Author:
Zicchetti, Matteo

Title:
The epistemology of meta-theoretic properties of mathematical theories

consistency, soundness, categoricity

The Epistemology of Meta-theoretic Properties of Mathematical Theories:

Consistency, Soundness, Categoricity

By

Matteo Zicchetti

Department of Philosophy

University of Bristol

A dissertation submitted to the University of Bristol in accordance with the requirements of the degree of Doctor of Philosophy in the Faculty of Arts.

July 2022

Word count: 55496

Abstract

This dissertation investigates the epistemology of three meta-theoretic properties of mathematical theories: *consistency*, *soundness* and *categoricity*. These properties are essential for the integrity and significance of (many) inquiries into mathematical subject matters and are of great importance in the philosophy of mathematics.

Although these meta-theoretic properties have been extensively investigated in mathematical and philosophical logic, they have received little of their well-deserved attention in the epistemology of mathematics. This dissertation aims to make some progress concerning some important epistemological issues about these meta-theoretic properties.

This thesis is divided into three parts: Part I focuses on the epistemology of consistency, investigating issues concerning belief and knowledge of consistency statements. Part II investigates the role of soundness statements for the coherence of mathematical inquiries. Finally, Part III provides an investigation of categoricity.

Acknowledgements

I am glad to have the opportunity to thank those people who supported me and influenced me during the past four years. This journey hasn't been easy for many reasons: from the intrinsic solitude and uncertainty of research to a global pandemic and several mental health issues. However, I feel *privileged* to find myself in the current situation.

The content of this dissertation has been presented at different seminars and conferences online but also in Munich, Bristol, Paris, Gdansk, Moscow, Urbino and Novara. I am grateful to all the participants for their feedback and questions.

I am thankful to my (present and past) supervisors: Leon Horsten, Catrin Campbell-Moore, Kentaro Fujimoto, Johannes Stern and Philip Welch. Their continuous support, kindness, expertise and supervision made me into the researcher I am today.

In particular, I am deeply indebted to Leon Horsten for supporting me from the beginning of this PhD and being an excellent guide. Moreover, chapter 4 is based on our joint paper (Horsten and Zicchetti, 2021) and our numerous discussions have influenced chapter 5. I am grateful to Kentaro Fujimoto for discussing issues in the

philosophy of mathematics and proof-theory with me and for his patience and support during the – nearly endless – revisions to my article (Zicchetti, 2022a), which I revised as chapter 5. I owe my gratitude to Catrin Campbell-Moore for her supervision and support during the later stages of this PhD, for repeatedly discussing this dissertation with me, and for always providing critical yet supporting feedback. I am also indebted to Martin Fischer for the numerous discussions and collaboration, which resulted in the article (Fischer and Zicchetti, 2022), revised as chapter 6.

Finally, I am deeply grateful to Crispin Wright, who agreed to be my advisor during my visiting stay at NYU. The ideas that led to my work in chapters 1, 2 and 3 originated during our discussions.

I want to express my gratitude to other colleagues and philosophers for their invaluable philosophical exchanges, for the collaboration (and for the support!) during my PhD: Nikolaj Pedersen, Marianna Antonutti, Martin Fischer, Dan Waxman, Henri Galinon, Benedict Estaugh, Hannes Leitgeb, Graham Leigh, Volker Halbach, Julien Murzi and Hartry Field. I would also like to thank a couple of friends for bearing with me: Francesca, Theo and Simone, Matteo e Nicola.

I am grateful to the South, West and Wales Doctoral Training Partnership and the Arts and Humanities Research Council for generously funding this research project (grant reference AH/L503939/1).

I dedicate this dissertation to my wife, Franziska, who has (for the better and the worse) been with me through this journey.

Author's declaration

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's Regulations and Code of Practice for Research Degree Programmes and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

Signed: Date: 13.07.2022

Publications

Some of the ideas discussed and employed in chapter 2 are based on ideas introduced and developed in (Zicchetti, 2022b), my article "The Moderate Conception, Other Minds and Knowledge Closure" (under review). I am the sole author of this article.

Chapter 4 is a revised version of the article "Truth, Reflection, and Commitment", published in the edited collection *Modes of Truth. The Unified Approach to Truth, Modalities, and Paradox* (Horsten and Zicchetti, 2021). The original article is co-authored with Leon Horsten. Although chapter 4 is a revised version of the article (Horsten and Zicchetti, 2021), revisions have been kept minimal: I adapted the phrasing and notation to the context of this dissertation, and when relevant, I considered new literature. However, the formulation of the surveyed technical results has not changed. The published article is a result of equal collaboration.

Chapter 5 is a revised version of the article "Cognitive Projects and the Trustworthiness of Positive Truth", published in *Erkenntnis* (Zicchetti, 2022a). I kept the revisions to a minimum. The article's final version results from discussions with my supervisors and colleagues during the first two years of my PhD in Bristol.

Chapter 6 is a revised version of the article (Fischer and Zicchetti, 2022), "Truth, Categoricity and Determinateness' (under review), co-authored with Martin Fischer. This article results from a long-term collaboration and many discussions with Martin during the third year of my PhD. The article under review is a result of equal collaboration.

I assume full responsibility for the claims and possible errors or mistakes in the present dissertation.

Contents

Introduction

Agents engage in a plethora of inquiries into different mathematical subject matters. In such inquiries in mathematics, agents usually also employ mathematical theories. In doing so, they rely on (either implicitly or explicitly and for the moment informally) the presupposition that the theories employed in such inquiries are in *good epistemic standing*. The thought that theories are in good epistemic standing can be understood, for instance, as the thought that such theories are *reliable*. Furthermore, reliability can be spelled out – both formally and informally – in many ways. This dissertation investigates the epistemology of three meta-theoretic properties that (at least partially) capture the reliability of theories: *consistency*, *soundness* and *categoricity*. Most inquiries in mathematics rely at least on the consistency of the theories employed. Other investigations sometimes rely on the stronger assumption of the soundness of such theories, i.e., that everything provable in the theory is also true. Another important meta-property of (some) mathematical theories is that such theories are (in a sense to be made precise much later in this dissertation) categorical: that they are about a particular subject matter. Propositions expressing these meta-theoretic properties are essential for the integrity and significance of our mathematical investigations and of great interest in the philosophy of mathematics. These three meta-theoretic properties have received extensive attention in

both philosophical and mathematical logic and the philosophy of mathematics and have generated several debates in the literature. This dissertation investigates some relevant epistemological issues concerning these meta-theoretic properties. This dissertation is divided into three parts, each focusing on a separate meta-theoretic property: Part I is a broad investigation into the epistemology of consistency, whereas Part II focuses on a specific issue in the epistemology of soundness statement, and Part III focuses on categoricity.

This dissertation focuses on four different *topics* (to be introduced shortly). Although these topics are related to one another, the literature on these topics generated separate and self-standing philosophical debates. For this reason, these four topics are kept separate and self-contained. Some are related to consistency, some to soundness, and others to categoricity. The presentation of the topics follows and mentions the main literature and debates within each context. The following sections introduce and discuss four broad issues:

(A) Implicit acceptability of reflection principles: the case of consistency.

(B) Soundness arguments for consistency.

(C) Epistemic value of global reflection principles.

(D) Internal Categoricity and determinacy.

This introduction presents the key ideas of each topic together with this dissertation's claims. Therefore, this presentation is going to be rather superficial. A detailed presentation of the topics and this dissertation's approach can be found in the summaries of Parts I, II and III.

(A) Implicit acceptability of reflection principles: the case of consistency

The process of extending axiomatic theories via so-called proof-theoretic reflection principles has been investigated in proof theory ever since Gödel 'discovered' the incompleteness phenomena. One of the first aims of such investigations was to 'limit' or 'circumvent' the incompletability of arithmetic (and mathematics). To my best knowledge, this process started with Turing (1939), and continued, for instance, with Feferman (1962) and Feferman and Spector (1962).[1] After that, philosophers investigated the extension of axiomatic theories via the addition of reflection principles in connection with many – and sometimes unrelated – questions in the philosophy of mathematics and philosophy of language. One important example is Feferman's investigation of *predicativity*.[2] More recently, the process of extending theories via reflection principles has received considerable attention in the context of axiomatic theories of truth: this has received some attention within the investigation of *disquotationalism* about truth, where disquotationalism roughly claims that disquotational principles for truth are somewhat more basic than so-called compositional principles for truth. Philosophers aimed to justify compositional principles for truth, employing reflection principles, starting with a (base) disquotational theory. Examples of such investigations are for instance (Halbach, 2001, 2009), (Fischer et al., 2017) and (Horsten and Leigh, 2017).[3] Additionally, philosophers investigated reflection principles in the context of *deflationism* about truth, particularly concerning the conservativeness challenge of deflationism. This debate generated a considerable

[1] We will briefly survey some of these results in chapter 4.

[2] This has been investigated for instance in (Feferman, 1964) and (Schütte, 1964, 1965). An investigation of predicativity or predicative acceptability would exceed the scope of this dissertation.

[3] We will say more about this in chapter 4.

amount of literature and discussions.[4] However, the main interest of this dissertation is not deflationism about truth but a (to some extent) purely epistemological issue about reflection principles.

This work will be related to the so-called *Implicit Commitment Thesis*, which Solomon Feferman championed. Informally, this thesis claims that the acceptance of a theory **S** implies the acceptability of principles formulated in the language of **S** even if such principles are logically independent of **S**. Such principles are for instance consistency statements and other reflection principles. Feferman claimed instances of this thesis in different places. He claimed that reflection principles should be understood as expressing *trust* in axiomatic theories:

> 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. (Feferman, 1962, p. 261)

Additionally, Feferman believed that reflection principles are *implicitly acceptable*:

> 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. (Feferman, 1991, p. 2)

[4] As we will point out later, this dissertation focuses on general epistemological issues. For a glimpse into the debate about deflationism and conservativeness, see (Ketland, 1999, 2005, 2010), (Tennant, 2002, 2005, 2010), (Cieśliński, 2010, 2017b, 2018). See also (Field, 1999) and more recently (Nicolai, 2015), (Waxman, 2017), (Fujimoto, 2017, 2021), (Picollo and Schindler, 2021) and (Murzi and Rossi, 2018). For an introduction into the issue of deflationism about truth, see for instance (Horwich, 1998) and (Horsten, 2011).

Logicians and philosophers have been quite successful in analysing what is implicit in accepting such principles using proof-theoretic methods. Philosophers started to focus on purely epistemological issues concerning the implicit commitment thesis.[5] However, this epistemological interest is relatively new, and much more work is needed to make substantial philosophical progress. Part I of this dissertation focuses on several issues in the epistemology of consistency and is divided into three chapters.

Chapters 1 and 2 focus on presenting a non-evidentialist epistemology of mathematics as a general framework to investigate the epistemology of consistency. Chapter 1 (broadly and for now informally) focuses on the issue of *justifying* propositions such as consistency statements. This chapter argues for the following claims:

1. The warrant to believe the consistency of our accepted theories in *epistemic foundational projects* is an *entitlement*.

2. Entitlement constitutes an *epistemic obligation* to believe in the consistency of the relevant theories.[6]

Chapter 2 focuses on the additional issue arising within entitlement-based epistemological theories: whether such theories can *coherently claim* that propositions warranted through entitlement cannot be known. Part of chapter 2 will be devoted to presenting the intuition behind the claim that entitled propositions cannot be known. After that, the chapter argues for (and defends) the following claim (the details of this issue will be discussed in chapter 2):

[5]Some examples of this investigation are for instance (Franzén, 2004a), (Galinon, 2014), (Dean, 2014), (Horsten and Leigh, 2017), (Fischer et al., 2017, 2019), (Nicolai and Piazza, 2018) and (Horsten, 2021).

[6]That is, entitlement is not simply an *epistemic permission* to believe the consistency of the relevant theories.

3. It is coherent to claim that entitled mathematical propositions cannot constitute mathematical knowledge.[7]

These claims – to be made fully explicit later in the dissertation – are in the spirit of some recent work done in the epistemology of mathematics (Shapiro, 2004, 2011), (Galinon, 2014), (Pedersen, 2016), (Wright, 2016) and (Horsten, 2021).

(B) Soundness Arguments for Consistency

An additional issue within the context of the epistemology of consistency pertains to the so-called *soundness arguments* for consistency. Such arguments aim at inferring the consistency of a theory **S**, employing a soundness claim for **S** as one of the argument's premises. Although these arguments are generally accepted as *valid* (or even *sound*), philosophers share the intuition that such arguments are *epistemically defective*. Chapter 3 focuses on the issue of whether soundness arguments are *cogent*. Roughly – more detail in chapter 3 – a valid argument D is cogent, just in case, if D's premises are justified, then agents can in principle acquire a warrant to believe D's conclusion in virtue of D's premises being warranted and D being valid. As we will show, the answer to the issue of cogency will depend on the background epistemological position concerning the *superstructure* of warrant. We will present the two central positions with respect to this topic: *conservativism* and *liberalism*, and argue for the following:

4. *Conservativism* evaluates soundness arguments as non-cogent.

5. *Liberalism* evaluates soundness arguments as cogent.

[7] I investigate and defend this claim within the epistemology of perception in (Zicchetti, 2022b). Some of the ideas in that article are used in chapter 2.

(C) Epistemic value of global reflection principles

Quite recently, philosophers have been discussing the role of particular soundness statements called *global reflection principles* in expressing the *trustworthiness* of theories and conforming to *norms* of *epistemic practices*. In particular, it has been argued by Fischer et al. (2017, 2019) that theories of truth formulated in (some suitable) non-classical logic are trustworthy because (in some sense to be made explicit later) they are compatible with their global reflection principle. Fischer et al. (2019) argued that such theories are epistemically superior to (some) theories of truth formulated over classical logic. The second part of this dissertation focuses on this issue and is divided into two parts. Chapter 4 introduces the relevant philosophical context around soundness statements and global reflection principles by briefly surveying some of the relevant results in mathematical and philosophical logic about extensions of theories by soundness statements. This chapter is a revised version of (Horsten and Zicchetti, 2021). Chapter 5 focuses explicitly on axiomatic theories of truth and the epistemic role that global reflection principles play in expressing trustworthiness. The work of this chapter is closely related to the work provided by Leigh (2016), Horsten and Leigh (2017), and Fischer et al. (2017, 2019). Following (and expanding on) (Fischer et al., 2017, 2019), I investigate two questions: (a) whether there are acceptable theories (of truth) in classical logic that are trustworthy; (b) what is the role of trustworthiness in relation to specific epistemic norms. This chapter argues for the following claims:

6. Theories of positive truth (and falsity) formulated over classical logic are trustworthy.

7. To believe untrustworthy theories is *epistemically blameworthy*.

The first claim contributes to the work by Fischer et al. (2017), where the authors provided trustworthy theories of truth formulated over some weak non-classical logic. Moreover, this work defends theories proposed by Leigh (2016) and Horsten and Leigh (2017) against the worry that these theories might be untrustworthy. The second claim further explains why and to what extent trustworthy theories are epistemically superior to untrustworthy theories. This chapter is a revised version of (Zicchetti, 2022a).

(D) Internal Categoricity and determinacy

Historically, the meta-theoretic property of categoricity has played an essential role in discriminating between two types of theories: (for the moment quite informally) theories that are about a unique subject matter and theories that are not about a unique subject matter. This quote by Isaacson might help explicate this matter:

> Mathematicians study two sorts of structures, which I shall call *particular structures* and *general structures*. The distinction is marked by [the] use of the definite and indefinite articles. We speak of *the* natural numbers and *a* group. Particular structures include the natural numbers, the Euclidean plane, the real numbers. General structures include groups, rings, fields, metric spaces, topologies. The particularity of a particular structure consists in the fact that all its exemplars are isomorphic to each other. (Isaacson, 2011, p. 18)

When considering arithmetic, traditionally, philosophers have strong intuitions about its determinacy: arithmetic is usually accepted to be about a *particular* subject matter, the natural numbers. Moreover, in most case, *arithmetical truth* is accepted as

determinate – more detail in chapter 6. The categoricity of arithmetic is important and has been sometimes employed to argue that arithmetic is conceptually and epistemically superior if compared to set theory.[8] Additionally, one could argue that we have *prima facie* good mathematical evidence to support this acceptance; Dedekind's categoricity theorem shows that second-order arithmetic is categorical in the sense that *all full models* of second-order arithmetic are isomorphic. Moreover, isomorphism implies that arithmetical statements have the same truth value across these models.

However, it is well-known that Dedekind-style categoricity arguments come with some significant philosophical and epistemological limitations. Such results rely on a restriction of the class of models considered to be only the full second-order models, i.e. those models in which the second-order domain is the full power set of the first-order domain. Philosophers have pointed out that the restriction to this 'special' type of model has some crucial philosophical drawbacks.[9] To name a few of these issues: the restriction to full models seems unjustified or justified from an outside perspective because the resources of second-order arithmetic do not suffice to define or determine full models; the interpretation of full models has to be provided using external means. Another issue with Dedekind-style categoricity theorems is that they are strongly impredicative. To circumvent these problems, some philosophers proposed an *internal* version of categoricity based on a philosophical position called

[8] See (Feferman, 2014).
[9] See (Parsons, 2008), (Isaacson, 2011), (Button and Walsh, 2016, 2018) and (Button, 2022).

internalism.[10]

This dissertation's third and last part focuses explicitly on internal versions of categoricity. More precisely, it focuses on Parsons-style categoricity theorems, i.e., theorems proved for first-order theories in first-order logic.[11] The main philosophical aim of Part III is to investigate the following (for now informal) question: Do Parsons-style categoricity theorems provide *determinacy* of truth? This chapter argues for the following claim:

8. Parsons-style categoricity provides *internal* determinacy. However, *no external* determinacy and no *standardness* are obtained.

This claim is closely connected to work done by Button and Walsh (2016), Väänänen (2012); Väänänen (2020), Väänänen and Wang (2015) and Hamkins and Yang (2013). As we will argue, these claims do not contradict the main result in (Hamkins and Yang, 2013) that (to put it succinctly) *satisfaction is not absolute*.

However, before focusing on the issue of determinacy, there is a particular problem concerning Parsons-style categoricity theorems that needs attention. Button and Walsh (2016, 2018) argued that such theorems are inadequate as internal categoricity theorems because they are not general enough. The main reason for this loss of generality is the choice of first-order logic instead of second-order resources. Chapter 6 aims to provide a general version of Parsons-style categoricity. Chapter 6 introduces and investigates a *truth-theoretic* Parsons-style categoricity theorem em-

[10] Core ideas have already been discussed by Parsons (1990), Parsons (2008, p. 112). Väänänen (2012) also uses the term 'internal categoricity'. Finally, internalism is introduced and discussed by Button and Walsh (2016, 2018).

[11] It focuses on arithmetical theories and leave an investigation of Parsons-style categoricity over set theory for the future.

ploying a primitive, axiomatic notion of truth. This should provide a more general and acceptable version of a Parsons-style categoricity theorem in first-order logic. This work will relate to and expand on the work done by Button and Walsh (2016, 2018), Mount and Waxman (2021), Simpson and Yokoyama (2013) and Feferman and Hellman (1995). Chapter 6 is a revised version of (Fischer and Zicchetti, 2022).

A final comment on this dissertation's methodology

Before starting this investigation with the first chapter, a brief remark on this dissertation's methodology is essential. We will understand 'epistemology of mathematics' as being much broader than just the issue of explaining why and how we have justified mathematical beliefs and (possibly) mathematical knowledge. Moreover, this approach aims to provide a more unified investigation of these issues by combining techniques from philosophical logic with some philosophical concepts, frameworks and methodology employed in general epistemology. Combining these methodologies will be a fruitful way forward in *both* epistemology of mathematics and general epistemology.

Part I

Cornerstones Propositions and the Epistemology of Consistency

Summary of Part I

This is a summary of the content of chapters 1, 2 and 3. In contrast to the previous introduction, this summary will be more detailed. A thorough discussion of the issues is available in each chapter.

Chapter 1

The aim of chapter 1 is to introduce entitlement theories as a *non-evidentialist* epistemology of mathematics and to discuss some of the relevant epistemological issues that arise in this context. The core of such epistemology is the claim that there are at least two types of warrant for mathematical propositions:[12] *evidential* and *non-evidential*. This chapter investigates foundational theories employed in so-called *epistemic foundational projects* – following Shapiro (2004) – and considers whether there is any warrant to believe the proposition that (the relevant) foundational theories in such projects are consistent. Within the context of such projects, we will introduce a well-known kind of *scepticism* that aims at showing that there is no warrant whatsoever to believe that (the relevant) foundational theories are consistent.[13]

As we will point out, the so-called sceptic argues that for any given foundational theory **S** in a given project, there is no warrant whatsoever to believe that **S** is consistent. This argument is pressing because the proposition that **S** is consistent turns out to be a *cornerstone* of any such epistemic foundational project, where cornerstones are propositions (in a way to be made explicit in chapter 1) essential for

[12]We will use the term 'warrant' as synonymous with 'justification'.

