philosophy and Phenomenological Research **4** (1944), no. 3, 341–376.

[Tar83] Alfred Tarski, *The concept of truth in formalised languages (1935). in: A. tarski, logic, semantics, metamathematics.*, pp. 152–278, Hackett, 1983.

[TB20] Oliver Tatton Brown, *Rigour, proof and soundness*, Ph.D. thesis, University of Bristol, 2020.

[TBng] ———, *Rigour and proof*, Review of Symbolic Logic (forthcoming), 1–29.

[Ten02] Neil Tennant, *Deflationism and the gödel phenomena*, Mind **111** (2002), no. 443, 551–582.

[Ten05] ———, *Deflationism and the gödel phenomena: Reply to ketland*, Mind **114** (2005), no. 453, 89–96.

[Thi94] Udo Thiel, *Zur diskussion. leibniz and the concept of apperception*, Archiv für Geschichte der Philosophie **76** (1994), no. 2, 195–219.

[Thi11] ———, *The early modern subject: Self-consciousness and personal identity from descartes to hume*, Oxford University Press, 2011.

[Thu94] William P. Thurston, *On proof and progress in mathematics*, Bulletin of the American Mathematical Society **30** (1994), 161–177.

[TS96] Anne Troelstra and Helmut Schwichtenberg, *Basic proof theory*, Cambridge University Press, 1996.

[Tur39] Alan Turing, *Systems of logic based on ordinal numbers.*, Proceedings of the London Mathematical Society **45 (ser. 2)** (1939), 161–228.

[Tym79] Thomas Tymoczko, *The four-color problem and its philosophical significance*, Journal of Philosophy **76** (1979), no. 2, 57–83.

[vA09] Mark van Atten, *Monads and sets. on gödel, leibniz, and the reflection principle. in g. primiero and s. rahman (eds), judgement and knowledge. papers in honour of b.g. sundholm.*, pp. 3–33, College Publications, 2009.

[Vau16] Robert C. Vaughan, *Goldbach's conjectures: A historical perspective*, Open Problems in Mathematics (John Forbes Nash, Jr. and Michael Th. Rassias, eds.), Springer International Publishing, 2016, pp. 479–520.

[vdVH13] Joanna van der Veen and Leon Horsten, *Intrinsic justifications for large-cardinal axioms*, European Journal for the Philosophy of Religion **5** (2013), 117–138.

[vF80] Bas van Fraassen, *The scientific image*, Clarendon Press, 1980.

[vF84] ———, *Belief and the will*, Journal of Philosophy **81** (1984), 235–256.

[vF89] ———, *Laws and symmetry*, Oxford, England: Oxford University Press, 1989.

[vF95] ———, *Belief and the problem of ulysses and the sirens*, Philosophical Studies **77** (1995), 7–37.

[Wan61] Hao Wang, *Process and existence in mathematics*, Essays on the foundations of mathematics (Bar-Hillel et al., ed.), Magnus Press, 1961, pp. 328–351.

[Wan96] ———, *A logical journey. from gödel to philosophy*, MIT Press, 1996.

[Wax17] Daniel Waxman, *Deflationism, arithmetic, and the argument from conservativeness*, Mind **126** (2017), 429–463.

[Wes03] Kenneth R. Westphal, *Epistemic reflection and cognitive reference in kant's transcendental response to skepticism*, Kant Studien **94** (2003), no. 2, 135–171.

[Wey18] Hermann Weyl, *Das kontinuum. kritische untersuchungen über die grundlagen der analysis*, Veit, 19818.

[WH16] Philip Welch and Leon Horsten, *Reflecting on absolute infinity*, Journal of Philosophy **113** (2016), 89–111.

[Wil78] Bernard Williams, *Descartes: The project of pure enquiry*, Pelican, 1978.

[Wil95] Andrew Wiles, *Modular elliptic curves and fermat's last theorem*, Annals of Mathematics **141** (1995), no. 3, 443–551.

[Wil00] Timothy Williamson, *Knowledge and its limits*, Oxford University Press, 2000.

[Wil01] Michael Williams, *Problems of knowledge: A critical introduction to epistemology*, Oxford University Press, 2001.

[Wil07] Timothy Williamson, *The philosophy of philosophy*, Wiley-Blackwell, 2007.

[Wit22] Ludwig Wittgenstein, *Tractatus logico-philosophicus logisch-philosophische abhandlung*, Kegan Paul, 1922.

[WL17] Bartosz Wcisło and Mateusz Łełyk, *Notes on bounded induction for the compositional truth predicate*, Review of Symbolic Logic **10** (2017), 455–480.

[Wol47a] Harry Austryn Wolfson, *Philo: Foundations of religious philosophy in judaism, christianity, and islam vol. i*, Harvard University Press, 1947.

[Wol47b] ———, *Philo: Foundations of religious philosophy in judaism, christianity, and islam vol. ii*, Harvard University Press, 1947.

[Woo98] Hugh Woodin, *The tower of hanoi*, Truth in mathematics (H. Dales and G. Oliveri, eds.), Oxford University Press, 1998, pp. 329–351.

[Woo01] Hugh Woodin, *The continuum hypothesis. i*, Notices of the American Mathematical Society **48** (2001), 567–576.

[Woo17] ———, *In search of ultimate-l: the 19th midrasha mathematicae lectures*, Bulletin of Symbolic Logic **23** (2017), 1–109.

[Wri94] Crispin Wright, *About "the philosophical significance of gödel's theorem": Some issues*, The Philosophy of Michael Dummett (Brian McGuiness and Gianluigi Oliveri, eds.), Kluwer Academic Publishers, 1994, pp. 167–202.

[Wri02] Crispin Wright, *(anti-)sceptics simple and subtle: G.e. moore and john mcdowell*, Philosophy and Phenomenological Research **65** (2002), 330–348.

[Wri04a] ———, *Intuition, entitlement and the epistemology of logical laws*, Dialectica **58** (2004), 155–175.

[Wri04b] ———, *Warrant for nothing (and foundations for free)?*, Proceedings of the Aristotelian Society. Supplementary Volume **78** (2004), 167–212.

[Wri12] ———, *Replies part iv*, Mind, meaning, and knowledge. Themes from the philosophy of Crispin Wright. (A. Coliva, ed.), Oxford University Press, 2012, pp. 451–486.

[Zic23] Matteo Zicchetti, *Cognitive projects and the trustworthiness of positive truth*, Erkenntnis **88** (2023), 3527–3550.