[13]This is going to relate to traditional discussions of scepticism by Pryor (2000, 2004) and Wright (2002, 2004, 2012) in the context of the epistemology of perception, but also to more recent discussions about scepticism in mathematics by Pedersen (2021) and Horsten (2021).

the integrity and significance of such projects. We will present one of the standard responses to this kind of scepticism, the so-called *entitlement theories*. Entitlement theories propose a *concessive* response to the sceptic: they concede to her that there is no evidential warrant to believe that (the relevant foundational theory) **S** is consistent. Nevertheless, agents are *non-evidentially* warranted – i.e. entitled – to believe that **S** is consistent. This part of chapter 1 should be understood as a simple reconstruction of the sceptical challenge in mathematics but also of the standard response to it provided by entitlement theories.[14]

The second part of chapter 1 focuses on one of the important epistemological issues arising in the context of entitlement theories. In particular, it focuses on the issue of the *epistemic force* of entitlement.[15] This is the question of whether entitlement constitutes an *epistemic permission* or an *epistemic obligation* to believe the relevant cornerstone. For the relevant case of the proposition that **S** is consistent, this amounts to whether agents are epistemically permitted or obligated to believe that (the relevant theory) **S** is consistent. This issue has received some attention in the epistemology of perception and within the context of scepticism.[16] As this part points out, the question about the force of entitlement has no immediate or trivial answer. This has to do with the nature and properties of entitlement. Entitlement is a non-evidential warrant, which is *unearned* and (in a sense to be made clearer later) *beyond* evidential support; it is never a result of evidential work. This chapter argues that agents are epistemically obligated to believe that **S** is consistent. Although

[14]This scepticism and the entitlement-based response are well-known in the literature about the epistemology of perception. Moreover, Pedersen (2021) recently discussed this scepticism in mathematics. Our reconstruction is mainly going to follow Pedersen's. Horsten mentions mathematical scepticism in (Horsten, 2021) (although without discussing it).

[15]We adopt the term 'force' for lack of a better term.

[16]See (Wright, 2004, 2012), (Jenkins, 2007), (Pedersen, 2008) and (Volpe, 2011).

details will be provided later, here is the gist of our claim: agents, who are engaging in the relevant project and consider the question of whether **S** is consistent, are *epistemically blameworthy* just in case they either believe that **S** is inconsistent or they fail to believe that **S** is consistent (by either doubting or by being open-minded about whether **S** is consistent). We will say that a doxastic attitude φ is epistemically blameworthy just in case φ-ing fails to conform to epistemic norms of the foundational project.[17] We will argue that agents are epistemically obligated to believe that **S** is consistent, insofar as failing to believe that **S** is consistent constitutes an epistemically blameworthy behaviour.

Chapter 2

Chapter 2 focuses on an additional issue arising in the context of entitlement theories: the issue of the entitlement's *epistemic pedigree*, i.e., the question of whether the entitlement type of warrant is good enough to constitute knowledge (if for instance compared with the evidential type of warrant).[18] Let us be more explicit about the focus of this chapter: its aim is not going to be to determine whether entitled propositions can be known, but of determining whether entitlement theories can coherently claim that entitled propositions cannot be known. In this sense, instead of focusing on the *epistemological* issue of providing a theory of knowledge that determines whether entitlement constitutes knowledge, this chapter focuses on the (somewhat) *meta-epistemological* issue of determining whether the epistemological claim that entitlement cannot constitute knowledge is coherent. This investigation

[17] Here this investigation roughly follows (Brown, 2018, 2019). This notion of an attitude being blameworthy will play a role also in chapter 5.

[18] In contrast to the issue of the entitlement's force, the issue of pedigree has received little attention in both general epistemology and in the epistemology of mathematics. This issue has been investigated by Smith (2020) and myself (Zicchetti, 2022b) within the epistemology of perception.

is significant because Smith (2020) recently argued that the claim that entitlement cannot constitute knowledge is bound to be incoherent. Roughly, this chapter aims to defend this claim against the worry of incoherence.[19] We will present the two positions within entitlement theories that directly address the issue of pedigree: the *moderate* and the *full-blooded* conception of entitlement.[20] The moderate conception claims that entitled propositions cannot be known, whereas the full-blooded conception claims that entitled propositions can be known.[21]

Our main aim here is to support the moderate conception of entitlement by arguing that it is coherent to claim that entitled propositions cannot be known. We will present the motivation to endorse the moderate conception and (hopefully) make the case that this idea is intuitive enough. As it turns out, the moderate conception of entitlement takes *evidential support* to be necessary for knowledge, whereas the full-blooded conception does not. After presenting the intuition behind the moderate conception, the chapter considers an important worry against the moderate conception:[22] that the moderate conception of entitlement is incoherent with *prima*

[19] An attempt to do so in the context of the epistemology of perception is to be found in (Zicchetti, 2022b). Some of the ideas present there are going to be also presented in chapter 2.

[20] There is a third position, called the *weak* conception of entitlement, endorsed for instance by Wright and Pedersen separately in (Wright, 2004) and (Pedersen, 2021). However, this dissertation will not discuss the weak conception because this position does not even allow for entitlement to be a warrant to believe propositions. According to the weak conception, entitlement is only a warrant to rationally accept the cornerstones. The proponent of the weak position provides an immediate negative answer to the issue of pedigree, by arguing that rational acceptance never results in knowledge. Although it would be interesting to investigate this position, this would exceed the scope of this dissertation.

[21] To my best knowledge, the moderate conception has not been endorsed yet. I endorse and defend this conception (although in the epistemology of perception) in (Zicchetti, 2022b). Moreover, to my best knowledge none of these positions has been explicitly endorsed. However, as we will point out, both Shapiro and Horsten separately claim in (Shapiro, 2011) and (Horsten, 2021) that entitled propositions can be known. This (at least) indicates that some full-blooded conception of entitlement is presupposed in the background.

[22] Such worries are collected in (Smith, 2020).

facie natural closure principles for knowledge, and that should be therefore rejected. After presenting the worry and making the assumptions needed for its formulation explicit, it is argued that the proponent of the moderate conception has good independent reasons to reject the (discussed) principles of knowledge closure. This chapter's conclusion will be the following: there is no immediate incoherence concerning the moderate conception, and the worries expressed by Smith (2020) can be resisted in a principled way.

Chapter 3

Chapter 3 investigates so-called 'soundness arguments for consistency'. Such arguments aim at inferring the consistency of a theory **S**, employing a soundness claim for **S** as one the argument's premises. Although these arguments are generally accepted as *valid* (or even *sound*), philosophers have strong intuitions that such arguments are *epistemically defective*. Girard (1987) claims that such arguments do not have any epistemic value. Dummett and Wright are also unconvinced that such arguments are epistemically good, as they both claim – separately in (Dummett, 1978) and (Wright, 1994) – that these arguments are *uninformative*. Recently, Piazza and Pulcini (2013) argued that *all* such arguments are (in a sense to be discussed later) *ill-founded*. Although the intuition that soundness arguments are epistemically defective seems to be correct, most of the epistemological narrative behind the reason why and to what degree such arguments are defective still needs to be made fully explicit and investigated. This chapter aims to make some progress in our understanding of why and to what degree soundness arguments can be evaluated (in one way or another) as epistemically defective. We will investigate this question by focusing on the issue of whether soundness arguments are *cogent*. We will say

that a valid argument is *cogent* just in case, if (a) the argument's premises are warranted and (b) the argument is valid, agents can, in principle, acquire a warrant to believe its conclusion in virtue of (a) and (b). This understanding of cogency will be similar to what is sometimes called 'transmissiveness'.[23] This chapter investigates whether soundness arguments are cogent from the background provided by the two epistemological positions about the *superstructure* of warrant: *conservativism* and *liberalism*. As we will explain later, these are positions about the superstructure of warrant, i.e., about what enabling conditions must be in place for a warrant to play its justificatory role. These positions have received extensive attention in the epistemology of perception but (to my best knowledge) have never been discussed within the epistemology of mathematics.[24] This chapter argues for the following claims:

1. Liberalism evaluates soundness arguments as cogent.

2. Conservativism evaluates soundness arguments as non-cogent.

After providing this analysis, the chapter concludes with some remarks on the significance and ramifications of the presented results.

[23]This notion has been discussed in the epistemology of perceptual warrant. See (Wright, 2002, 2003, 2012), (Pryor, 2000, 2004), and (Dretske, 2005) for a discussion. See instead (Moretti and Piazza, 2018) for an introduction to this topic.

[24]For a conservative, see Wright (2004, 2012, 2014). For a liberal, see Pryor (2000, 2004). For a general presentation and reconstruction of these two positions, see (Neta, 2010).

Chapter 1

Mathematical Scepticism and Entitlement Theories

1.1 Introduction

Agents constantly engage in inquiries into different subject matters. Using the terminology proposed by Crispin Wright, they engage in *cognitive projects*.[1] Informally, a project consists of a (collection of) question(s) and a (collection of) procedure(s) one might competently execute to answer the project's question. Projects have *cornerstones*: propositions essential for the integrity and significance of the project. For a given project about some subject matter D, we have the following understanding of cornerstones:

[1]For Wright's original presentation, see (Wright, 2004). From now on, we will refer to such inquiries simply as 'projects'.

(cornerstone) A proposition p is a cornerstone just in case the absence of warrant to believe p would imply the absence of warrant to rationally claim warrant to believe D propositions.[2]

Cornerstones include propositions expressing the proper functioning of the relevant cognitive faculties, the reliability of the instruments are theories employed in the project, the soundness of relevant principles of inference. Let us be more explicit about the importance of cornerstones by considering some inquiry into the empirical world. We have a strong informal intuition that perception can provide evidential warrant for (at least some) propositions about (common-sense) facts about the empirical world. In this case, the proposition that *perception is a reliable source of evidence* is a cornerstone; if there is no warrant to believe that perception is a reliable source of evidence, there is no warrant to rationally claim to be warranted to believe (at least some of) the ordinary propositions about the world. This is so precisely because of the role that perception plays in this project and by (cornerstone).[3] Our interest here is not in ordinary projects but in particular projects in mathematics.[4] Recently, Horsten discusses an example of cognitive projects in mathematics:

> For example, we might (admittedly somewhat ridiculously) identify the cognitive project of number theory with first-order Peano Arithmetic (**PA**), or, somewhat pedantically, with discovering facts about the natural numbers on the basis of proofs in **PA**. (Horsten, 2021, p. 741)

[2]This is essentially Wright's formulation. Wright (2004, p. 168) claims that "a proposition [is] a cornerstone for a given region of thought just in case it would follow from a lack for warrant for it that one could not rationally claim warrant for any belief in that region".

[3]Another typical cornerstone of such inquiries is the proposition that *there is an external world*.

[4]For investigations of mathematical inquiries as cognitive projects, see (Fischer et al., 2019), (Horsten, 2021), (Pedersen, 2016, 2021), (Wright, 2016) and (Zicchetti, 2022a).

Although we could discuss virtually any project in mathematics following Horsten's example, here we are interested in so-called *epistemic foundational projects*. For the understanding of such projects, we will follow Shapiro:

> The idea is that the foundation provides one way in which the mathematical propositions in question could have become known. It does not matter whether anyone came to know the propositions via the proposed foundation. One goal of the enterprise is to show that the mathematical propositions are a priori knowable. (Shapiro, 2004, p. 24)

Importantly, these projects aim to provide – as Shapiro (2004, p. 21) claims – an epistemic foundation of mathematics. This investigation focuses on even weaker epistemic foundational projects, aiming to provide a way in which the mathematical propositions in question could have been *warranted* in principle. For any mathematical subject matter D, an epistemic foundational project aims at providing a way in which mathematical propositions about D are *a priori* warranted, where by 'a priori' is roughly meant 'by purely mathematical means'. This is important for this investigation and for a better understanding of what Shapiro means in the quote. It is not significant whether agents actually know (or justifiably believe) propositions about D by other acceptable means. The project only aims at providing an alternative, purely mathematical route to this knowledge (or justified belief).[5]

Consider a project about some mathematical subject matter D. Assume that in this project, agents employ some foundational theory $\mathbf{S}$ about D to answer the project's

[5] As a side remark: this might be why Shapiro – right after the quoted passage – calls this epistemic foundational project a kind of 'reconstructive epistemology'. For completeness, we should point out that Shapiro contrasts these epistemic projects with other two foundational projects: *ontological* and *mathematical*.

questions. Following Horsten's example with arithmetic, let us assume that agents aim at providing answers to questions about D by virtue of proofs and refutations in $\mathbf{S}$. As a side remark, one should point out that for this discussion, it is not essential to be fully explicit about what D and $\mathbf{S}$ are. These could be respectively some arithmetical subject matter and an arithmetical theory – as in Horsten's example –, or some set-theoretic subject matter and some set theory.[6] Since we are focusing on foundational projects, it makes sense to think of D as (informally) being virtually all of (relevant) mathematics and $\mathbf{S}$ as being some foundational theory, which is supposed to be rich, expressive enough to decide questions about D by means of proofs and refutations. Again, the details of which foundational theory exactly $\mathbf{S}$ is will not be relevant for the discussion about mathematical scepticism in the following section. Think of $\mathbf{S}$ as "your favourite" foundational theory.

What is crucial for our discussion is the following: in these foundational projects, agents employ $\mathbf{S}$ to provide (at least defeasible) warrant for mathematical propositions about D. Using Wright's understanding of projects, together with Shapiro's focus on foundational ones, we can expand on Horsten's example and provide an informal understanding of what the project looks like. One can think about this project as the following pair:

$$\langle \text{ "Questions about the subject matter } D\text{", "Proofs and refutations in } \mathbf{S}\text{"} \rangle$$

Here, the first element of the pair is supposed to be a collection of questions about D that agents want to answer. On the other hand, the second element of the pair is the set of procedures to be competently executed to provide answers to the project's

[6] We only need to assume that $\mathbf{S}$ is acceptable by well-known and almost unanimously accepted mathematical standards: $\mathbf{S}$ is supposed to be some first-order, recursively axiomatisable theory, for which Gödel's Incompleteness theorems hold.

questions. As pointed out earlier, this is an epistemic foundational project: answering questions about D employing proofs in our foundational theory would determine what propositions about D can be warranted in principle. One should not read too much into the brief description of the project since this is just a (hopefully) illustrative means to clarify how one thinks about Wright-style projects.

It should be noted that in the context of Wright-style projects, we think of *evidence* and *evidential warrant* in a project P as being whatever warrant is provided by the chosen set of procedures in P. That is, *it is a part of the setup* that what is evidentially warranted (in P) is exhausted by the chosen set of procedures (of P). In our relevant foundational project, proofs and refutations in $\mathbf{S}$ are going to exhaust what count as evidence for mathematical propositions about D. This is not taken to mean that proofs in $\mathbf{S}$ are the only type of evidence to provide warrant for propositions about D *across all projects*. This hopefully helps clarify the issue: in other projects, one might take additional sources of evidence to provide justification to believe propositions about D. For example, one could have a project, in which agents accept *testimony* as an additional source of evidence to believe propositions about D. However, this would arguably be another project, since testimony should not be able to provide the 'right' type of warrant, when one is interested in the question of what is justifiable in principle, by purely mathematical means. To say a long story short: in Wright-style projects, what counts as evidence is always determined by the set of procedures in a project, and in our case, this is proofs and refutations in the given foundational system. As we will see in the following section, the fact that the set of procedures (in a given project) exhausts what counts as evidence to believe propositions about the relevant subject matter (in the project) is going to

be crucial to fully understanding the sceptical challenge.

In these foundational projects, the proposition expressing that **S** is consistent is a cornerstone; if there is no warrant to believe that **S** is consistent, then there is no warrant to rationally claim to be warranted to believe D propositions; this is so precisely because in these projects proofs in **S** are the chosen procedure to provide evidence to believe propositions about D.

It is almost unanimously accepted that any good epistemology of mathematics must address the questions of how and to what extent the cornerstones of our foundational projects are warranted. Alternatively, in other words, any good epistemology has to explain how it is so that the mathematical methods employed in our projects are reliable. An epistemology of mathematics that addresses such issues is sometimes called *dissident*. This is for instance the terminology employed (although in full generality and not restricted to mathematics) by Roland:

> Call an epistemology for a practice P dissident if and only if it not only addresses questions concerning the modes of justification (i.e., epistemic norms and standards) operative in P but also has the resources to address the question of the reliability [...] of those modes of justification. (Roland, 2007, p. 432)

Following Roland, an epistemology is *quietist* just in case it is not dissident. Again, it is quite intuitive to think that our epistemology (of mathematics) should be dissident, or in other words, that we should care as epistemic agents about questions concerning the reliability of our methods. From now on, this investigation operates

under this *working hypothesis*.[7] Moreover, we should note that some type of dissident epistemology is needed to understand the real force of *scepticism*: as we will explain in the following section, the sceptic challenges us to provide an argument to show that the relevant cornerstones of our foundational projects, such as the proposition that **S** is consistent, is warranted. However, the *quietist* would not even be willing to accept the challenge and would claim that there is no obligation to accept the sceptical challenge. However, from a dissident perspective, we seem obligated to take the challenge seriously.[8] The next section discusses mathematical scepticism in the context of our foundational projects.

1.2 Mathematical Scepticism

As sketched in the previous section, agents engage in the proposed epistemic foundational project and employ the foundational theory **S** to provide answers to questions about D. As mentioned earlier, proofs and refutations in **S** are what exhausts what counts as evidence *in this project*. Due to the structure of the project and the understanding of (cornerstone), the proposition that **S** is consistent is a cornerstone of the foundational project.

Now, the sceptic tries to argue that, for any given foundational project about (some relevant subject matter) D employing some foundational theory **S** to provide warrant

[7]The phrase 'working hypothesis' is used because there is still no knock-down argument against a quietist epistemology. To my best knowledge, the only instance of quietist epistemology seems to be provided by *naturalism* or *second philosophy* as in (Maddy, 2007). To provide an argument for dissident epistemology will amount to having a thorough discussion of naturalism, which cannot be given here.

[8]This has been discussed extensively in the philosophy of scepticism. See (Wright, 2002, 2004), (Williams, 1988, 2013) and (Pedersen, 2021).

to believe propositions about D, *there is no warrant to believe that* **S** *is consistent.* We can capture the essence of the sceptical argument in this way:[9]

(1) There is no way to obtain evidence for the proposition that **S** is consistent, using the given standard of evidence in the project, in this case, proofs in **S**.

(2) If there is no warrant in the evidential sense given by the procedures in the project to believe that **S** is consistent, then there can be no warrant whatsoever to believe that **S** is consistent.

(3) Therefore, there can be no warrant whatsoever to believe that **S** is consistent.

There is nothing special about the mathematical sceptic *per se* since it is just an instance of much general scepticism. To see this, let us consider a more well-known case of sceptical argument focusing on a project about the empirical world, aiming at answering questions about (common-sense) facts about the world, and which employs perception as the procedure to provide warrant to believe propositions about the world:

(P1) There is no way to obtain evidence for the proposition that perception is reliable using the given standard of evidence in the project, in this case, perception.

(P2) If there is no warrant in the evidential sense given by the procedures in the project to believe that perception is reliable, then there can be no warrant whatsoever to believe that perception is reliable.

(P3) Therefore, there can be no warrant whatsoever to believe that perception is reliable.

[9]This is meant as a reconstruction to understand how the sceptical argument is supposed to work. Here we follow the reconstruction provided by Pedersen (2021).

As pointed out by Pedersen (2021, p. 231), these sceptical arguments are instances of the following general *sceptical template*:

(*Context*) Proposition c is a cornerstone proposition for (the relevant subject matter D).

Step 1 There is no evidential warrant to believe c.

Step 2 If there is no evidential warrant to believe c, there is no warrant at all to believe c

Step 3 Therefore, there is no warrant at all to believe c

This scepticism attacks cornerstone propositions of many (and quite general) cognitive projects, such as the proposition that *we are not brains in vats*, the proposition that *we are not deceived by an evil demon*, the proposition that *there is an external world*. By targeting the cornerstones of our projects, sceptical arguments employing this template apply in the case of mathematical projects as well.[10]

Going back to the mathematical case, we can see that the sceptic might argue for (1) just by using the fact that **S** is an acceptable theory, for which the incompleteness theorems hold. For any such foundational theory **S** satisfying natural properties (which virtually all of our accepted mathematical theories satisfy), **S** *can never* provide evidence for its consistency by means of proofs in **S**.[11] The argument for (1) is quite general and in this context, it applies virtually to any consistent first-order theory, which satisfies the properties needed for the incompleteness theorems

[10] In his (Pedersen, 2021) he investigated the proposition expressing that **S** is satisfiable, for some mathematical theory **S** employed in some mathematical (non-foundational) project.

[11] More precisely, **S** cannot prove its canonical consistency statement. There are some technicalities involved with respect to the formulation of the consistency statement, however, these are not relevant to our discussion. Moreover, **S** is supposed to be rich enough to represent basic syntactic properties.

to hold. (2) is sometimes called the *sceptical lemma*: in full generality, it says that the only warrant available to believe the relevant cornerstone must be given by the chosen procedure(s) in the project. In this particular context, this amounts to the claim that the only warrant available in the foundational project to believe that **S** is consistent must be provided by a proof in **S**. With this, she claims that if there is no warrant to believe that **S** is consistent using proofs in **S**, then there is no warrant at all to believe that **S** is consistent. Finally, the sceptic infers (3) by *modus ponens*. Finally, since the proposition that **S** is consistent is a cornerstone of the project, the sceptic draws the following conclusion from (3):

(scepticism) There can be no warrant to rationally claim to be warranted to believe any D proposition.

The sceptical conclusion, if it cannot be resisted, would put any (foundational) projects in danger. Although it should be quite clear why (scepticism) is *bad*, let us be more explicit about this. First of all, the prior sceptical conclusion (3) would show that our dissident epistemology is inadequate; it is unable to provide an argument to show that the (relevant) cornerstone(s) of the foundational project(s) is warranted. Looking at (scepticism) more closely, we can see that it does not threaten the existence of warrants for D propositions: as Wright and Pedersen correctly point out separately in (Wright, 2012) and (Pedersen, 2021), (scepticism) does not present a *first-order* sceptical conclusion: it does not threaten the possibility of having a warrant for D propositions. Here is Pedersen's reconstruction:

> Note that scepticism [...] presents a higher-order rather than a first-order challenge. It targets *rational claims to warrant* rather than possession of [first-order] warrant. The first-order conception of scepticism is widespread. First-order scepticism could in a certain sense be addressed

> by our *de facto* good epistemic fortune. For, suppose that some form of externalism – (*de facto*) reliabilism, say – is right. In that case, if our environment was generally conducive and our belief-forming processes reliable, then our beliefs would be warranted. (Pedersen, 2021, p 231)

As we can see, (scepticism) does not say anything about whether we have warrant to believe D propositions. It targets our warrant to make rational claims about being warranted to believe ordinary propositions. In this sense, we agree with both Wright and Pedersen that this scepticism presents a kind of *intellectual challenge*; it threatens our intellectual integrity and our justification to make rational claims in our foundational projects. As Wright points out, by threatening our warrant to make rational claims to be warranted to believe ordinary propositions, this higher-order scepticism also threatens our warrant to make rational claims to know ordinary propositions about the relevant subject matter:

> What is put in doubt by [the] sceptical argument is not our possession of any knowledge or justified belief [...] [but] rather our right to claim knowledge. (Wright, 2004, p 210)

It should be clear by now that any good (dissident) epistemology must provide an argument to resist the sceptic. Fortunately, there are ways to try and resist the sceptical argument. Of course, one possibility to resist the sceptical conclusion is to argue that the sceptical argument is unsound, and accordingly to argue that at least one of its premises is false. The following section briefly presents the solution proposed by entitlement theories against the sceptical argument. The next section focuses on the entitlement-based response to the sceptic, which aims at resisting the sceptical conclusion by arguing that (2), the sceptical lemma, is false. However, let us say a few words about the alternative strategy: to argue that the first premise

is false. To argue that (1) is false, i.e., that there is a warrant to believe that **S** is consistent, one has at least two options. The first option might be to extend the relevant theory **S** to some theory $\mathbf{S}'$ (by the addition of some acceptable axioms motivated in some way or another), so that $\mathbf{S}'$ provides a warrant to believe that **S** is consistent, by means of a proof in $\mathbf{S}'$. This welcomes a certain kind of *regress*; the sceptic would (probably) argue that we have a warrant to rationally claim to be warranted to believe that **S** is consistent by means of a proof in $\mathbf{S}'$ just in case $\mathbf{S}'$ is reliable. After that, the sceptic would (probably) construct an analogous sceptical argument against the consistency of $\mathbf{S}'$. Of course, the devil is going to be in the details. However, this option does not seem particularly promising.[12]

The second option would be to *extend* the collection of procedures acceptable in the project, to provide a warrant to believe that **S** is consistent. However, to avoid a similar kind of regress happening with the first option, the new procedure or capacity to provide warrant to believe that **S** is consistent will have to outstrip the method of *mathematical proof* altogether. As a possibility, one might argue that propositions such as the one expressing that **S** is consistent are warranted by some capacity along the lines of *rational intuition* for instance. However, in these cases, one would have to provide a notion (and a theory) of the new capacity able to provide the warrant to believe that **S** is consistent. And even after this is done, there is no guarantee that the sceptic could not possibly apply her sceptical argument to the new capacity.[13]

[12] We will see later that this type of regress is going to be used to argue that the warrant to believe that **S** is consistent is an *entitlement*.

[13] Although investigating these strategies is of philosophical interest, it would exceed the scope of this dissertation. Thanks to Leon Horsten and Dan Waxman for separately suggesting to be explicit about this.

1.2.1 Anti-sceptical Argument for Entitlement

This section is devoted to the investigation of the entitlement-based strategy against the sceptic, which aims at resisting the sceptical conclusion by arguing that the sceptical lemma is false. First of all, entitlement theories accept that the antecedent of the conditional claim in (2) is true: they accept that there is no evidential warrant for the proposition that **S** is consistent. However, entitlement theories argue that there is an additional *non-evidential* type of warrant, called *entitlement*, for the proposition that **S** is consistent. Entitlement theories aim to show that, within any project *that employs* **S**-proving *as the standard method to provide evidential warrant*, the statement expressing the consistency of **S** is an entitlement. Informally, we can say that **S** is consistent just in case there is no statement p (in the language of **S**, such that both p and its negation are provable in **S**. Formally, the consistency of **S** can be straightforwardly formalized in **S** as the statement that there is no proof of a contradiction in **S**. For the discussion of the entitlement strategy, we assume that the entitlement theorist can freely switch from the informal consistency statement to the formalised version and *vice versa*.[14]

Before presenting the anti-sceptical argument for entitlement, let us introduce the relevant notion of entitlement employed by the entitlement theorist against the sceptic:

[14] As pointed in footnote 11, there are some technicalities involved with respect to the formulation of the consistency statement in **S**. However, these are not relevant for the philosophical point that the entitlement theorist wants to make. We assume that the entitlement theories works with a standard notion of provability satisfying the usual Löb's derivability conditions. See (Halbach, 2014) and (Boolos, 1993).

For a project P, a subject matter D, and a method E to provide evidential warrant for D propositions, a proposition p is an entitlement of P just in case the following holds:

(i) p is a cornerstone of P,

(ii) there is no independent reason to believe that p is not the case,

(iii) any attempt to provide an evidential warrant by means of E for p would involve either epistemic circularity or a commitment to an infinite regress of the justificatory process.[15]

We argued earlier that the proposition that **S** is consistent is a cornerstone of any project that employs **S**-proving as the standard method to provide evidential warrant for D-beliefs. So this proposition satisfies (i). With respect to (ii): we can safely assume that agents do not have any independent reason to believe that **S** is inconsistent. We can assume this because, in this project, **S** is the accepted theory employed to answer the project's questions; if agents had an independent reason to believe that **S** were inconsistent, they would not employ **S** to provide warrant for their D-beliefs. Moreover, we also have that agents engaging in the foundational project employing **S** do not have any evidence provided by their chosen standard in the project – that is, by proofs in **S** – that **S** is consistent. This follows simply by incompleteness considerations; as pointed out earlier in this chapter, we assume that **S** is a first-order, acceptable theory, satisfying the conditions needed for the incompleteness theorems to obtain. To put it briefly: agents do not have any evidence to believe that **S** is consistent because the relevant formal statement of the

[15]This presentation is quite similar to the one given in (Wright, 2004) and in (Pedersen, 2021).

consistency of **S** is independent of **S**.[16] Then, to resist the sceptical conclusion, entitlement theories have to show that the proposition that **S** is consistent satisfies (iii): that any attempt to provide a warrant *by means of the chosen method E* for the belief that **S** is consistent would either involve some circularity or some infinite regress in the justificatory process. Entitlement theories argue that the proposition that **S** is consistent satisfies (iii) in the following manner:

> (Context) Agents engage in P and investigate D. In P, **S**-proving is the accepted standard method to provide evidential warrant for D-beliefs.

(1*) If **S** is consistent, then **S** cannot provide evidence to believe in its consistency, employing an **S**-proof. Therefore, there is no possibility in P to provide an evidential warrant to believe that **S** is consistent.

(2*) Reacting to this, agents might start a new project, P', expanding on the previous project P, to provide an evidential warrant to believe that **S** is consistent. To do so, agents employ some other theory $\mathbf{S}'$ stronger than **S**. In P' agents take proofs in $\mathbf{S}'$ as providing evidential warrant to believe mathematical propositions. Assume that agents provide an $\mathbf{S}'$-proof of the consistency of **S**.

(3*) However, in the expanded project P', agents have warrant to rationally claim to be warranted to believe that **S** is consistent by means of an $\mathbf{S}'$-proof, only if there is a warrant to believe that $\mathbf{S}'$ is consistent.

[16] As a side remark, one might worry that also the formal statement expressing that **S** is *inconsistent* might be an entitlement. After all, if **S** is consistent, there is also no evidence provided by proofs in **S** that **S** is inconsistent. However, the inconsistency of **S** is obviously not a cornerstone of the foundational project, so that (i) is not satisfied. Additionally and more importantly, we will see in section 1.3.2 that agents have good reasons against believing that **S** is inconsistent. As I will show later, believing that **S** is inconsistent is going to violate what I will call norm of *epistemic responsibility*.

(4*) If $\mathbf{S}'$ is consistent, then $\mathbf{S}'$ cannot provide evidence to believe in its consistency, employing an $\mathbf{S}'$-proof. Therefore, there is no possibility in P' to provide an evidential warrant to believe that $\mathbf{S}'$ is consistent.

(5*) Therefore, agents do not have any warrant to rationally claim to be warranted to believe that $\mathbf{S}$ is consistent by means of an $\mathbf{S}'$-proof.

The entitlement theorist argues that agents are going to repeat this reasoning in principle *ad infinitum*: agents are bound to expand their projects continuously. However, given that the foundational theories in the expanded projects are acceptable so that they cannot prove their consistency, agents will not be able to have any evidential warrant for the relevant cornerstone in each project. Therefore, they will not be able to have a warrant to rationally claim to be warranted to believe the consistency of the (starting) foundational theory $\mathbf{S}$. That is – so the entitlement theorist – the proposition that $\mathbf{S}$ is consistent satisfies (iii) and is thereby an entitlement of P.[17]

At this point, it is instructive to pause and make a few remarks concerning the anti-sceptical argument. First, for this reasoning to work, our chosen theories $\mathbf{S}$, $\mathbf{S}'$, $\mathbf{S}''$, ... must satisfy the properties needed for the incompleteness theorems to hold. This is not problematic because virtually any first-order theory employed in our projects satisfies these properties. However, an additional remark is quite important: the assumption that *proving D-propositions in our accepted theory is the standard method of producing evidence for D propositions* is made by *both* the sceptic and the entitlement theorist. From this assumption (together with the fact the incompleteness theorems hold for our theories) *both* the sceptic and the entitlement

[17]Let us point out that Pedersen presents a similar argument in (Pedersen, 2021) for the claim that the proposition that our accepted theories are *satisfiable*, i.e., have at least some interpretation, is an entitlement.

theorist reason for (1*), (4*). This is also significant for the formulation of (2*); the entitlement theorist keeps the informal standard method to provide evidential warrant *fixed*: proving D-propositions in our accepted theory. This is not an issue in this context. However, it is crucial that *both* the sceptical and anti-sceptical arguments are formulated *relative* to the chosen standard method E to provide evidential warrant in *each* project. For this reason, entitlement theories provide a *concessive* answer to scepticism; they concede to the sceptic that there is no evidential warrant for the belief that **S** is consistent, relative to our chosen standard method of producing evidence.

1.2.2 The Core of Entitlement Theories

Hopefully, it was clear enough from the previous section that entitlement theories resist the sceptical conclusion – that there is no warrant whatsoever to rationally claim to be warranted to believe propositions about D – by arguing that the relevant cornerstones are an *entitlement*. In our particular case, this amounts to the claim that the proposition expressing that the (relevant foundational) theory **S** is an entitlement. This is the *core* of entitlement theories.

At this point, it is helpful to be more explicit about entitlement: mathematical entitlement theories distinguish between two types of warrant in these mathematical projects. The standard type of warrant is *evidential*, that is, it is provided by means of the chosen standard to provide evidence in the foundational project, that is, proofs in **S**. Normally, mathematical propositions about D can be warranted evidentially, that is, can be supported by evidence provided by means of a proof in **S**. In contrast to this type of warrant, an entitlement to believe that **S** is consistent

is non-evidential, insofar as (by the very understanding given in the previous section) it cannot be provided by the standard of evidence in our foundational project. As pointed out, this is the *concessive* part of the entitlement-based response to the sceptic: both the entitlement theorist and the sceptic agree that in any foundational project employing (some suitable theory) **S** to provide evidential warrant for mathematical propositions (about the relevant subject matter), the proposition that **S** is consistent cannot be supported by evidence. This is crucial: entitlement for a proposition p never consists in possession of evidence for p.[18] In this sense, entitlement is always an *unearned* warrant; it never results from some *evidential work*, from any investigation of the relevant subject matter D.[19] As hinted in the previous section, entitlement results from the fulfilment of a somewhat *negative* clause: a cornerstone p is an entitlement, absent a warrant to believe that p is false, and provided that any attempt to justify p would result in some epistemically vicious process.[20] This ends the exposition of the core of entitlement theories.

In addition to this core, entitlement theories differentiate themselves with respect to the issue of what attitude is warranted by the entitlement type of warrant: *weak* conceptions of entitlement claim that entitlement is a warrant to *rationally accept* the (relevant) cornerstones. On the other hand, *moderate* and *full-blooded* conceptions

[18]For this, see (Wright, 2014) and (Pedersen et al., 2020).

[19]Sometimes, entitlement is described as an *internalist* type of warrant, as pointed out in (Pedersen et al., 2020). Although this dissertation does not focus on the issue of whether entitlement is internalist or not, one should still point out that, if this characterisation were to be correct, then this type of entitlement would be substantially different from Burge-style entitlement, which is understood as an *externalist*, *evidential* type of warrant. Furthermore, this is significant for other discussions of mathematical entitlement in (Shapiro, 2011), (Fischer et al., 2019) and (Horsten, 2021). For a discussion of the tension between these two conceptions of entitlement see (Pedersen et al., 2020; Pedersen and Graham, 2020), and for Burge's notion see (Burge, 2003, 2020).

[20]For some presentations of these properties of entitlement, see (Wright, 2004, 2012), (Smith, 2013), (Pedersen, 2021), and (Horsten, 2021).

of entitlement claim that entitlement is a warrant to *believe* the (relevant) cornerstones.[21] One of the main motivations (as reported by Wright, who is the leading proponent of the weak conception) to choose the weak conception over the other two is the intuition that rational belief must be warranted evidentially (Wright, 2004, pp. 192-4). On the other hand, proponents of the moderate and full-blooded conceptions are more liberal concerning what counts as a warrant to rationally believe a proposition. They depart from the idea that rational belief is tied to evidence and accept that in the case of the (relevant) entitled cornerstones, rational belief is warranted non-evidentially. The discussion in the next section presupposes as a background either a moderate or a full-blooded conception. This is not problematic for the present investigation; the following section investigates the issue of the *force* of entitlement, i.e., the question of whether entitlement constitutes an *epistemic permission* or an *epistemic obligation* to *believe* the (relevant) cornerstones. Of course, this question is of any philosophical interest only if entitlement is a warrant to believe cornerstones *to start with*. For if someone endorses a weak conception of entitlement, she would (trivially) answer the question of the entitlement's force negatively: entitlement would neither constitute a permission nor an obligation to believe the cornerstones because – so the weak conception – entitlement only warrants rational acceptance, and rational acceptance is different from belief! The reader, who only accepts the weak conception, is free to understand this investigation of the entitlement's force as conditional on the possibility that either the moderate or full-blooded position is acceptable.

[21] Although the moderate and full-blooded conception agree that entitlement is a warrant to believe the cornerstones, these conceptions disagree about whether entitled cornerstones can be known. This issue is going to be the focus of chapter 2. For a general discussion of all three conceptions see for instance (Smith, 2020).

Before continuing, we should point out that this chapter (and with respect to the issue of force) does not aim to adjudicate the dispute between weak, moderate and full-blooded conceptions. So, the decision to presuppose either a moderate or full-blooded conception (for what is to follow) is only needed for methodological purposes and nothing more.[22]

1.3 The Force of Entitlement

This section aims to investigate the issue of the entitlement's force.[23] As pointed out earlier, this issue is understood as the question of whether the entitlement type of warrant constitutes an epistemic permission or an epistemic obligation to believe the (relevant) cornerstones. Although the next sections will be fully explicit about what is meant by 'epistemic permission', 'epistemic obligation' (and other related notions), let us introduce the main idea more generally and informally: to investigate a particular instance of the issue of the entitlement force, and focus on the cornerstone that the (relevant) foundational theory **S** is consistent. The question of whether entitlement constitutes a permission or an obligation to believe that **S** is consistent will be understood as the issue of determining whether agents are either (i) epistemically permitted to believe the negation of the cornerstone, i.e., to believe that **S** is inconsistent, or (ii) epistemically permitted to be open-minded

[22] Although an investigation of the distinctions between weak (on the one hand) and moderate and full-blooded conception (on the other) would be of philosophical interest, it will not be pursued here because it is not relevant to this investigation. For a discussion of the weak position see for instance (Pedersen, 2008) and (Jenkins, 2007). However, such an investigation would have to provide a principled discussion of the relevant distinctions between belief and rational acceptance, and of their possible connection to evidential support. See for instance (Cohen, 1992) and (Engel, 1998) for two very well-known presentations of the main distinctions. See also (Van Fraassen, 1980) for a conception of acceptance much similar to belief.

[23] As pointed out in the introduction, the term 'force' is employed for lack of a better term.

about whether **S** is consistent.[24] To put it succinctly, we will argue for the following claim (Epistemic Obligation):

(EO) Agents are epistemically obligated to believe that (the relevant theory) **S** is consistent.

We will argue for (EO) by showing that agents are neither permitted to believe that **S** is inconsistent nor permitted to be open-minded about whether **S** is consistent. Either believing that **S** is incosistent or being open-minded about whether **S** is consistent will amount to an *epistemically blameworthy* attitude by failing to conform to natural norms of the epistemic project (details to be introduced later).

Before continuing, a few clarifications are needed: this approach does not consider all epistemically possible situations, in which agents can be. For instance, it does not consider situations where agents do not (fully) believe that **S** is consistent, but are nevertheless to some degree confident that **S** is consistent. This investigation focuses on *full* belief and (failure thereof).[25] We will argue that (EO) is exactly what we should expect; it is in harmony with Wright's additional claim (to be discussed in more detail later) that entitlement for a cornerstone is such that it excludes the epistemic possibility of being open-minded about whether the cornerstone obtains.

1.3.1 Permission, Obligation and Epistemic Blame

Are agents obligated to believe the proposition that **S** is consistent when this proposition is an entitlement? This question has no immediate, trivial answer, and one reason for this is (first of all) that the meaning of 'obligation' is still unspecified.

[24] I will say more about the notion of 'open-mindedness' involved.

[25] To investigate *quantitative* notions of belief would be certainly of philosophical interest. However, this would require a separate investigation and is therefore left open for future work.

So, the first obvious thing to do is to present the relevant understanding of 'obligation'. Foundational projects – understood as cognitive inquiries – are *epistemic practices*: they have aims and goals, which are pursued by the agents engaging in them. Practices have norms, where norms can be informally seen as 'rules' that regulate the practice (to some degree). That practices have norms should be fairly acceptable. Epistemic norms can be informally understood as 'rules' regulating explicit epistemic dimensions of the practice:[26] This is quite similar to the informal understanding proposed for instance by Henderson

> People develop and deploy epistemic norms - normative sensibilities in light of which they regulate both their individual and community epistemic practice. (Henderson, 2020)

To be more precise, we follow Pollock and understand epistemic norms as "describing when it is epistemically permissible to hold various beliefs." (Pollock, 1987, p. 61). Expanding on Pollock, we can see that epistemic norms also describe when it is not permissible to hold certain beliefs. Consider for instance some project inquiring into the (empirical) world: a natural norm is for instance the following (No Dogmatism):

(ND) For any proposition p, one should refrain from believing p against compelling evidence that p is false.

This norm 'says' that it is not epistemically permissible to believe p against compelling evidence that p is false. In this sense, epistemic norms have normative force. In an epistemic practice X with epistemic norms $R_1, ...R_n$, agents engaging with X *ought to* follow the epistemic norms when forming their beliefs. We will expand

[26]To the best of my knowledge, it is widely accepted that epistemic practices are (at least partially) regulated by epistemic norms. A separate issue is to provide a principled way to tell different kinds of norms apart. We will not focus on this issue here. See for instance (Kauppinen, 2018) for an investigation of this issue.

on Pollock here and have a more encompassing informal understanding of epistemic norms, endorsing that epistemic norms describe when it is epistemically permissible (alternatively not epistemically permissible) to hold various *epistemic attitudes* (in addition to belief).[27] So there seems to be an intuitive sense in which we can think of agents as being *epistemically obligated* to hold their epistemic attitudes in accordance with the (relevant epistemic norms). A way to understand the notion of 'epistemic obligation' is by looking at the notion of 'epistemic blame' (Obligation in terms of Blame):

(O-in-B) An agent is *epistemically obligated* to φ p just in case not φ-ing p is epistemically blameworthy.

We say that otherwise, agents are permitted to φ p. Of course, this does not tell us enough about the notion of epistemic obligation involved. However, it tells us already that this notion of obligation is (probably unsurprisingly) weak in the following sense: the obligation to φ p does not mean that agents *cannot in principle* fail to φ p. It only means that agents, who fail to φ p, are epistemically blameworthy, i.e., can be justifiably epistemically blamed by their peers.[28] Finally, if agents fail to conform to (the relevant) epistemic norms of the practice, then they are epistemically blameworthy. This understanding of blameworthiness follows the informal idea spelled out by Brown:

> When subjects violate epistemic standards or norms, we sometimes judge them blameworthy rather than blameless. (Brown, 2018, p. 389)

And by Boult:

[27] For investigations in this direction see for instance (Nottelmann, 2007) and (Rettler, 2017).

[28] For this investigation, it will not be relevant to precisely discuss who the peers are. However, let us say (for clarity) that I understand peers as agents with (i) the same cognitive capacities and (ii) the same access to (the total) evidence as the target agent.

> Epistemic evaluation is a familiar part of ordinary life. We routinely judge others to be irrational, or unjustified in holding certain beliefs. We regard others as doing something they should not when they suspend judgment on a matter about which there is unequivocal evidence. Just as we have characteristic ways of responding to one another for moral failings - for example, we sometimes blame others for moral wrongdoings - it seems there are characteristic ways of responding to one another for epistemic failings. (Boult, 2021b)

Brown and Boult separately argue that there is a specific *epistemic kind of blame.*[29] When thinking about the (ND) norm, it is intuitive to think that "we might judge a subject blameworthy for dogmatically continuing to believe a claim even after receiving evidence which undermines it." (Brown, 2018, p. 389) Before continuing with the investigation of the issue of the force of entitlement (in the context of our foundational project), let us reformulate an updated version of the claim (Epistemic Obligation):

(EO*) If agents do not believe that (the relevant theory) **S** is consistent, then they are epistemically blameworthy.[30]

Since in this context to be epistemically blameworthy can be seen as a result of failing to conform to (the relevant) epistemic norms, the argument for (EO*) is going to show that either believing that **S** is consistent or being open-minded about

[29]One might worry that there is no genuine epistemic kind of blame. Such worries might be motivated by the idea that blame is a *moral* concept (Kauppinen, 2018). This dissertation does not provide any argument for the claim that there is a specific epistemic kind of blame, and it assumes that the epistemic account of blame is coherent enough. Finally, this investigation of the entitlement's force should be seen as an *application* of epistemic blame to the context of entitlement theories. For recent investigations of epistemic blame, see (Schmidt, 2021), (Boult, 2020, 2021a) and (Boult, 2021b).

[30]This is similar to (EO) with the distinction that I simply substituted the informal understanding of 'epistemic obligation' in the relevant position.

whether **S** is indeed consistent fails to conform to (the relevant) epistemic norms. Let us focus now on foundational projects in mathematics.

1.3.2 Epistemic Responsibility

As pointed out at the beginning of this chapter, (the) agents (that we consider) engage in some epistemic foundational project to answer the informal question of what propositions about the (relevant mathematical) subject matter D can be warranted in principle. To provide answers to this question, agents employ some foundational theory **S** about D and take proofs in **S** as providing evidence to believe propositions about D. In this project, an epistemic norm such as (Mathematical No Dogmatism) seems to be intuitive:

(MND) For any proposition p about D, one should refrain from believing p in the presence of compelling mathematical evidence against p.

In such projects, an example of compelling evidence against p would be a proof in the foundational theory **S** of the negation of p. Using different terminology, we can understand epistemic norms such as (MND) as expressing that agents should be responsive to *defeaters*.[31] Using the terminology employed for instance by Pritchard in (Pritchard, 2016), we can distinguish between (at least) two types of defeater: *undercutting* and *overriding*.[32] An undercutting defeater is supposed to be any compelling consideration that counts against the reliability of the source of evidence employed. When discussing the case of perception, Pritchard brings this as an example of undercutting defeater:

[31] For a recent introduction to defeaters in epistemology see (Moretti and Piazza, 2017).

[32] These two types of defeaters are sometimes also called *first-order* defeaters (Carter, 2016). Moreover, overriding defeaters are sometimes called 'rebutting'. For a traditional discussion of the distinction between undercutting and overriding defeaters (although in a slightly different terminology) see for instance (Pollock, 1970).

> [I]t could be that the defeater is that one's perceptual faculties are not functioning as they ought [...]. (Pritchard, 2016, p. 3069)

On the other hand, an overriding defeater does not count against the reliability of the source of the evidence, but rather provides compelling counter-evidence against the target proposition. Using this terminology, we can see that a proof in **S** of the negation of p is going to count as an overriding defeater for p.[33] So, more precisely, (MND) can be understood as claiming that agents should be responsive to overriding defeaters. It seems intuitive to judge agents as blameworthy if they are not responsive to such defeaters (in the relevant project).

However, norms such as (MND) do not help determine whether agents are epistemically permitted to believe that **S** is inconsistent, or permitted to be open-minded about the issue of whether **S** is consistent. This is so because, as pointed out earlier in this chapter, the entitlement to believe that **S** is consistent (and in general to believe cornerstones) is a non-evidential type of warrant: it never results from any evidential work. Remember also that entitlement theories agree with the sceptic that, for any foundational project, there is no evidential warrant (relative to the employed standard) to believe that **S** is consistent, or to believe its negation (and in general to believe the relevant cornerstones): agents neither have evidence for the proposition that **S** is consistent, nor against it. Nevertheless, we can use a different argument to make the case that agents should believe that the theories *that they employ to make cognitive achievement in their projects* are (at least) consistent. Considering the following epistemic norm of (Epistemic Responsibility):

[33] On the other hand, an undercutting defeater in our foundational project would be any compelling consideration or evidence that the source of evidence in the target project is unreliable. Compelling considerations or evidence that **S** is inconsistent would count as such undercutting defeater.

(ER) Agents should be in a position to rationally claim to be warranted to believe propositions, for which they have evidential warrant.

Let us be more explicit about (ER). Consider our foundational project employing **S**: agents engage in the (relevant) foundational project to make progress with respect to the question of what propositions about D can be warranted in principle. Now, assume that there is some proof in **S** of some proposition p about D, so that agents can conclude (within this project) that p is one of the propositions about D that can be warranted in principle. Given that there is a proof of p in the foundational theory **S**, it seems pretty natural to want not only that agents are in a position to claim that they are warranted in believing p precisely because they provide a proof in **S** of it, but that they are also rational in doing so. And this is the intuition that (ER) captures. By adopting (ER) we can show that agents are epistemically obligated to believe that **S** is consistent. As mentioned previously, this amounts to show that (i) believing that **S** is inconsistent or (ii) being open-minded about whether **S** is consistent fails to conform to (ER). What follows treats the two cases (i) and (ii) separately (for clarity).

(i) To believe that S is inconsistent. Assume that some *epistemic peer*[34] engaging in the relevant foundational project about D employing **S** to provide evidential warrant to mathematical propositions about D also believes that **S** is inconsistent. One can see that believing **S**, in this case, violates (ER): if the agent believes that **S** is inconsistent and nevertheless employs **S** to provide evidential warrant to believe propositions about D, she results in the following position:

[34]We assume that the agent is rational and that she is an epistemic peer, for these issues to be of interest. Otherwise, one could 'dismiss' the agent as being not on a pair with the others, as being unreasonable or irrational.

She cannot rationally claim to be warranted to believe the propositions that are evidentially warranted by proofs in **S**.

And this immediately clashes with (ER). It does clash because it is not rational to claim to be warranted to believe p – in virtue of a proof in **S** of p – whilst believing that **S** is inconsistent. Her belief that **S** is inconsistent seems to act like an undercutting defeater, insofar as it undermines the reliability of **S** and thereby the evidential support that proofs in **S** provide to believe propositions about D.[35] Since case (i) results in a violation of (ER), (i) is epistemically blameworthy.

(ii) To be open-minded about whether S is consistent. What about the second possibility, that an epistemic peer is open-minded about whether **S** is consistent whilst employing **S** to provide evidential warrant to believe propositions about D? First of all, let us say a few words about open-mindedness. We say that an epistemic peer is open-minded about whether p, insofar as she suspends judgments with respect to p, and remains open and responsive to new evidence for either p or its negation.[36] Although to a lesser degree, (ii) is at odds with (ER) for the following reason: due to her open-mindedness about whether **S** is consistent, the epistemic peer should be equally open-minded about whether **S** provides evidential warrant to believe mathematical propositions about D through proofs in **S**. It this sense, being open-minded about whether **S** is consistent plays the role of an undercutting defeater, by undermining the reliability of **S** – although not as substantially as in (i). Finally, she has to be open-minded about whether she is in any posi-

[35] It does so because we work under the assumption that the peer is reasonable.

[36] Wright (2012) points out that open-mindedness is weaker than agnosticism or scepticism about whether p, insofar as open-mindedness (as understood) includes being open and being responsive to whatever new compelling evidence. This is not always the case when discussing agnosticism or scepticism because these attitudes involve doubt about whether p and agents might not be responsive to all compelling evidence.

tion to rationally claim to be warranted to believe propositions about D by means of proofs in $\mathbf{S}$. Therefore, (ii) is epistemically blameworthy. We can state our conclusion again: (EO) agents are epistemically obligated to believe that $\mathbf{S}$ is consistent.

A few clarifications and remarks on the limitations of the claim (EO) are in order: this investigation did not argue that any epistemic peer, who fails to believe that $\mathbf{S}$ is consistent, is epistemically blameworthy. When arguing for the claim (EO), we implicitly restricted our interest to epistemic peers, who fail to believe that $\mathbf{S}$ is consistent, *after considering the issue of whether* $\mathbf{S}$ is consistent. This is significant because we do not want to imply that epistemic peers, who never even thought about the issue of consistency, are epistemically blameworthy. Importantly, an epistemic peer that fails to believe that $\mathbf{S}$ is consistent by failing even to consider the issue does not fail to conform to (ER) and is therefore not to blame.

A second clarification is with respect to the *scope* of the entitlement's force. In the arguments for (EO), we always kept a project fixed and reasoned within the project. In this way, I argued that, relative to some project P employing some theory $\mathbf{S}$, epistemic peers are epistemically obligated to believe that $\mathbf{S}$ is consistent. This, however, should not be expected always to be the case *across* projects. This leaves the possibility open that two different communities, C and C', engaging in two different foundational projects P and P', and employing two different foundational theories $\mathbf{S}$ and $\mathbf{S}'$, might disagree about what is permissible to believe across P and P'. Let us point out that this is not a shortcoming of our solution; this is another consequence and philosophical issue arising within the context of entitlement theories. This seems to have to do with what Pedersen calls the *generosity* of entitlement (Pedersen, 2022). Entitlement applies to many propositions and – so

Pedersen points out – it applies to cornerstones c and c' of two different projects, such that c and c' would be incoherent if taken together.[37] Several philosophical issues arise from the generosity of entitlement: from the issue of epistemic relativism (for instance of relativism concerning the value of entitlements) to the issue of evaluating peer disagreement about cornerstones across projects.[38]

1.4 Conclusion

This chapter introduced and discussed a non-evidentialist epistemology of mathematics. To do so, the first part of the chapter has been devoted to presenting epistemic foundational projects, the notion of a cornerstone and mathematical scepticism. The chapter presented entitlement theories as a non-evidentialist response to the sceptic and discussed the notion of entitlement. The second part of the chapter discussed the issue of the entitlement's force. It focused on the issue of whether epistemic peers are epistemically obligated to believe that the relevant theories employed in their foundational project are consistent, and concluded that they are obligated to do so. We supported this conclusion by arguing that failing to believe that the relevant theories employed in the project are consistent violates natural epistemic norms. In particular, it violates (Epistemic Responsibility). Indeed, the relevant chapter did not provide an additional, independent motivation for (ER) other than an appeal to it being natural or intuitive. But now we can at least

[37] As Pedersen points out, entitlement might even apply (in two different projects) to both *sceptical hypotheses and their negations*.

[38] Although these issues are of great interest in general and for anyone, who endorses entitlement theories, to investigate these issues would exceed the scope of this dissertation. However, these would be immediate questions concerning to peer disagreement about cornerstones: Is such disagreement *faultless*, so that neither one of the disagreeing peers can be held blameworthy? Is this disagreement *deep*, so that it does not have any rational solution? And if it is deep, what kind of deep disagreement is it? For a discussion of these issues see for instance (Kölbel, 2004), (Davis, 2014), (Ranalli, 2018), (Martin, 2019), (Lynch, 2010), and (Smith and Lynch, 2020).

see more clearly why it is so that (ER) is intuitive: it provides an analysis of the force of entitlement capturing our informal intuitions about how we should evaluate situations such as (i) and (ii): we should be able to judge agents as epistemically blameworthy, if they employ a theory **S** (or whatever relevant source of evidence E) *whilst* believing that **S** (or whatever relevant E) is unreliable, for this behaviour undermines the epistemic practice. This behaviour even undermines the possibility of other agents to be in a position to rationally claim warrant to believe evidentially warranted propositions. This is so because the belief that **S** is inconsistent can be (informally) seen as an undercutting defeater of any proposition p provable in **S**. The belief that **S** is inconsistent attacks the reliability of **S**. As we have seen, being open-minded about whether **S** is consistent does a similar undermining (although to a lesser degree).

Chapter 2

Entitlement's Pedigree and the Moderate Conception[1]

2.1 Introduction

Chapter 1 considered the issue of the entitlement's force: focusing on the proposition expressing that **S**, the theory employed in the relevant foundational project, is consistent, it argued that agents engaging in the project are epistemically obligated to believe that **S** is consistent, insofar as they are epistemically blameworthy if they fail to believe that **S** is consistent, after considering the matter. This chapter focuses on a separate related issue that naturally arises in the context of entitlement theories: entitlement's *pedigree*. One can understand this issue as whether the entitlement type of warrant is 'good enough' (or has 'the right pedigree') to constitute knowledge, or more simply, as the question of whether entitled propositions can be known. Entitlement theories invoking a *moderate* conception of entitlement claim

[1]Some of the ideas present in this chapter – particularly in section 2.3 – are contained in (Zicchetti, 2022b).

that entitled propositions should be believed but cannot be known. In contrast, the *full-blooded* conception of entitlement makes the additional claim that entitled propositions can be known.

This chapter is devoted to the moderate conception, and in particular, it focuses on a (somewhat) *meta-epistemological* issue related to the question of pedigree: to determine whether *the claim* – made by the proponent of the moderate conception – that entitled propositions cannot be known is tenable.[2] This interest is motivated by the fact that the moderate conception of entitlement seems to capture quite natural intuitions about the nature of entitlement and of the relevant differences between entitlement and evidential warrant. However, despite this conception's intuitiveness, it has been recently argued that the moderate conception is nevertheless untenable: Smith (2020) argues that the moderate conception of entitlement is not tenable because it is incoherent with natural principles of knowledge closure. In essence, this chapter aims to provide a first preliminary defence of the moderate conception of entitlement against these worries. Nevertheless, this work is essential for entitlement-based epistemological theories that do not want to go full-blooded; for if Smith is correct in his objection, there is no hope – as he argues – for coherent entitlement theories that do not adopt the full-blooded conception. To put it succinctly, if this defence is in good standing, it provides some hope for an entitlement

[2]This is meant as a clarification: we call this a meta-epistemological issue for the following reason: to focus on the issue of pedigree directly would mean to investigate an epistemological question: the issue of whether entitlement constitutes knowledge. However, this chapter focuses on whether a specific attitude towards the issue of pedigree – the moderate conception – is tenable.

theory based on the moderate conception.[3].

Here is the chapter's structure (details are in the relevant sections): section 2.2 introduces the moderate conception and presents the intuition behind the claim that entitled propositions, *in contrast to evidentially warranted propositions*, cannot be known. After that, section 2.3 discusses Smith's main worry that the moderate conception is incoherent with *prima facie* intuitive principles of knowledge closure. This chapter argues that although Smith is correct in claiming that the moderate conception is incoherent with his proposed principles, the proponent of the moderate conception has good independent reasons to reject them.

2.2 The Moderate Conception of Entitlement

As explained in chapter 1, the core of entitlement theories is provided by the acceptance of a non-evidential type of warrant, called entitlement, for the cornerstones of (the relevant) projects. In the context of foundational projects about (the relevant subject matter) D and employing some foundational theory $\mathbf{S}$, we saw that the proposition that $\mathbf{S}$ is consistent is a cornerstone of the project and that, moreover, agents are entitled to believe that $\mathbf{S}$ is consistent. As Smith points out, welcoming the idea that agents are entitled to believe (relevant cornerstones) immediately "raises a further question: Could [agents] be in a position to know a proposition to which [they are] merely entitled?" (Smith, 2020, p. 284) Focusing on examples of

[3]However, we will not discuss the full-blooded conception in any depth. Moreover, we will not attempt to adjudicate the dispute between moderate and full-blooded conceptions. To do so would involve a substantive investigation of both the moderate and full-blooded theory of knowledge, which exceeds this investigation's scope. Finally, this chapter does not aim to provide a fully-fledged characterisation of the moderate conception and its epistemology

cornerstones from other projects in the context of the epistemology of perception, he makes the following remark:

> Could I be in a position to know, for instance, that there are other minds? If I have an ordinary earned justification for believing a proposition, and the proposition is true, and the situation is normal and I'm not in any way Gettiered, then, conventionally, this would be regarded as enough for me to be in a position to know the proposition. If I believed the proposition, and my belief was appropriately based upon my justification, then it would qualify as knowledge. But what if the justification in question is not an earned one, but an entitlement instead? Can entitlement supply the needs of knowledge in the same way that earned justification can? (Smith, 2020, p. 284)

This issue is essential for non-evidentialist, entitlement-based epistemological theories. As one can see from the quote, the crux of the issue is that entitlement is *unearned* and not supported by evidence. This issue is equally relevant in the context of entitlement theories in mathematics. Since the proposition that **S** is consistent is merely entitled in the relevant foundational projects, the same question arises. We can reason by analogy with the case presented by Smith and reason in the following manner: assume that agents have an evidential warrant to believe a mathematical proposition about D. Moreover, assume that the proposition is true and that the agents are not Gettiered. Under these conditions, entitlement theories endorse that agents are in principle in a position to know the target proposition about D. But assuming that everything else is equal, the warrant for the relevant proposition is an entitlement. Could agents be in a position to know the relevant entitled proposition?

Proponents of different conceptions of entitlement provide contrasting answers to this question. The so-called moderate conception answers this question negatively. In contrast, the full-blooded conception answers it positively.[4] To be more explicit, let us briefly consider the some foundational project – as in the previous chapter – employing some theory **S**. The moderate entitlement theorist and the full-blooded theorist are going to disagree about whether agents can know that **S** is consistent: the moderate theorist will claim that agents cannot know their entitlements, whereas the full-blooded theorist accepts that agents can in principle know their entitlements. Importantly, these two conceptions of entitlement are also going to disagree about whether the deductive consequences of entitlements together with evidentially warranted propositions can be known. One such example would be the case where some mathematical proposition p is independent from **S**, but provable in **S** *together* with the formalised statement expressing the consistency of **S**. With respect to this issue, the full-blooded entitlement theorist will claim that p can in principle be known by means of the new proof provided in the extension of **S**, although it is a consequence of – *inter alia* – an entitlement. On the other hand, the moderate theorist will deny that p can be known as a consequence of the relevant proof. To put it succinctly, the full-blooded conception understands entitlements as possibly extending the scope of what we can know. On the other hand, the moderate conception understands entitlements as only enriching the scope of what we can rationally believe. Finally, we should point out that the moderate and full-blooded conceptions agree that the deductive consequences of evidentially warranted beliefs

[4]As pointed out in chapter 1 for the weak conception this question does not even arise since it denies that entitlement is a warrant to believe propositions. Finally, it should be noted that the moderate conception should not be confused with the position called *moderatism*, presented for instance in (Coliva, 2015).

can amount to knowledge.

In the context of foundational projects in mathematics, there is no explicit mention or discussion neither of the moderate nor of the full-blooded conception of entitlement. However (and this is going to be speculative), Shapiro seems to endorse a moderate conception of entitlement when discussing Wright-style entitlement for cornerstones:

> Wright argues that a rational agent is only entitled to 'trust' [...] For Wright, rational trust in a proposition p is a state distinct from both the adoption of p as a working hypothesis and outright belief in (or knowledge that) p. [...] I would say instead that [the agent] is entitled to outright belief in [the cornerstone]. (Shapiro, 2011, p. 146)

Shapiro seems to endorse (at least) a moderate conception since he clearly states that entitlement is a warrant to believe (the relevant propositions).[5] In contrast to Shapiro, Horsten seems to endorse a full-blooded conception of entitlement. In his recent article, Horsten (2021) discusses several issues within the epistemology of consistency and adopts the framework of Wright's cognitive projects.[6] Towards the end of his discussion, he makes the following claim:

> *If you are not in a Gettier situation*, then you have more than justified true belief in the axioms of [arithmetic] (and in theorems of [arithmetic]): you *know* them. And then, after your process of reflection, you have acquired more than an epistemically entitled true belief in [the consistency

[5]He endorses at least a moderate conception because he does not tell us much about whether agents can be in a position to know (the relevant) cornerstones or not.

[6]The same framework has been adopted in chapter 1, and in (Wright, 2004), (Wright, 2012), (Wright, 2014), (Wright, 2016), and (Pedersen, 2016, 2021).

> of arithmetic]: you *know* this proposition. You have acquired knowledge of a cornerstone proposition of your cognitive project. (Horsten, 2021, p. 750, emphasis in the original text)

As already mentioned, to ascribe the two conceptions of entitlement to these two philosophers is to some extent speculative, since there is no mention or discussion of these positions in their work. However, even if speculative, this seems at least to suggest that philosophers do rely on such positions (possibly implicitly) when discussing crucial issues in the epistemology of mathematics.

Our aim here is to provide a preliminary defence of the moderate conception against important worries presented by Smith (2020). As pointed out previously, it is essential to provide such a preliminary defence for the following reason: if Smith's argument that the moderate position is inherently incoherent, the only position available, within the context of entitlement theories, would be the full-blooded conception. So, a defence of the moderate conception is essential to provide an alternative position to the entitlement theorist. However, one might wonder why we bother defending the moderate position: one would argue that the full-blooded conception is simply better since it gives agents a warrant to rationally believe their cornerstones and even to know them. However, as Smith points out, the full-blooded conception faces an additional issue:

> If one acknowledges a [sceptical] hypothesis to be epistemically possible, then to simply reject it without investigating it or unearthing any evidence against it would seem, in general, to be a dogmatic, irresponsible way of proceeding. To claim that there is a special epistemic commodity – entitlement – that we all possess and that makes this conduct somehow

> acceptable in the case of sceptical hypotheses can seem like pure wishful thinking. Call this the too-good-to-be-true problem. (Smith, 2020, p. 282)

Let us be explicit about this issue: all entitlement theories face some version of the "too-good-to-be-true problem": the moderate conception faces the issue of explaining why rational belief in the cornerstones is responsible. However, as pointed out in chapter 1, it follows from acceptable epistemic norms of our epistemic practices that we should believe in the cornerstones. On the other hand, the full-blooded conception faces a strong version of the "too-good-to-be-true problem": the full-blooded conception makes the notion of entitlement quite strong, indeed as strong as evidential justification when it comes to knowledge. As Smith (2020, p. 285) points out, according to the full-blooded conception, "entitlement effectively provides us with all of the epistemic goods that earned justification does – what separates entitlement from earned justification is its source, and not its strength or value". There are good reasons to think this position might fail to be responsible and be too committed. With the "too-good-to-be-true problem" in mind, the weaker, moderate conception is *prima facie* more acceptable.

However, is the moderate conception coherent to begin with? Smith (2020) argues that the moderate conception of entitlement should be rejected because it is incoherent with *prima facie* natural principles of knowledge closure. If Smith's argument were correct, there would be no dispute to adjudicate; following Smith, the only available position for the entitlement-based epistemology would be the full-blooded one. For this reason, it is essential to provide (at least) a preliminary defence of the

moderate conception. The remainder of this chapter is going to focus precisely on this. The following section presents Smith's main worry.

2.3 Closure and the Argument against the Moderate Conception[7]

Entitlement-based epistemological theories concede to the sceptic that there is no warrant, given the standard of evidence in our project, to believe the (relevant) cornerstones. Nevertheless, the entitlement theorist argues that agents have a non-evidential warrant, an entitlement, to believe the cornerstones. In addition, the moderate conception concedes that agents cannot know the (relevant) cornerstones. So, even if agents justifiably, by means of an entitlement, believe a true cornerstone and (some relevant) anti-Gettier conditions are in place, this belief does not amount to knowledge. Finally, this is the *only* concession that the moderate entitlement theorist makes: they claim namely that for any ordinary proposition p (i.e., not a cornerstone) of the (relevant) subject matter, evidentially warranted, true belief that p does amount to knowledge, if (some relevant) anti-Gettier conditions are in place. Focusing on this, Smith considers as a case to argue against the moderate conception the ordinary proposition that *my friend is upset* and the cornerstones that *there are other minds*:

> Many of the propositions that we are intuitively in a position to know will entail the propositions to which we are putatively entitled. On the moderate conception, I am in a position to know that my friend is feeling upset, given appropriate evidence and conditions, but I am not in

[7]This section follows ideas from (Zicchetti, 2022b).

> a position to know that there are other minds, even though the former obviously entails the latter. Consider the following principle: If I know a proposition p, and p obviously entails p, then I must be in a position to know p. The principle is very natural – it is odd to think that I could know p, but not even be in a position to know an obvious consequence q. And yet, if we adopt the idea of moderate entitlement, then this principle must be abandoned. (Smith, 2020, pp. 290-291)

Let us explain the relevant parts of the quote: Smith claims that, according to the moderate conception, we are in a position to know that *our friend is upset*. However, the moderate conception endorses that we cannot know that *there are other minds*, although the proposition that my friend is upset *obviously entails* that there are other minds. To put it more explicitly, according to Smith, the moderate entitlement theorist endorses the following claim, for any ordinary proposition p, cornerstone q and any agent I (*No Cornerstone Knowledge*):

(NCK) I can know p but cannot know q, although p obviously entails q.

From the background of (NCK), Smith argues that (*Simple Closure*) must be abandoned (for any two propositions p and q and any agent I):

(SC) If I knows p, and p obviously entails q, then I is in a position to know q.[8]

However, many authors have pointed out that (SC) is inadequate for the following reason: if an agent has no grip whatsoever on the fact that p obviously entails q, it is highly implausible that she nevertheless must be in a position to know that q.[9]

[8]Principles of knowledge closure have been extensively investigated. See for instance (Dretske, 2005, 2006) and (Alspector-Kelly, 2019) for some well-known arguments against closure. For defences of closure, see for instance (Feldman, 1995), (Hawthorne, 2004; Hale and Wright, 2005) and (Williamson, 2000). See (Luper, 2020) for an introduction to this topic.

[9]See (Alspector-Kelly, 2019).

However, since Smith argues that the principle of knowledge closure is quite natural and should be accepted, I am going to be charitable and assume that he has a more refined, acceptable principle in mind, such as (*Refined Closure*):

(RC) If I knows p, and p obviously entails q, then, if I competently deduces q from p, so that I comes to believe q based on the competent deduction from p, while retaining her knowledge that p throughout, then I is in a position to know q.

Principles similar to (RC) have received extensive attention (also in the epistemological literature also outside of the context of entitlement theories).[10] (RC) improves on simple closure; it avoids the obvious worry that we had against (SC) by requiring that the agent competently deduces q from p. Moreover, (RC) does not claim that agents must know all the logical consequences of p. (RC) simply captures the weaker intuition that agents can in principle extend what they know by performing (competent) deductions.

For the moment quite informally, Smith argues that the moderate conception is untenable because it is incoherent with principles such as (RC). In other words, the moderate conception is untenable because it is at odds with the natural idea that we can always in principle extend what we know by performing competent deductions (or inferences). The following section presents and discusses Smith's worry in more detail. It focuses on the two cornerstones that *there are other minds* and that *there is an external world* and introduces so-called I-II-III arguments.[11] After discussing the actual examples in (Smith, 2020), we will attempt to provide a I-II-III arguments

[10]This principle is sometimes called "Intuitive Closure". For discussions of this principle, see for instance (Hawthorne, 2004), (Williamson, 2000), (Silins, 2005), (Tucker, 2010) and (Alspector-Kelly, 2019).

[11]This discussion will follow my (Zicchetti, 2022b).

for (our relevant) mathematical cornerstone that (our relevant foundational theory) **S** is consistent. We do so because (to my best knowledge) there is no worked-out example of a I-II-III mathematical argument. Moreover, there are some difficulties in constructing a I-II-III argument for the proposition that **S** is consistent (more details in the next section). We aim to construct the best case for a Smith-style I-II-III argument in the mathematical case and to argue that the moderate conception is nevertheless coherent.

2.3.1 I-II-III arguments

Probably the most (in)famous example of a I-II-III argument is Moore's "proof" of the existence of the external world. Here is a simple reconstruction of the argument's structure:[12]

[MOORE]

(I) I have hands

(II) That I have hands obviously entails that there is an external world

(III) Therefore, there is an external world.

Let us say a few words about the argument: in I-II-III arguments the proposition of type (I) is supposed to be an ordinary proposition, i.e., a proposition about the subject matter D. In contrast to this, the proposition of type (III) is supposed to be the target cornerstone (relevant for the project about D). Finally, the conditional in (II) is supposed to be an *obvious* entailment between (I) and (III). First, let us say

[12]This and other I-II-III arguments have been extensively investigated in the literature on scepticism. For a discussion of Moore's proof, see for instance (Wright, 2002, 2003, 2004, 2014). For the original argument, see (Moore, 1939).

here that this interpretation of (II) can be questioned (and probably also resisted): it is not entirely clear what it means for an entailment to be obvious. A way to make sense of the entailment between (I) and (III) is to understand it as a *conceptual truth relative to some given conception* or philosophical position. Although in full generality (II) is by no means obvious, it seems acceptable that *relative to a Moorean common-sense realism*, the entailment between (I) and (III) is obvious since it is a common-sense fact that hands are empirical objects, external to us, then there is an external world, i.e., empirical objects external to us if there are hands. The argument investigated by Smith is understood as such I-II-III argument:

[OTHER MINDS]

(I) My friend is upset

(II) That my friend is upset obviously entails that there are other minds

(III) Therefore, there are other minds.

Here again, the proposition of type (I) is supposed to be an ordinary proposition, i.e., a proposition about the subject matter D. In contrast to this, the proposition of type (III) is supposed to be the target cornerstone (relevant for the project about D). Finally, the conditional in (II) is supposed to be an *obvious* entailment between (I) and (III). In general, these arguments are simple instances of a I-II-III template for any (relevant) ordinary p and cornerstone q:

[TEMPLATE]

(I) p

(II) p obviously entails q

(III) Therefore, q.

A few clarifications are in order: first of all, at no point we defend such arguments or the notion of conceptual implication involved. This section aims to charitably presuppose that Smith's argument is in good standing: that is, we assume with Smith that the moderate entitlement theorist would accept (II) as a (or the relevant kind of) conceptual truth, and that she would also accept that the I-II-III argument is valid.

Given the assumption that the moderate conception accepts the I-II-III argument, Smith shows that the moderate conception must reject (RC) in the following way: for an agent I, assume that I knows the proposition of type (I) that my friend is upset. Remember that this is not a problematic assumption even for the moderate conception, as she endorses that we can know ordinary propositions such as (I). Given the assumption that the moderate conception accepts the I-II-III argument, she can (competently) infer (III) from (I) and (II) simply by performing *modus ponens*. Assume also that by performing this inference, she comes to believe the proposition of type (III) (and that she still knows the proposition of type (I)). Now, by the relevant instance of (RC), which I repeat here:

(*) If I knows that my friend is upset, and moreover, the proposition that my friend is upset obviously entails that there are other minds, then, if I competently deduces that there are other minds from the proposition that my friend is upset, so that I comes to believe that there are other minds based on the competent deduction from the proposition that my friend is upset, while retaining her knowledge that my friend is upset throughout, then I is in a position to know that there are other minds.

Smith concludes that I is in a position to know that there are other minds. Let us say that I agree with Smith that the moderate conception is incoherent with (RC); as we said earlier, the moderate conception concedes to the sceptic that agents are not in a position to know (the relevant) cornerstones. However, with (RC) the moderate theorist seems to be forced to endorse that agents are in a position to know (the relevant) cornerstones, *de facto* implying a full-blooded conception of entitlement.

However, in the context of entitlement-based epistemologies, it turns out that (RC) is not as intuitive as it seemed – or as Smith claims. We can see that the proponent of the moderate conception has a good and crucially *independent* reason to argue that (RC) *fails* in cases such as [MOORE] and [OTHER MINDS]. We can see that (RC) fails in the I-II-III arguments as a result of the failure of the principle of (*Evidential Transmission*):

(ET) If p obviously entails q, and w is an evidential warrant for p, then w is an evidential warrant for q.

The moderate entitlement theorist claims that evidential warrant is not transmitted by means of conceptual implication. This claim is motivated by the intuition that whatever evidential warrant w (relative to the chosen set of procedures in the project) agents have for the ordinary proposition p, the conceptual implication between p and q *does not imply that w is an evidential warrant for the cornerstone*! This idea is not new and has been extensively investigated: as both Wright (2002, 2003) and Dretske (2005) point out, failure of principles such as (ET) is an absolutely pervasive phenomenon.[13] In the case of [MOORE], the natural intuition behind the

[13]See for instance (Wright, 2002) and (Dretske, 2005). Although they discuss the same issue, unfortunately, Wright calls principles such as (ET) 'transmission principles' whereas Dretzke calls them 'closure principles', which sometimes generates some unnecessary confusion.

argument for the failure of (ET) is the following: although *agents perceive* hands, and although having hands obviously entails that there is an external world – relative to common-sense realism – this does not imply (by any means) that *agents perceive* that there is an external world. The moderate entitlement theorist argues for the failure of (RC) closure, by claiming that (ET) fails in the case *where q in the relevant implication is a cornerstone*. And since no evidential warrant can be transmitted to the cornerstone and evidential warrant is necessary for knowledge, (RC) fails. This to be a good preliminary response to Smith's first worry. This response is a concessive one: it concedes to Smith that the moderate conception is incoherent with (RC), however, it shows that moderate entitlement theorists have good reasons to reject (RC) in full generality due to the failure of (ET).[14]

Now, does the same worry apply in the mathematical case? In other words, can a mathematical I-II-III argument be constructed, such that (i) the argument can be accepted to be in good standing and (ii) the type (II) implication between an ordinary proposition and the relevant mathematical cornerstone is a conceptual truth relative to some conception or philosophical position? These questions do not have immediate or trivial answers. The following section introduces and discusses a preliminary attempt to construct such I-II-III argument for the cornerstone that (the relevant foundational theory) **S** is consistent. Importantly, at no point, we will argue for (or endorse) I-II-III arguments in mathematics; we will argue that *even if we grant that such I-II-III arguments are possible*, proponents of the moderate conception still have good independent reason (along Dretzke and Wright) to reject the relevant principle of knowledge closure. Finally, let us say that (to my best

[14]This is the same conclusion I reached in (Zicchetti, 2022b).

knowledge) there is no discussion of I-II-III arguments for cornerstones of mathematical projects, such as consistency statements. Therefore, most of this work will be speculative and quite exploratory.

2.3.2 A I-II-III Argument for Consistency?

This is an attempt to present a Smith-style worry for the mathematical moderate entitlement theorist. Focusing on some (foundational) project about the relevant subject matter D, I will focus on a I-II-III argument for the proposition that **S** is consistent. Quite clearly, the issue is whether we can find an ordinary proposition p, for which agents have *evidence* in the sense specified in the project, i.e., by means of a proof in **S**, such that we can build the relevant instance of [TEMPLATE]. First of all, it is important to repeat that the proponent of a Smith-style worry does not need to find a proposition p, warranted by a proof in **S**, such that p logically implies that **S** is consistent. We know that given the fact that we assume that **S** is acceptable and incomplete, the proposition that **S** is consistent can never be provable in **S**, so that it is logically independent of any of the ordinary propositions p provable in **S**. Therefore, if the entailment between the ordinary proposition and the cornerstone is understood as logical, then mathematical I-II-III arguments for the consistency of **S** (where **S** is our foundational theory in the foundational project) are impossible.[15] Importantly, the conditional in (II) is supposed to express some conceptual truth relative to some conception or philosophical position, which is (again quite importantly) an informal conception in the background. So, the proponent of the Smith-style worry has to provide a I-II-III argument with an obvious, conceptual entailment between (I) and (III), where (III) is the proposition that **S** is consistent.

[15]Thanks to Catrin Campbell-Moore for suggesting to be fully explicit about this.

This is a first attempt to present a I-II-III argument for the consistency of **S** that the proponent of a Smith-style worry might employ. Again, we do not endorse this argument. We only aim to provide one for the proponent of Smith-style worry, to investigate the consequences for the moderate conception of entitlement. The first instance of [TEMPLATE] could be of the following form:

(First attempt)

(I) n is not (the code of) a proof in **S** of a contradiction.

(II) That n is not (the code of) a proof of a contradiction in **S** obviously entails that **S** is consistent.

(III) Therefore, **S** is consistent.

We assume that the proponent of the I-II-III argument understands (I) informally, as the interpretation of the following formal statement in **S**: $\neg Proof_S(n, \bot)$, where informally $Proof_S(x, y)$ is a recursive predicate representing provability in **S** and is read as 'x is (the code of) a proof of y in **S**'. (III) is the informal proposition that **S** is consistent that the proponent of the I-II-III arguments reads off of the formal consistency statement: $\forall x \neg Proof_S(x, \bot)$. Finally (II) is supposed to be the conceptual entailment between (I) and (III). Clearly, this argument will not do; although this argument could be unacceptable for many reasons, one of the most obvious ones is that the implication in (II) is false: for (I) to possibly conceptually imply (III), we would need (at least) the collateral information that (I) holds for all n. Moreover, one would need to have the collateral information that proofs in **S** are standardly finite. A more acceptable version of the I-II-III argument could be of the following form:

(Second attempt)

(I*) For all n, n is not (the code of) a proof in **S** of a contradiction.

(II*) That for all n, n is not (the code of) a proof in **S** of a contradiction obviously entails that **S** is consistent.

(III) Therefore, **S** is consistent.

This argument seems to do better than the first attempt, insofar as (I*) tries to overcome the limitations of (I), by allowing (I*) to be a more general statement, stating that for all n, n is not (the code of) a proof of a contradiction. The idea is that for each n we have in **S** a proof that n is not (the code of) a proof of a contradiction. By reflecting on this fact, the relevant agent in the Smith-style worry makes the informal claim that for all n, n is not (the code of) a proof in **S** of a contradiction. It is crucial that this statement is informal. The informal quantifier in the natural language formulation is supposed to be "external", for if (I*) were to be understood internally, as formalised in **S**, it is again clear that the formalised version of (I*) does not imply that **S** is consistent. This is again for incompleteness reasons. Provided that (I*) is acceptable for the proponent of Smith-style worry, the next thing to do would be to provide a philosophical position or conception to argue that (I*) understood as an *informal claim conceptually implies that* **S** *is consistent*.[16] Let us focus on (II*) now. To claim that (II*) is a conceptual truth, the proponent of this Smith-style worry will have to present some philosophical arguments to provide the missing information (or in other words, to provide the relevant philosophical or conceptual context), so that (II*) can be seen as a conceptual truth. Here, we will

[16]Let us be fully explicit: we do not endorse that the fact that, for each n, **S** proves that n is not (the code of) a contradiction provides evidence for the informal claim (I*). However, we suspect that the proponent of a I-II-III argument in this style would have to reason in a way similar enough to my example.

not go into the details of providing such an argument: what we need here is simply that such philosophical arguments for the conceptual truth of (II*) are possible and available for the proponent the I-II-III argument. And such philosophical arguments are clearly possible. Of course, one would need to provide an independent philosophical justification and motivation for such an argument. Moreover, the question is going to arise – and rightly so! – about whether and why these philosophical arguments are in good epistemic standing. Nevertheless, this is not important for this discussion; it is only needed that the proponent of the Smith-style worry can provide a philosophical argument that provides the collateral information to explain how it is so that (II*) is a conceptual truth. And this is certainly possible.[17]

For this investigation, suppose that the proponent of a the I-II-III argument is able to argue for the (Second attempt)-version of the I-II-III arguments. In addition, suppose that the moderate entitlement theorist understands the I-II-III argument proposed by her adversary and agrees on (I*), (II*) and (III). Now, suppose that the following is the case, so that the antecedent of (RC) is satisfied:

- (i) The moderate entitlement theorist competently deduces (III) from (I*) and (II*) by means of the conceptual implication between (I*) and (III).
- (ii) For all n, the moderate entitlement theorist knows that n is not (the code of) a proof in **S** of a contradiction.
- (iii) By deducing (III) she comes to believe (III) and also retains knowledge (for each n) that n is not (the code of) a proof in **S** of a contradiction.

[17]We should point out that the antecedent of (II*) seems to require some kind of infinitary reasoning, because there is an infinite number of n, such that **S** proves that n is not (the code of) a proof in **S** of contradiction. This issue can be avoided, as the antecedent of (II*) can be formalised and expressed by finitary means in very weak arithmetical theories. See (Fischer, 2021) and (Łełyk and Nicolai, 2022).

Can the proponent of the I-II-III argument use (RC) to show that the moderate entitlement theorist must endorse that agents are in a position to know (III)? The answer is: no. According to the example, the agent knows, for each n, that n is not (the code of) a proof in **S** of a contradiction. That is, the agent knows a series of propositions, $\varphi_0, \varphi_1, \varphi_2 \ldots$, where each φ_i expresses that i is not (the code of) a proof in **S** of a contradiction. Therefore, (RC) cannot be used because in this I-II-III argument q is not obviously entailed by a proposition, but by a series of propositions. Now, the proponent of the I-II-III argument might be tempted to employ a stronger principle of closure of knowledge with respect to an obvious entailment between a collection of propositions Γ and a cornerstone q:

(RC^+) If I knows each p in Γ, and Γ implies q, then, if I competently deduces q from Γ, so that I comes to believe q on the basis of the competent deduction from Γ, while retaining her knowledge of each p throughout, then I is in a position to know q.

Of course, this principle could be thought of as the following: ($\mathsf{RC}^{<n}$) if Γ is restricted to finite collections of propositions. However, this is not enough for the proponent of the I-II-III argument, as she has to consider the case where Γ is an infinite collection of propositions, namely the series $\varphi_0, \varphi_1, \varphi_2 \ldots$. Let us call this principle (RC^{ω}). It seems that the proponent of the I-II-III argument is going to need such a principle. With this principle, she would be able to employ the following instance:

(K) If for each n I *knows* that n is not a proof in **S** of a contradiction – and (i), (ii) and (iii) obtain –, then I is in a position to know that **S** is consistent.

If the proponent of the I-II-III argument can argue for this principle (and if everything else she said is in good standing), then she can show that the moderate

entitlement theorist is going to end up – by means of (K) – endorsing that agents are in a position to know that **S** is consistent, thereby contradicting her moderate conception and *de facto* implying a full-blooded conception. Of course, the final crux of the whole issue is going to be – for the proponent of such I-II-III argument – to explain how agents can perform such inferences: this principle is quite strong and in the relevant case, it resembles the ω-rule (when Γ is not finite). Traditionally, philosophers have been arguing that such inferences cannot be followed or performed. However, Warren (2020) seems to suggest, against our traditional intuitions, that it might be possible to follow some kind of infinite reasoning.[18] However, we should point out that the proponent of the I-II-III argument can only formulate the relevant rule (RC^{ω}) – and also (K) – by finitary means. This can be done for instance in standard way by adopting *predicates* with the relevant meaning. Examples of such finitary formulations of somewhat infinitary informal reasoning are known from the literature on axiomatic theories of truth and have received considerable attention within the context of *truth* predicates, but also in the context of epistemic predicates such as *believability*.[19]

This should provide at least a preliminary sketch and attempt to spell out how the proponent of a I-II-III argument would reason, to threaten the moderate conception of entitlement. Finally, it is worth repeating that at no point this chapter endorses the I-II-III argument or the reasoning involved in it.

[18]The proponent of the I-II-III argument might need to follow a similar kind of reasoning as in (Warren, 2020). An investigation of Warren's claim and strategy requires independent attention.

[19]See for instance (Cieśliński, 2017b, 2018).

Finally, we can see that, *even if we grant all this to the proponent of the I-II-III argument*, the moderate entitlement theorist has a good independent reason to argue that (RC^+) fails. Similarly to the I-II-III arguments discussed in the previous section, the relevant principle of knowledge closure fails as a result of the failure of the principle of (*Evidential Transmission*) – expanded to treat the relevant case:

(ET^+) If the collection of propositions in Γ together obviously entails q, and for each p_i in Γ, w_i is an evidential warrant for p_i, then what combines the warrants for the collection of the propositions in Γ is an evidential warrant for q.

We can see that (ET^+) is a generalisation of (ET): whereas (ET) roughly claimed that a warrant w for a proposition p is transmitted to q by means of obvious entailment between p and q, (ET^+) claims that the warrants for each proposition p in Γ are transmitted to q by means of obvious entailment between Γ and q. As in the previous case, the proponent of the moderate conception claims that evidential warrant is not transmitted under this kind of conceptual implication, that is, that (ET^+) fails for this I-II-III arguments. In our mathematical case, the intuition is quite similar to the cases discussed by Dretzke: although (i) for each n, there is an **S**-proof of φ_n, so that each φ_n is evidentially warranted and (ii) (I*) conceptually entails (III), this does not imply by any means there is evidential warrant by means of an **S**-proof of the consistency of **S**. For there is none – if **S** is consistent to start with – by Gödel's Incompleteness theorem. It is the failure of (ET^+) that provides the proponent of the moderate conception with a reason to reject (RC^+).

This should show that there is at least no immediate worry that the moderate conception (in general or in mathematics) is incoherent. In the relevant cases, even if we grant that the proponent of the I-II-III argument can construct an acceptable argument – which is still debatable and rightly so – the proponent of the moderate

conception has good independent reasons to explain why the relevant knowledge closure principles must be rejected.

2.4 Conclusion

This chapter investigated the meta-epistemological question of whether the claim (made by the moderate conception of entitlement) that entitled propositions cannot be known is coherent. It aimed to provide a preliminary defence of the moderate conception against the worries presented by Smith (2020) of its untenability. To provide such defence was essential to preserve at least one alternative position to the full-blooded conception. This chapter argued that the moderate entitlement theorist has good reasons for her main claim and that moreover, she has good independent reasons to resist the worries presented by Smith. In contrast to the defence of the moderate position in the context of the philosophy of perception, which seemed to be quite simple, the situation concerning the mathematical context is much more complex. As pointed out, the proponent of the I-II-III arguments has to make several assumptions to argue that the moderate conception is incoherent. Of course, the proponent of the moderate conception can refute to go along at many points: as mentioned before, some of the assumptions made by the proponent of the I-II-III argument in mathematics need further motivation and moreover, this dissertation does not endorse any of the reasoning performed by the proponent of the I-II-III argument. Nevertheless, we showed that even if we grant the proponent of the I-II-III argument can coherently and forcefully make his case and justify all her assumptions, the moderate conception still has a good independent reason to reject the conclusion drawn by the proponent of the I-II-III argument. This provides an alternative to the full-blooded position for the entitlement theorist, who wishes to

provide a principled response to the sceptical challenge, but nevertheless wants a less committed position in the light of the "too-good-to-be-true" problem.

Chapter 3

Soundness Arguments for Consistency

3.1 Introduction

Chapters 1 and 2 introduced and discussed some relevant issues of a non-evidentialist epistemology of mathematics. The present chapter focuses on an additional, separate, relevant topic in the context of an epistemology of consistency (to be introduced in the next section). This chapter is devoted to an epistemological analysis of such arguments. To provide such analysis is of philosophical interest, at least for the following reason: many philosophers seem to share the intuition that soundness arguments for consistency are somewhat *epistemically defective*. Although this intuition seems to be correct, most of the epistemological narrative and intricacies behind why and to what extent these arguments are defective still need to be uncovered and investigated. However, one might wonder why bothering to provide an analysis of soundness arguments to explain why they are defective if virtually

everyone accepts that they are defective. Against this, it should be fairly clear that we should not be satisfied with our first (however strong) superficial intuitions. Our analysis aims to uncover significant implicit epistemological assumptions needed for the evaluation of such arguments as defective.

This chapter focuses on a standard understanding of the epistemic defectiveness of arguments known from the literature in general epistemology: the failure of *cogency*. We will say that a valid argument is cogent just in case agents can in principle acquire a warrant to believe its conclusion in virtue of (a) the argument's premises being warranted and (b) the argument being valid.[1] Although the issue of cogency or transmission has received a considerable amount of attention in general epistemology, these topics have not been investigated in the epistemology of mathematics yet.[2]

This chapter addresses the issue of cogency against the background of the two main positions about the *superstructure* of warrant: *conservativism* and *liberalism*. As we will explain later, these are positions about what enabling conditions must be in place for a warrant to play its justificatory role. These positions have been extensively investigated in the epistemology of perception. However, to my best knowledge, they have not been considered in the epistemology of mathematics yet.[3] The main claims of this chapter are the following:

[1]Sometimes philosophers call cogent arguments – in the introduced sense – *transmissive*. This topic has received extensive attention in the epistemology of perception. See for instance (Wright, 2002, 2003, 2012), (Pryor, 2000, 2004), (Dretske, 2005) (Silins, 2005, 2007), (Neta, 2007, 2010), (Coliva, 2008), (Tucker, 2010). For an introduction to the relevant issues, see (Moretti and Piazza, 2018). As pointed it out in the introduction of this dissertation, the terms 'warrant' and 'justification' are employed as synonymous.

[2]To my best knowledge, the only exception is Waxman (b).

[3]For a conservative, see Wright's position in (Wright, 2012). For a liberal, see (Pryor, 2000, 2004).

1. Liberalism evaluates soundness arguments as cogent.

2. Conservativism evaluates soundness arguments as non-cogent.

The end of the chapter briefly comments on some of the possible epistemological implications and philosophical ramifications of these claims.

3.2 The shared Intuition on Soundness Arguments

Soundness arguments for consistency have been discussed in philosophy at least since the 60s.[4] A soundness argument for the consistency of **S** is an argument for the consistency of **S**, which employs a *detour by arguing first* that **S** is sound, i.e., that all theorems of **S** are true. Informally, we can think of a soundness argument for the consistency of **S** as the following:

> If the axioms of **S** are true and the rules of inference of **S** are truth-preserving, then all theorems of **S** are true. Since all theorems of **S** are true, **S** does not prove any false statement. Therefore, **S** is consistent.[5]

Most philosophers agree that such arguments are valid and sound for most of our accepted theories. However, when it comes to the epistemic value of soundness arguments, most philosophers have a strong intuition that such arguments are in some sense *defective*.[6] Girard (1987, p. 226) claims that such arguments have little epistemic value. Dummett and Wright are more explicit about why the argument is defective. According to them, soundness arguments are not *informative*:

[4] To my best knowledge, one of the first discussions of a soundness argument for consistency is to be found in (Myhill, 1960).

[5] A similar version of this argument is presented by Dummett (1978).

[6] For completeness, one should acknowledge that Shapiro (1998, p. 505)) seems to be the only exception, as he seems to suggest that soundness arguments might have some explanatory force. The investigation of whether soundness arguments are explanatory would exceed the scope of this dissertation.

> [S]uch a general form of consistency proof cannot, of course, be expected to be genuinely informative." (Dummett, 1978, p. 194)

> "[A] 'genuinely informative' consistency proof must (presumably) do more than elicit, in this trivial way, an implication of the presupposition that the axioms of the system are true and its rules of inference sound." (Wright, 1994, p. 176)

Finally, more recently Piazza and Pulcini (2013) argued that *all* such arguments – which they call CvS-proofs (where 'CvS' stands for 'Consistency-via-Soundness') – are *ill-founded*:

> [T]he CvS-proof is uninformative in the sense of being ill-founded. Indeed, we spot some circularities which afflict the conceptual structure of the standard CvS-proof.

In their investigation, Piazza and Pulcini (2013) focus on the paradigmatic example where the theory **S** is some first-order axiomatisation of Peano arithmetic, **PA**. In this context, they focus on the soundness argument for the consistency of **PA** formalised in a suitable theory of truth **T** extending **PA**. More specifically, Piazza and Pulcini (2013) focus on the paradigmatic example where the soundness argument for the consistency of **PA** is formalised in the theory of truth **CT**.[7] This theory expands the language of a arithmetic with a truth predicate T, together with axioms governing its behaviour. As they point out, this theory of truth extends arithmetic with a truth predicate, which applies to (code of) sentences and formulas of the language of **PA**:

[7] Although Piazza and Pulcini (2013) call this theory PA^{Tr}, this theory is standardly known in the literature as **CT**. See for instance (Halbach, 2014), (Cieśliński, 2017b) for the standard formulation of **CT**. For this reason, I will simply use the name **CT** instead PA^{Tr}.

> In a nutshell, [**CT**] is obtained from **PA** by: (i) extending the language by means of the truth predicate $Tr(x)$ and (ii) adding a cluster of axioms á la Tarski for deductively managing the new predicate. (Piazza and Pulcini, 2013, p. 167)

As the authors correctly point out, extending arithmetic with suitable principles for a primitive truth predicate provides the required strength to prove 'the' global reflection principle for arithmetic, i.e., that everything provable in arithmetic is true. And from that, one can conclude that arithmetic is consistent.[8] Crucially, one has also to expand the arithmetical induction schema to allow instances of induction in the expanded language with the 'new' truth predicate.

After reconstructing and briefly surveying the formalisation of the soundness argument for the consistency of **PA**, Piazza and Pulcini (2013) conclude with the following remark concerning the question of whether this formalised version of the argument provides agents with a warrant to believe that **PA** is consistent:

> On the one hand, the acceptance of the consistency statement [for **PA**] crucially relies on the trustworthiness of the formal soundness expressed within [**CT**]; on the other, one needs to *assume* the consistency of [**CT**] in order to trust all its theorems and, specifically, the formal soundness itself. (Piazza and Pulcini, 2013, p. 168)

As pointed out earlier in the introduction of this dissertation, Piazza and Pulcini are onto something right. However, most of the epistemological narrative for evaluating the formalised version of the semantic argument still needs to be made explicit. Moreover, as we will point out later, their diagnosis does not seem entirely correct.

[8] For one example of such a proof, see (Halbach, 2014).

In the present investigation, we follow Piazza and Pulcini (2013) and consider their discussed paradigmatic case: to provide a soundness argument for the consistency of **PA**, where the premises of the soundness argument (and in particular the soundness statement for **PA**) are justified by proofs the theory of truth **CT**. As I will show, their diagnoses for this precise case of the soundness argument is not entirely correct.[9] However, before investigating this paradigmatic case of the soundness argument, the relevant notion of *cogency* has to be introduced. This is the focus of the following section.

3.3 Epistemic Defectiveness and Cogency

One natural way to understand the epistemic defectiveness of arguments is, informally, by means of the impossibility of employing them to acquire justification to believe their conclusion. Arguments that can be employed to acquire such justification for their conclusion are called *cogent*. More explicitly, we spell out cogency in the following manner (*Cogency*):

(cogency) An argument D is cogent just in case, (a) D is valid, (b) D's premises are warranted, and (c) one can in principle acquire a first-time warrant to believe D's conclusion *in virtue of* (a) and (b).[10]

[9]We should point out that Piazza and Pulcini (2013) explicitly focus on the case where **CT** expands **PA**, where the schema of induction is accepted for formulas of arbitrary syntactic complexity. However, it should be mentioned that (Heck, 2015) showed that the soundness argument for the consistency of **PA** can be formalised in **CT** over the weaker theory $\mathbf{I\Sigma_1}$, that is, **PA** but with induction schema restricted to Σ_1 formulas. This has been also pointed out by Leigh and Nicolai (2013). Moreover, Łełyk (2022, Corollary 56) showed that the soundness argument can be proved in $\mathbf{CT_0}$, that is, **CT** over $\mathbf{I\Delta_0}$, where $\mathbf{I\Delta_0}$ is like **PA** but with the induction schema restricted to bounded formulas. To provide an epistemological analysis of this version of the semantic argument, i.e., this argument formalised over $\mathbf{I\Sigma_1}$ (or even $\mathbf{I\Delta_0}$) is left open for the future.

[10]Sometimes an argument's cogency is understood as its ability to provide an *additional independent* warrant for its conclusion, or to provide an additional warrant to *raise our confidence* in

Let us unpack the notion of cogency a bit more. First, cogency is a principle about the possibility of warrant acquisition. In this sense, this principle is quite different to (*Closure*):

(closure) A valid argument D satisfies closure just in case, whenever D's premises are warranted, D's conclusion is also warranted.

An argument satisfies closure just in case whenever there is a warrant for each of D's premises, there is a warrant for D's conclusion. According to this understanding of closure, some arguments will satisfy closure, while nevertheless failing to be cogent.[11] For illustration, let us look at some valid arguments satisfying closure but failing to satisfy cogency. Clearly, all valid circular arguments fail to be cogent, while nevertheless satisfying closure, where a circular argument is an argument, whose conclusion appears as one of its premise(s). This is a trivial example of such arguments:

(i) p

(ii) p implies p

(iii) Therefore, p.

Where (i) and (ii) are the premises and (iii) is obtained by performing *modus ponens* from (i) and (ii). Now, assuming that (i) and (ii) are warranted, closure is trivially satisfied; trivially, there is a warrant for the argument's conclusion precisely because the argument's conclusion appears as one of the argument's premise(s). However, the argument fails to be cogent, because the justification for the conclusion cannot

its conclusion. For a discussion of cogency along these lines, see (Okasha, 2004), (Chandler, 2009) and (Moretti, 2010). To investigate this would exceed the scope of this dissertation.

[11] That closure and cogency are quite different properties has been recognised and made explicit for instance by Wright (2002) and Dretske (2005).

be acquired by virtue of (a) and (b). For if (a) is satisfied, i.e., the (i) is warranted, the conclusion (iii) is also warranted (because it is the same proposition as in (i)), without the need of (b) to obtain.[12]

Additionally, it is important to be explicit about the fact that cogency here is about warrant acquisition *in principle*: as Coliva (2008) – as well as Moretti and Piazza (2018) – point out, cogency is not meant to capture how agents *actually justify their beliefs*, or whether agents *de facto* are justified in their beliefs.[13] Cogency is about whether agents *can in principle* acquire a warrant to believe the conclusion of a valid argument, whenever (a) and (b) obtain. One could follow Coliva (2008) and Moretti and Piazza (2018) and say that (cogency) is about *propositional* justification and not about *doxastic* justification.[14] A few words of clarification are in order: propositional justification is a justification (that an agent might have) *to believe* a target proposition. On the other hand, doxastic justification refers to the justification agents have for *their beliefs*. Importantly, propositional justification to believe a target proposition p does not necessarily entail that one's belief that p is justified. This might happen for many reasons: a trivial case is the one where there is a propositional justification to believe p, however, the target agent never even considered p and therefore never formed the relevant belief that p. A second, non-trivial case is the one where there is a propositional justification to believe that p and the target agent has formed the belief that p, but nevertheless her belief is

[12]For the same explanation of why circular arguments fail to be cogent see (Moretti and Piazza, 2018).

[13]Thanks to Hannes Leitgeb for suggesting to make this point explicit.

[14]However, two philosophers investigate cogency with respect to doxastic justification: Silins (2005) and Tucker (2010). To investigate doxastic investigation would exceed the scope of this dissertation.

not justified. This happens for instance when the agent bases her belief wrongly.[15] The distinctions, intricacies and relations between these two types of justifications are not relevant to this investigation. What is important is to keep in mind that in this investigation cogency is discussed with respect to propositional justification.

There are many examples of philosophically significant non-circular arguments, which are putatively evaluated as failing to be cogent. Probably the most (in)famous example of a (putatively) non-cogent argument is Moore's "proof" of the existence of the external world – this argument has been presented in chapter 2.[16] The question that we are interested in now is whether one can argue in a principled way that the soundness argument for the consistency of arithmetic is cogent (resp. non-cogent), according to the understanding of (cogency) introduced here. The following section is devoted to this issue.

3.4 Is the Soundness Argument Cogent?

As pointed out in the introduction, a soundness argument for consistency is an argument aiming at showing the consistency of some target theory by employing the statement that the target theory is sound in some relevant step in the argument. Following the strategy employed also by Piazza and Pulcini (2013), we assume that the statements in the soundness argument for the informal claim that **PA** is consistent are warranted by proofs in a suitable theory of truth **CT** extending **PA**. Now we can ask whether the soundness argument is cogent in the following way:

[15] This is according to what is called for instance by Turri (2010) the 'orthodox view'.

[16] This argument has been extensively investigated in the literature on scepticism. For a discussion of Moore's proof see for instance (Wright, 2002, 2003, 2004, 2014). For the original argument see (Moore, 1939).

can agents acquire a first-time warrant to believe that **PA** is consistent in virtue of (a) the premises of the soundness argument being warranted and (b) the argument being valid?

At this point, we have (at least) two different epistemological positions about the superstructure of warrant, according to which this question can be investigated and answered: liberalism, and conservativism. Earlier on we said that these positions are about the superstructure of warrant. What we mean exactly by it is that these positions are about *what enabling conditions* have to be in place, for a warrant to have its justificatory force. For our present purposes, however, it is enough to formulate these positions with respect to what enabling conditions have to be in place for evidential warrant – by means of proofs in the truth theory **CT** – to have its justificatory force.

(CT-Conservativism) A **CT**-proof of p provides a warrant to believe that p just in case there is, *as an enabling condition*, a prior warrant to believe a number of propositions, whose negation is incompatible with **CT**-proofs constituting evidence to believe that p.

(CT-Liberalism) A **CT**-proof of p provides a warrant to believe that p just in case there is no prior warrant to believe in a number of propositions incompatible with **CT**-proofs constituting evidence to believe that p.

Let us point out that these positions are not new: they have been extensively investigated in the context of the epistemology of perceptual warrant. The formulation of conservativism and liberalism with respect to proofs in **CT** is a relativisation (or instance) of the respective conservative and liberal positions in general as for-

mulated by Neta (2010).[17] Let us be more explicit about the intuition behind these two positions. Liberalism and Conservativism are two opposing positions about what enabling conditions must be in place in order for warrants to play their justificatory role. On the one hand, the conservative stance has high standards with respect to what is needed for warrants to have their force. In the case of proofs in **CT** the conservative argues that such proofs are evidence in support of the relevant proposition (proved in **CT**) only if there are as an enabling condition warrants to believe a number of propositions expressing that **CT** is in *good epistemic standing*, where the negation of such propositions would be incompatible with proofs in **CT** having justificatory force. One obvious example of such a proposition expressing that **CT** is in good standing is for instance the proposition that expresses that **CT** is consistent (this has been extensively in chapters 1 and 2). On the other hand, the liberal stance is much less demanding with respect to what is needed for the existence of warrant given by proofs in **CT**: it only requires the absence of warrant to believe that **CT** is *in*consistent.

Once we have these two positions and the notion of (cogency), we provide a precise answer to the question of whether and why the soundness argument for the consistency of **PA** is cogent or not. To avoid confusion, let us repeat the informal soundness argument here:

If the axioms of **S** are true and the rules of inference of **S** are truth-preserving, then all theorems of **S** are true. Since all theorems of **S** are true, **S** does not prove any false statement. Therefore, **S** is consistent.

[17] As pointed out in an earlier footnote, these positions have been discussed for instance by Pryor (2000, 2004), Wright (2004, 2012), Coliva (2008, 2015) and Neta (2007, 2010).

What is important here is that in our paradigmatic case (following Piazza and Pulcini (2013)) evidence for the informal conclusion of the soundness argument that **PA** is consistent is given by proofs in **CT** of the relevant formalised statements.[18] In other words, we take the warrant to believe the premises of the soundness argument to be given by proofs in **CT**. Additionally, both the conservative and liberal share the following context:

(*Context*) The soundness argument is valid and moreover, its premises are all warranted by proofs in the (suitably formulated) theory of truth **CT**.

Given this context, *conservativism evaluates the soundness argument for the consistency of* **PA** *as failing to be cogent.* Let us analyse the conservative diagnosis in more detail. Within the specified context, conservativism makes the following claim:

1. Any proof in **CT** of a proposition p provides a warrant to believe p just in case there is, as an enabling condition, a prior warrant to believe – *inter alia* – the proposition that **CT** is consistent.[19]

Obviously, conservativism makes this claim for a proof in **CT** of any of the premises of the soundness argument – the claim in 1. is for any proof in **CT**. Now, focusing for simplicity on the statement expressing that **PA** is sound, i.e., that everything provable in **PA** is true, conservativism makes the following claim, which is just an instance of her first claim:

[18] An example of such proof can be found in Halbach (2014). Moreover, Piazza and Pulcini (2013) propose a sketch of the proof's idea.

[19] This is so because the consistency of **CT** is interpreted as being a minimal requirement for **CT** to be in good epistemic standing.

2. The proof in **CT** of the statement that **PA** is sound provides a warrant to believe that **PA** is sound only if there is a prior warrant to believe (at least) that **CT** is consistent.

From the second claim and the assumption made in (*Context*) that all premises of the soundness argument are warranted by proofs in **CT**, conservativism concludes with the following:

3. There has to be a prior warrant to believe that **CT** is consistent.

But then, since **PA** is a sub-theory of **CT** – this is given by the paradigmatic case chosen by Piazza and Pulcini (2013) and by the fact that the suitable truth theory extends arithmetic – the prior warrant to believe that **CT** is consistent is a warrant to believe that **PA** is consistent. So, conservativism can conclude the following:

4. There has to be a prior warrant to believe that **PA** is consistent.

This gives us that the soundness argument fails to be cogent; if the premises of the argument are warranted by means of proofs in **CT**, then there has to be a prior warrant for the argument's conclusion, namely for the statement that **PA** is consistent – this is by 1. through 4. Therefore, agents cannot acquire a first-time warrant to believe that **PA** is consistent in virtue of (a) the argument's premises being warranted and (b) the argument being valid. That is:

5. The soundness argument for the consistency of **PA** is not cogent.

In contrast to conservativism, liberalism evaluates that the soundness argument for the consistency of **PA** is cogent. It does so because liberalism makes the following claim:

1*. Any proof in **CT** of a proposition p provides a warrant to believe p just in case there is, as an enabling condition, *no* prior warrant to believe the proposition that **CT** is *inconsistent*.

As one can see, this is simply the 'liberal version' of 1., the conservative claim. Now, assuming that agents do not have any prior, independent warrant to believe that **CT** is inconsistent, **CT** proofs simply provide warrant to believe the propositions proved. The soundness argument for the consistency of **PA** is cogent, according to liberalism, simply because liberalism has no way of getting to claim 3.

Now, let us say a few words about these diagnoses. There is a sense, in which this analysis seems to do exactly what it should: the diagnosis of the soundness argument is analogous to the *conservative diagnosis* of Moore's argument for the existence of the external world provided by Wright (2002). On the other hand, the diagnosis provided by liberalism is analogous to the *liberal diagnosis* of Moore's argument provided by Pryor (2000, 2004). However, it should be added that, although liberalism and conservativism disagree about whether the soundness argument for the consistency of **PA** is epistemically defective or not – in the sense of failure of (cogency) – the two parties agree that the soundness argument is *dialectically ineffective*, insofar it cannot be employed to overcome doubt about whether **PA** is consistent. For suppose an agent has reasonable doubt that **PA** is consistent. Then she should be equally reasonably in doubt about whether **PA** is in epistemic good standing. For instance, she should have reasonable doubt that proofs in **PA** provide any warrant to believe the propositions proved. This is so because we assumed that the agent has reasonable doubt about whether **PA** is consistent. But then, the agent cannot overcome this doubt by employing the soundness argument, because

in our paradigmatic case the premises of the argument are warranted by proofs in **CT**, which extends **PA**! And in particular, the proof of the relevant soundness statements for **PA** employs *arithmetical* induction in the expanded language with the truth predicate. To put it succinctly: doubt that **PA** is consistent cannot be overcome by employing induction in an extension of **PA**. As an interesting fact, here the situation is analogous to the one described by Pryor (2000, 2004), where liberalism about perception diagnoses Moore's proof as cogent, but dialectically ineffective.

Before concluding, a few remarks are in order: Conservativism concludes that agents cannot in principle acquire a warrant to believe the conclusion of the soundness argument by virtue of (a) the premises being warranted and (b) the argument being valid. This does not mean that agents cannot acquire warrant to believe that **PA** is consistent. According to conservativism, the premises of the soundness argument are warranted just in case there is a prior warrant to believe (at least) that **PA** is consistent. However, whatever this warrant is, it cannot be acquired in virtue of the soundness argument. Secondly, the conservative diagnosis does not conclude that the soundness argument is circular, on the contrary. It may be helpful to restate the diagnosis provided by Piazza and Pulcini:

> On the one hand, the acceptance of the consistency statement [for **PA**] crucially relies on the trustworthiness of the formal soundness expressed within [**CT**]; on the other, one needs to *assume* the consistency of [**CT**] in order to trust all its theorems and, specifically, the formal soundness itself. (Piazza and Pulcini, 2013, p. 168)

It is important to see that CT-Conservativism does not claim that the consistency of **CT** has to be *assumed*. This is important, because according to the quoted pas-

sage, the soundness argument is *circular*, insofar as its conclusion, the proposition that **PA** is consistent, has to be assumed as one of the premises. In contrast to this, conservativism claims that a warrant to believe the argument's conclusion is needed, to have a warrant to believe the argument's premises. That is, conservativism diagnoses the soundness argument as non-circular.

There is a final brief remark that is worth making before concluding. This is not meant as a fully worked out argument, but as a suggestion for possible future investigations. This has to do with whether the soundness argument is *uninformative*. Possibly one could argue that the soundness argument is uninformative, insofar as it satisfies the so-called *information-dependence* template. Wright argues that sometimes arguments fail to be cogent because they satisfy the so-called *information-dependence* template.[20] Roughly, a valid argument D satisfies this template just in case there is a special kind of information dependence between the warrant for D's premises and the warrant for D's conclusion (*Information Dependence*):

(ID) A warrant provided by some body of evidence, E, to believe a proposition p is information dependent, just in case the existence of the warrant for p provided by E requires some collateral information, I.

One can argue that from the background provided by (CT-conservativism), the soundness argument satisfies the information-dependence, insofar as the warrant provided for the (relevant) premises in the soundness argument by means of a proof in **CT** *is information dependent*, that is, it depends on the existence of collateral information – the evidence for the prior warrant to believe that **CT** is consistent. In

[20]This is discussed for instance in (Wright, 2002, 2003, 2012).

other words, the *justificatory force* of proofs in **CT** depends on collateral evidence that **CT** is consistent.

3.5 Conclusion

This chapter aimed to provide the first steps into an epistemological analysis of soundness arguments. After presenting the shared intuition that soundness arguments are epistemically defective, it focused on the case investigated by Piazza and Pulcini (2013): the soundness argument for the consistency of **PA**. This chapter argued that the epistemological position called *conservativism* evaluates the argument as non-cogent. On the other hand, *liberalism* evaluates the argument as cogent. Nevertheless, both positions agree that the argument is dialectically ineffective. To my best knowledge, this analysis provides a first explanation of why and to what extent the soundness argument for the consistency of **PA** can be evaluated as epistemically defective. Moreover, when 'epistemically defective' is specified by (cogency), we have a clear way two evaluate the issue. Moreover, this analysis seems to provide further support for conservativism over liberalism (at least in the mathematical case) because the former seems to capture our informal philosophical intuitions. This also opens the possibility to apply this approach to other arguments in mathematics.

Part II

The Epistemological Value of Reflection Principles

Summary of Part II

The second part of this dissertation focuses on the epistemological investigation of soundness statements, where soundness statements for a given theory **S** are statements expressing that everything provable in **S** is true. Part II is split into two parts: chapter 4 is a revised version of my co-authored paper (Horsten and Zicchetti, 2021), and chapter 5 is a revised version of my (Zicchetti, 2022a).

Chapter 4

Chapter 4 introduces the relevant historical and philosophical context with concerning reflection principles and soundness statements. After a brief presentation of some technicalities, the chapter briefly surveys some important results in mathematical and philosophical logic about reflection principles and soundness statements. The aim is to survey some of the relevant results in philosophical logic about so-called reflection principles and to highlight their epistemological significance. In particular, it aims to show that the epistemic notion of acceptance of (or commitment to) a theory has been playing a crucial role in the philosophical argumentation for reflection principles and their iteration. Finally, this chapter will not go into technical detail about the results because it focuses on highlighting what is relevant for our epistemological investigation.

Chapter V

Chapter 5 focuses specifically and uniquely on the role played by a particular type of soundness statement, called the *global reflection principle*, in cognitive projects involving axiomatic theories of truth. In particular, it follows Fischer et al. (2017,

2019) and focuses on the thesis that such global reflection principles play a crucial role for the *trustworthiness* of our theories. This chapter has two main aims: to investigate a cluster of theories of truth in classical logic from the perspective of trustworthiness and to investigate the role of the trustworthiness of theories of truth in relation to epistemic norms of cognitive projects.[21]

First, it briefly presents the main technical results and philosophical claims provided by Fischer et al. (2017, 2019) about extensions of theories of truth with global reflection principles that theories of truth in non-classical logic are *trustworthy* – in a sense to be made explicit later. On the other hand, most theories of truth in classical logic are not trustworthy. Our aim is to provide a cluster of theories of truth that is trustworthy in the sense provided by Fischer et al. (2019) (details will be introduced in chapter 5). In a sightly technical section, we will show that the cluster of theories of positive truth (and falsity) formulated over classical logic is trustworthy – we will make this claim also formally precise. This contributes to the work done by Fischer et al. (2017), where the authors provided trustworthy theories of truth formulated over some weak non-classical logic. Moreover, this work defends theories proposed by (Leigh, 2016) and (Horsten and Leigh, 2017) against the worry that these theories might be untrustworthy.

After that, this chapter discusses some of the consequences with respect to the epistemology of cognitive projects involving axiomatic theories of truth. In doing so, it provides an argument from epistemology for the choice of theories of positive truth

[21]Some of this work is connected to the discussion in chapter 1.

over rivals theories of truth in classical logic. Finally, the chapter briefly addresses some difficulties and objections against theories of positive truth and falsity.

Chapter 4

Reflection Principles and Implicit Commitment[1]

4.1 Introduction

This chapter briefly surveys some of the relevant results in philosophical logic about so-called reflection principles and highlights their epistemological significance. The section "Mathematical reflection" briefly surveys some of the earliest results about the method of extending theories by reflection principles: it briefly surveys Turing's completeness theorem in (Turing, 1939), and Feferman's completeness theorems in (Feferman, 1962), but also his result on so-called *autonomous progressions* in (Feferman, 1964). This section aims to highlight the epistemic considerations made

[1]As pointed out at the beginning of this dissertation, this chapter is a revised version of the article "Truth, Reflection, and Commitment", published in the edited collection *Modes of Truth. The Unified Approach to Truth, Modalities, and Paradox* (Horsten and Zicchetti, 2021). The original article is co-authored with Leon Horsten. The ideas in that article – and hence in chapter 4 – originated from the many discussions with Leon. We both equally contributed to the ideas present in the article. Revisions are minimal: the phrasing and notation have been adapted to the context of this dissertation and when relevant, new literature has been considered. Moreover, one section of (Horsten and Zicchetti, 2021) has been omitted. For additional information about the origin of (Horsten and Zicchetti, 2021), see Publications.

by Feferman concerning the process of extending axiomatic theories with so-called reflection principles. Section "Reflecting on truth" investigates reflection principles in the context of axiomatic theories of truth. It surveys some more recent – but fairly well-known – results about reflection principles over theories of truth. Section "Reflecting on acceptance" surveys a different approach to implicit commitment, where a 'reflection principle' is understood more broadly, as an extension of the base theory by a new primitive notion suitably axiomatised. This section considers the work by Cieśliński (2017b) with his theory of believability. The conclusion of this chapter briefly comments on the fact that an epistemological analysis of reflection principles still needs substantive investigation. The first steps toward this type of investigation are taken in chapter 5.

A remark on the notation, conventions and technicalities: we will not go into technical details of the technical results, because the focus of this chapter is mainly to highlight what is relevant for our epistemological investigation. We keep the notation as standard as possible.[2] Concerning proof-theoretical background, some familiarity with a few basic formal systems of arithmetic, such as Peano Arithmetic **PA**, Elementary Arithmetic **EA** is going to be helpful. We will also presuppose some familiarity with some basic facts about Kleene's notation system $\mathcal{O}$ – when surveying some of the results. As pointed out, this chapter aims to highlight the philosophical and epistemological side of these results, and for that, the current presentation and level of formalism will suffice. The discussion of axiomatic theories of truth will assume some acquaintance with a handful of theories of truth, such

[2]For details concerning notation and some of the formalism about reflection principles see for instance (Franzén, 2004a,b). For notation relevant for theories of truth see for instance (Halbach, 2014) and (Cieśliński, 2017b).

as the compositional theory **CT**, the Kripke-Feferman system **KF**, and the Partial Kripke-Feferman system **PKF**. Some of these theories of truth will receive more attention in chapter 5. The following section focuses on theories formulated in the language of first-order arithmetic or extensions thereof, and at least as strong as **EA**. Moreover, when reasoning about arithmetical theories, we use common conventions and presuppose some familiarity with coding.

4.2 Reflection Principles and their Iteration

This section aims to introduce so-called proof theoretic reflection principles and the notion of a progression of theories, with some the relevant detail needed to understand the philosophical significance of the results to be surveyed in the rest of this chapter.

A proof-theoretic reflection principle for an arithmetical theory **S** is a formalised soundness statement for **S**: it expresses that everything provable in **S** is true. Since by Tarski's theorem of the undefinability of truth the language of arithmetic cannot define its own truth predicate, the soundness statement for our arithmetical theory **S** can only be approximated. We distinguish between the following types of reflection principles:

$$\mathsf{prov}_{\mathsf{S}}\ulcorner 0 = 1\urcorner \rightarrow 0 = 1 \qquad (\mathsf{CON}_{\mathsf{S}})$$

$$\mathsf{prov}_{\mathsf{S}}\ulcorner \varphi\urcorner \rightarrow \varphi \qquad (\mathsf{LRFN}_{\mathsf{S}})$$

$$\mathsf{prov}_{\mathsf{S}}\ulcorner \varphi\dot{x}\urcorner \rightarrow \varphi(x) \qquad (\mathsf{URFN}_{\mathsf{S}})$$

Here $\mathsf{prov}_{\mathsf{S}}(x)$ is a standard provability predicate for the given theory $\mathbf{S}$, reading informally as 'x is provable in $\mathbf{S}$'.[3] $\mathsf{CON}_{\mathsf{S}}$ is a so-called 'consistency statement' for $\mathbf{S}$. $\mathsf{LRFN}_{\mathsf{S}}$ and $\mathsf{URFN}_{\mathsf{S}}$ are respectively called local and uniform reflection principles for $\mathbf{S}$. The schema $\mathsf{LRFN}_{\mathsf{S}}$ is *local* because formulated for sentences, whereas $\mathsf{URFN}_{\mathsf{S}}$ is called *uniform* because it is formulated for open formulas (in this formulation with at most one free variable x).[4]

For a given theory $\mathbf{S}$ and a reflection principle R we let $\mathbf{R}[\mathbf{S}]$ be the theory resulting from adding R to $\mathbf{S}$. The iteration of reflection can be defined in the following manner:

$\mathbf{R^0}[\mathbf{S}]$ be $\mathbf{S}$;

for α a successor ordinal, $\mathbf{R}^{\alpha+\mathbf{1}}[\mathbf{S}]$ be $\mathbf{R}[\mathbf{R}^{\alpha}[\mathbf{S}]]$;

for λ a limit ordinal, $\mathbf{R}^{\lambda}[\mathbf{S}]$ be the union of all $\mathbf{R}^{\alpha}[\mathbf{S}]$ for $\alpha < \lambda$.

The first proof-theoretic results to be surveyed concern so-called *progressions of theories* generated via iteration of reflection principles. Before continuing with the presentation of the results, a few notions concerning Kleene's $\mathcal{O}$ should be introduced. We keep the technicalities to a minimum because we aim to survey the results and highlight some epistemological issues and aspects concerning them.

We let $|a|$ be the ordinal denoted by the ordinal notation a in Kleene's notation system $\mathcal{O}$. Ordinal notations are partially ordered by the relation $<_{\mathcal{O}}$. For two ordinal notations a and b, we have that $a <_{\mathcal{O}} b$ if and only if $|a| < |b|$.

[3] More precisely, $\mathsf{prov}_{\mathsf{S}}(x)$ is a Σ_1-formula, short for '$\exists x(\mathsf{prf}_{\mathsf{S}}(x, y))$', where $\mathsf{prf}_{\mathsf{S}}(x, y)$ is a Δ_0 formula expressing informally that x is a proof in $\mathbf{S}$ of y.

[4] The notation $\ulcorner \varphi \dot{x} \urcorner$ is a shorthand for $\mathsf{sub}(\ulcorner \varphi v \urcorner, \ulcorner v \urcorner, \mathsf{num}(x))$, informally standing for the result of substituting, in the code of the formula $\varphi(v)$, the code of the free variable v with the x^{th}-numeral.

A *path* P is a subset of $\mathcal{O}$ such that the following obtains:

(i) for any $a, b \in P$ either $a \leq_{\mathcal{O}} b$ or $b \leq_{\mathcal{O}} a$,

(ii) if $b \in P$ and $c \leq_{\mathcal{O}} b$ then $c \in P$.

For any $a \in \mathcal{O}$, a set $P = \{\, b \,|\, b <_{\mathcal{O}} a\}$ is called a path *within* $\mathcal{O}$. The *length* of a path P is the ordinal of the restriction of $<_{\mathcal{O}}$ to P. For any path P within $\mathcal{O}$, the order type of P, denoted as $|P|$, is less than ω_1^{CK}. A path P is a path *through* $\mathcal{O}$ if $|P| = \omega_1^{CK}$, where $\omega_1^{CK} = sup\{|a| : a \in \mathcal{O}\}$. The relation $<_{\mathcal{O}}$ is not recursively enumerable. However, for any a, the restriction of $<_{\mathcal{O}}$ to $\{\, b \,|\, b <_{\mathcal{O}} a\}$ is recursively enumerable.

A progression of a theory $\mathbf{S}$ is a primitive recursive mapping taking any ordinal notation a in some path in Kleene's ordinal notation system $\mathcal{O}$ to a Σ_1^0-formula φ_a that recursively enumerates the axioms of a theory $\mathbf{S}_a$, such that

1. $\mathbf{S}_0 = \mathbf{S}$;

2. $\mathbf{S}_{suc(a)} = \mathbf{S}_a + \mathbf{R^a}[\mathbf{S}]$;

3. $\mathbf{S}_{lim(a)} = \bigcup_{b<a} \mathbf{S}_b$.

Any transfinite progression yields a *progressive reflection sequence*: a sequence of theories of the form

$$\mathbf{S}_0, \mathbf{S}_1, \ldots \mathbf{S}_\omega, \mathbf{S}_{\omega+1}, \ldots \mathbf{S}_\alpha, \ldots,$$

where $\mathbf{S}_{\alpha+1}$ is an extension by reflection of $\mathbf{S}_\alpha$, and $\mathbf{S}_\lambda$, for limit ordinals λ, has as axioms the union of the axioms of earlier theories. The following section focuses

on surveying two main results: Turing's completeness theorem for consistency progressions and Feferman's completeness theorem for uniform reflection progressions.

4.3 Reflecting on Mathematical Theories

Turing investigated consistency progressions in his attempt to reduce or circumvent the incompleteness of arithmetic. He proved the following theorem (Turing, 1939):

Theorem 1. *For any true* Π^0_1 *sentence* φ *there is an* $a \in \mathcal{O}$ *such that* $|a| = \omega + 1$ *and* $\mathbf{S}_a \vdash \varphi$*. Moreover, there is a primitive recursive function that associates such an* a *with each true* Π^0_1 *sentence* φ*.*

Turing suggests that the transition from a theory $\mathbf{S}_a$ to $\mathbf{S}_{suc(a)}$ invokes some sort of reflection:

> We were able, however, from a given system to obtain a more complete one by the adjunction as axioms of formulae, seen intuitively to be correct, but which the Gödel theorem shows are unprovable in the original system; from this we obtained a yet more complete system by a repetition of the process, and so on. (Turing, 1939, p. 198)

However, the epistemological import of Turing's completeness theorem is limited. Theorem 1 only tells us that for any true Π^0_1 sentence φ there is a consistency progression with length $\omega + 1$, such that $\mathbf{S}_{\omega+1}$ proves φ. As Franzén (2004b) already pointed out, Turing's result does not provide us with a method of *recognising*, for any true Π^0_1 sentence φ, that it is true. Turing's proof associates with every true Π^0_1 sentence φ a consistency reflection sequence of length $\omega + 1$ that ends in a theory $\mathbf{S}_{\omega+1}$ proving φ. However, the axioms of $\mathbf{S}_\omega$ have a non-canonical definition; Turing

defines $\mathbf{S}_\omega$ in such a way that its consistency entails that φ is true. Even though Turing's definition of ω and "canonical" definitions of ω extensionally coincide, no $\mathbf{S}_n$ proves that this is so.[5]

In order to strengthen Turing's completeness result, Feferman employed progressions of uniform reflection. He proved (Feferman, 1962):

Theorem 2. *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 $S_a \vdash \varphi$.*

This result is known as Feferman's completeness theorem. His proof generates a *path* P within $\mathcal{O}$ of length $\omega^{\omega^{\omega+1}}$ such that the union of all theories associated with the notations in this path is arithmetically complete. As with Turing's completeness theorem, and for the same reasons, the epistemological import of Feferman's completeness proof is limited. Following Franzén, it would be false to say that Turing's and Feferman's results show how to eventually obtain every arithmetical truth by iterating reflection principles.[6]

4.3.1 Autonomous Progressions

The previously surveyed proofs shows that there is a sense in which progressions fail to capture how systems of a higher ordinal level are justified "from below". Kreisel argued that progressions should satisfy an additional *autonomy* condition: for every $\mathbf{S}_a$ in a progression, it should be provable in some $\mathbf{S}_b$ with $b <_\mathcal{O} a$ that a is

[5]For more on the philosophical significance of the use of non-canonical definitions see (Franzén, 2004a) and (Franzén, 2004b).

[6]Completeness depends on the choice of the path in $\mathcal{O}$. Feferman and Spector (1962) showed that there are paths *through* $\mathcal{O}$, such that corresponding uniform reflection progression does not even prove every true Π^0_1 sentence.

in $\mathcal{O}$ (Kreisel, 1960). A progression satisfying this additional criterion is called an *autonomous progression*.

Feferman and Schütte investigated autonomous progressions of predicative theories of analysis (Feferman, 1964), (Schütte, 1964, 1965). In particular, in (Feferman, 1964) Feferman investigated autonomous progressions via uniform reflection based on the systems **H** and **R**,[7] determining the *limit* of predicative reasoning.[8] These theorems are epistemologically more significant than the completeness theorems presented earlier. In contrast to the non-autonomous progressions, the autonomy condition seems to support the claim that one *can recognise*, by means of a proof in a previous stage of the progression, that for a limit a, a is an ordinal notation. Feferman seems to suggest that the extension of a theory **S** by addition of reflection principles might be the result of some process of reflection. He claims that accepting a reflection principle for **S** (and iterating this procedure) rests on our *attitude* towards **S**:

> In contrast to an arbitrary procedure for moving from $\mathbf{A}_K$ to $\mathbf{A}_{K+1}$, a reflection principle provides that the axioms of $\mathbf{A}_{K+1}$ shall express a certain trust in the system of axioms $\mathbf{A}_K$. (Feferman, 1962, p. 261)

We can see from the quoted passage that the process of extending theories with their respective reflection principles does not involve only mathematical intuition that the target theory is true; the extension of such theories involves some notion

[7] **H** is the extension of first-order Peano arithmetic **PA** with Kreisel's so-called *hyperarithmetic comprehension rule* (HCR): see (Feferman, 1964, p. 17) for Feferman's original formulation of the system **H** and of (HCR). **R** is a system of Ramified analysis, see (Feferman, 1964, p. 21 - 22,)

[8] In his (Feferman, 1964, p. 23, Theorem 6.10) Feferman determines the 'limit' of predicative reasoning to be the ordinal Γ_0. Here it is not important to go into the technical details concerning the so-called ordinal analysis of predicative theories.

of *trust* is the target theory. In later work, Feferman continued to emphasise that reflection principles involve some epistemic concepts:

> 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. (Feferman, 1991, p. 2)

Feferman here seems to suggest an epistemological route by means of reflection to extend what an agent trusts from a weaker base theory to stronger and stronger theories. Feferman claims that the acceptability of reflection principles is (in some sense) *implicit* in the acceptance of the base theory. This claim is now usually interpreted as the so-called *Implicit Commitment Thesis*. Informally and in general, this thesis claims that the acceptance of a theory **S** implies the acceptability of principles formulated in the language of **S** that are logically independent of **S**. Such principles are for instance consistency statements and other reflection principles. Logicians and philosophers have been quite successful in analysing what is implicit in the acceptance of such principles using proof-theoretic methods.[9] Moreover, recently philosophers focused purely on the epistemology of the implicit commitment thesis.[10] As pointed out earlier in this dissertation, chapter 1 can be seen as a partial take on the implicit commitment thesis: when focusing on consistency, the chapter argued that agents accepting some foundational theory **S** are epistemically obligated to

[9]For the most recent discussions and results with respect to issues in philosophical logic related to the implicit commitment thesis see for instance (Dean, 2014), (Nicolai and Piazza, 2018) and (Łełyk and Nicolai, 2022).

[10]Some examples of this more informal, epistemological are for instance (Galinon, 2014), (Fischer et al., 2019) and (Horsten, 2021).

believe the consistency of **S**.[11] This concludes the presentation of the historical and philosophical context. The following section surveys some of the main results concerning the iteration of reflection principles over axiomatic theories of truth. Our aim is to highlight the philosophical significance and epistemological aspects of the technical results to be surveyed.

4.4 Reflecting on Axiomatic Truth

Kreisel and Lévy (1968, p. 98) pointed out that the concept of *truth* is involved in the concept of reflection:

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

This was regarded as a problem:

> Literally speaking, the *intended* reflection principle cannot be formulated in **S** itself by means of a single statement. This would require a *truth definition* T_S, with a variable a over (Gödel numbers of, or, simply, over) formulas of **S**, and a definition of the proof relation $\mathsf{prov}_S(p, a)$ (read: p is (the Gödel number of) a proof of a in **S**). The reflection principle for **S** would be
>
> $$\forall p \forall a[\mathsf{prov}_S(p, a) \rightarrow \mathsf{T}_S(a)].$$

[11] For the details about this claim and arguments see chapter 1. To investigate the implicit commitment any further would exceed the scope of this dissertation.

> Such a truth definition T_S, does not exist [...] (Kreisel and Lévy, 1968, p. 98)

This difficulty was circumvented by *approximating* the intended reflection principle by means of the purely arithmetical principles LRFN_S and URFN_S. Alternatively, a primitive truth predicate T can be added to the language of arithmetic with the aim of formulating reflection principles more explicitly. Reflection principles were related to a philosophical discussion about the function and role of the concept of truth.

One seemingly essential role of truth is to express and reason with generalisations over statements. For this purpose, the truth predicate is understood as a device of quotation and disquotation. According to this understanding, the Tarski-biconditionals, i.e., formulae of the form $\mathsf{T}^{\ulcorner}\varphi^{\urcorner} \leftrightarrow \varphi$, play a pivotal role in theories of truth. One distinguishes between *typed* and *untyped* (or type-free) Tarski-biconditionals. In the typed case, the truth predicate is not itself allowed to occur in φ. For instance, if we start with **PA** as a base theory and add to **PA** the collection of all *typed* Tarski-biconditionals $\mathsf{T}^{\ulcorner}\varphi^{\urcorner} \leftrightarrow \varphi$ for $\varphi \in \mathcal{L}_{\mathsf{PA}}$, the resulting theory is called $\mathbf{TB}^{-}$.[12] If one wants to add to **PA** a collection of untyped Tarski-biconditionals, then, in order to avoid the liar paradox, one can either weaken the background logic, or restrict the collection of Tarski-biconditionals and preserve classical reasoning. One example of the former option is, for instance, to work in some weaker logic such as *Basic De Morgan logic* (BDM). This has been investigated in (Fischer et al., 2017). The truth theory formulated in BDM, where the Tarski-biconditionals

[12] If one allows arbitrary formulas φ containing the truth predicate T to occur in instances of the induction schema, the resulting theory is called **TB**.

are completely unrestricted, is called $\mathbf{TS}_0$.[13] On the other hand, if one aims to retain classical logic, there are several options for restricting the Tarski-biconditionals to avoid inconsistency. Here we mention and briefly discuss two possible restrictions. One possibility is to restrict the Tarski-biconditional to sentences φ in which the truth predicate only occurs *positively* (that is, in the scope of an even number of negation symbols). Adding this collection of biconditionals to **PA** results in the truth theory $\mathbf{PTB}^-$.[14] A similar option is to expand the language of the truth theory ($\mathcal{L}_{\mathsf{T}}$) with a primitive falsity predicate, thus generating the language $\mathcal{L}_{\mathsf{TF}}$. One then considers the sub-language $\mathcal{L}^+_{\mathsf{TF}}$, obtained by allowing the negation symbol from $\mathcal{L}_{\mathsf{TF}}$ only to prefix atomic arithmetical formulas. Moreover, we consider the truth biconditionals $\mathsf{T}\ulcorner\varphi\urcorner \leftrightarrow \varphi$ with φ restricted to $\mathcal{L}^+_{\mathsf{TF}}$, and the falsity biconditionals $\mathsf{F}\ulcorner\varphi\urcorner \leftrightarrow \overline{\varphi}$, where $\overline{\varphi}$ is the *dual* of φ. Adding these two collections of biconditionals to **PA** results in the theory $\mathbf{TFB}^-$. We will focus on theories of positive truth (and falsity) in chapter 5.

4.4.1 Compositionality and Implicit Commitment

The philosophical question now arises whether some such collection of biconditionals captures the *content of the concept of truth*. An affirmative answer to this question is defended, for instance, in (Horwich, 1990), (Halbach, 2001), (Horsten and Leigh, 2017). This position is called *disquotationalism*, as it endorses that the content of truth is captured by a simple and natural collection of Tarski-biconditionals, that is, by a disquotational theory of truth. If disquotationalism is correct, then the concept of truth is at bottom merely a device for quotation and disquotation. A standard

[13] Clearly, there are several non-classical logics that one can opt for, such as Strong, Weak Kleene logic. For background on these non-classical logics, see for instance (Priest, 2008). We will say a bit more on BDM in chapter 5.

[14] See (Halbach, 2014, section 19.3).

objection against this is that truth is *compositional*. According to this view, theories of truth should prove general, intuitive, semantic principles: that any conjunction is true just in case its conjuncts are both true, and so forth. However, these principles cannot be derived from a set of Tarski-biconditionals. The 'standard' typed compositional theory of truth is **CT**.[15] One example of a well-known, compositional typefree theory of truth in classical logic is **KF**;[16] one example of a compositional, typefree truth theory in non-classical logic is **PKF**.[17] A discussion of some of these theories – particularly of **KF** and some of its variants – is postponed to chapter 5.

According to the compositional intuition, disquotationalist views fall short of capturing the content of the concept of truth. This objection against disquotational truth theories applies to all the theories mentioned above: the 'slogan' would be that compositional truth outstrips disquotational typed truth. Evidence for this slogan is provided by giving *core* principles governing the concept of truth that disquotational theories cannot prove. Without further resources, it seems that there is no way out for the disquotationalist. At this point reflection principles entered the philosophical debate. The idea is that the compositional principles might be *implicit* in some collection of Tarski-biconditionals and that *reflection* can bridge the gap between disquotational and compositional truth. Quite interestingly, proof-theoretic investigations support this claim. In the typed context, Halbach observed that iterating uniform reflection over **TB** twice recovers typed compositional truth (Halbach, 2001, section 4):

Theorem 3. $\mathbf{RFN^2}[\mathbf{TB}] \vdash \mathbf{CT}$.

[15] See (Halbach, 2014, chapter 8).
[16] For presentations of this theory see (Cantini, 1989), (Feferman, 1991) and (Halbach, 2014).
[17] For a presentation of this theory, see (Halbach and Horsten, 2006).

This phenomenon extends to the classical typefree context (Horsten and Leigh, 2017, theorem 7):

Theorem 4. $\mathbf{RFN^2[TFB] \vdash KF}$.

Without doubting the importance of Theorem 4, some clarificatory remarks are in order: the version of **KF** used by Horsten and Leigh (2017) – although closely related to the usual formulation of **KF** – it is not outright equivalent to it. In the version of **KF**, derivable via two iterations of reflection from **TFB**, the compositional axioms are restricted to the positive fragment of the language, whereas in the case of the usual **KF** the compositional axioms are completely unrestricted. Therefore, although these two versions of **KF** are equivalent for the arithmetical part of the language, their truth predicate behaves differently. As we will also discuss and show in chapter 5, **TFB** and the version of **KF** adopted by Horsten and Leigh (2017) – called $\mathbf{KF_{pos}}$ in chapter 5 – can be consistently closed under unrestricted rules of *Necessitation* and *Conecessitation* for the truth and falsity predicates. In contrast to this, the well-known, standard version of **KF** is inconsistent with the addition of the two rules. Chapter 5 is devoted to the technical and philosophical investigation of $\mathbf{KF_{pos}}$ and **KF**.[18] It is of philosophical interest that the recovery of compositionality through reflection also extends to the type-free non-classical context (Fischer et al., 2017, corollary 1, section 3.2):

Theorem 5. $\mathbf{R^2[TS_0] \vdash PKF}$.[19]

Following the general idea that accepting a theory generates the possibility to accept stronger theories of which the acceptability is implicit in the acceptance of

[18] As already pointed out, these results and the related philosophical discussion about $\mathbf{KF_{pos}}$ and **KF** can be found in my (Zicchetti, 2022a).

[19] Due to the non-classical context, the uniform reflection principle for $\mathbf{TS_0}$ is formulated as a rule instead of an axiom. The details of this formulation are not relevant for this survey.

the weaker theory, we can see that, if one commits to disquotational truth, then one *implicitly* commits – via reflection – to compositional truth.

Iterating reflection does not only recover compositional principles from disquotational ones. As it is shown by Leigh (2016, theorem 1.4, theorem 1.5, section 1), iterating the process of reflection also increases the amount of provable transfinite induction. One can fix a natural notation system for ordinals up to and not including Γ_0 that can be presented as an *elementary ordinal notation system* in the sense of (Rathjen, 1997), and call it $\mathbf{O}$. Then both $\mathbf{O}$ and the ordering relation $\prec$ on ordinals defined by elements of $\mathbf{O}$ are definable in first-order arithmetic.

Definition 1. *(Transfinite induction). Let A be a formula.*

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

$$\mathsf{prog}(\lambda x A) \to A(t),$$

   where t is a notation in $\mathbf{O}$ for α, and $\mathsf{prog}(\lambda x A)$ 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 $\alpha < \Gamma_0$, the schema of transfinite induction up to α, $\mathsf{TI}_{\mathcal{L}}(< \alpha)$, is the collection of formulae*

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

Definition 2. *For a theory $\mathbf{S}$ and an (elementary) ordinal κ, let $\mathbf{S}^{\kappa}$ denote the extension of $\mathbf{S}$ by $\mathsf{TI}_{\mathcal{L}}(< \kappa)$.*

Definition 3. *For a theory* $\mathbf{S}$ *and (elementary) ordinal* κ*, let* $\mathbf{RFN}^{\kappa}[\mathbf{S}]$ *denote the theory* $\mathbf{EA} + \kappa$ *times iterated uniform reflection over* $\mathbf{S}$*.*

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

Theorem 6. *For all* $\kappa \in \mathbf{O}$ *with* $\kappa > 0$*:*

1. $\mathbf{CT}^{\varepsilon_\kappa} = \mathbf{RFN}^{\mathbf{1}+\kappa}[\mathbf{TB_0}]$*;*
2. $\mathbf{KF}^{\varepsilon_\kappa} = \mathbf{RFN}^{\mathbf{1}+\kappa}[\mathbf{TFB_0}]$.

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

Theorem 7. *For all* $\kappa \in \mathbf{O}$ *with* $\kappa > 0$*:*

1. *If* A *is an* $\mathcal{L}_{\mathsf{PA}}$*-formula provable in* $\mathbf{RFN}^{\mathbf{1}+\kappa}[\mathbf{TB_0}]$*,* $\mathbf{RFN}^{\kappa}[\mathbf{CT}]$*, or* $\mathbf{CT}^{\varepsilon_\kappa}$*, then* A *is a theorem of* $\mathbf{EA} + \mathsf{TI}(< \varepsilon_{\varepsilon_\kappa})$*.*
2. *If* A *is an* $\mathcal{L}_{\mathsf{PA}}$*-formula provable in* $\mathbf{RFN}^{\mathbf{1}+\kappa}[\mathbf{TFB_0}]$*,* $\mathbf{RFN}^{\kappa}[\mathbf{KF}]$*, or* $\mathbf{KF}^{\varepsilon_\kappa}$*, then* A *is a theorem of* $\mathbf{EA} + \mathsf{TI}(< \varphi_{\varepsilon_\kappa}(0))$*.*

The situation in the non-classical settings is similar. In (Fischer et al., 2017, proposition 3. subsection 3.3.) it is shown that two acts of uniform reflection over the theory called *Basic*, which is **EA** formulated in the language with the truth predicate $\mathcal{L}_{\mathsf{T}}$ with an induction rule for Δ_0^0-formulae and in BDM logic,[20] proves the principle of transfinite induction for the language $\mathcal{L}_{\mathsf{T}}$ for all ordinals up to and including ω^{ω}:[21]

[20] See (Fischer et al., 2017, section 2.2) for more details.

[21] Due to the non-classical settings, Fischer et al. (2017) formulate the version of uniform reflection needed for both theorems 8 and 9 – respectively (Fischer et al., 2017, Proposition 3)

Theorem 8. $\mathbf{R^2}[\mathbf{Basic}] \vdash \mathsf{TI}_{\mathcal{L}_\mathsf{T}}(\omega^\omega)$

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

Theorem 9. $\mathbf{R}^\omega[\mathbf{Basic}] \vdash \mathsf{TI}_{\mathcal{L}_\mathsf{T}}(< \omega^{(\omega^{\mathbf{2}})})$

In other words, transfinitely many iterations of uniform reflection over a non-classical truth theory still prove much less transfinite induction than just two iterations of uniform reflection over classical logic. This is because *Basic* is formulated in the non-classical logic BDM. Let us point out that some philosophers might interpret the mismatch between the result of "reflecting" over classical logic and of "reflecting" on the non-classical theories as a disadvantage of non-classical truth: one could argue that non-classical truth theories *cannot reproduce (possibly not even with reflection) the same mathematical reasoning that classical theories offer* (Halbach and Nicolai, 2017).[22]

This section's take-home message is that the employment of reflection principles to extend our accepted base theories results into a recovery of truth-theoretic principles and purely mathematical principles. If reflection principles are implicitly justified – by some version of Feferman's implicit commitment thesis – then compositional truth-principles and strong(er) principles of transfinite induction are also implicitly justified by the acceptance of the target base theory.

and (Fischer et al., 2017, Corollary 3) – differently. For the precise formulation of their uniform reflection see (Fischer et al., 2017, p. 2640).

[22]However, it should be noted that recent results in philosophical logic have been used to argue against the claim that non-classical truth is somewhat inadequate. For two novel investigations of this issue, see for instance (Fischer et al., 2021) and (Field, 2022). To discuss these results would exceed the scope of this dissertation.

4.4.2 The Global Reflection Principle

The reflection principles involved in the theorems discussed so far merely *approximate* the intended way of formalising soundness that was already articulated by Kreisel and Lévy (1968): 'the' *Global Reflection Principle* (GRP):

Definition 4. *The global reflection principle for a theory* **S**, *denoted as* $\mathsf{GRP}_{\mathsf{S}}$, *is the formula*

$$\forall x[\mathsf{sent}_{\mathsf{S}}(x) \wedge \mathsf{prov}_{S}(x) \rightarrow \mathsf{T}(x)].$$

If we look at theories formulated in non-classical logic such as $\mathbf{TS_0}$, then we see :

Theorem 10. *(Fischer et al., 2017, proposition 1) The uniform reflection principle and the global reflection principle are provably equivalent over* $\mathbf{TS_0}$.[23]

Since $\mathbf{TS_0}$ is arithmetically sound when uniform reflection is added, global reflection over $\mathbf{TS_0}$ is likewise sound. Moreover, this procedure can then consistently be repeated. However, the situation in classical logic is drastically different. The closure of classical truth theories under (GRP) for the whole language often forces some kind of inconsistency. This can either be outright inconsistency, or what is called *internal inconsistency*, i.e., the existence of a sentence φ, such that it is provable that $\mathsf{T}\ulcorner\varphi \wedge \neg\varphi\urcorner$. Although we will discuss the issue of closing theories of truth in classical logic under global reflection in chapter 5, it is illustrative to already mention some known facts about this issue: Halbach (2014) shows that **FS** is inconsistent with $\mathsf{GRP}_{\mathsf{FS}}$, where **FS** is the type-free, classical and fully compositional theory of truth Friedman-Sheard.[24]. Moreover, Fischer et al. (2019, p. 8) observed that the

[23]Due to the non-classical settings, Fischer et al. (2017, p. 2638) formulate the reflection principles as rules.

[24]For the first presentation of this theory see (Friedman and Sheard, 1987) and for a more recent one see (Halbach, 2014).

standard axiomatisation of **KF** is internally inconsistent with $\mathsf{GRP}_{\mathsf{KF}}$. Indeed, **KF** is internally inconsistent even with $\mathsf{GRP}_{\mathsf{FOL}}$, that is, global reflection principle for all sentences in the language $\mathcal{L}_{\mathsf{T}}$ provable in first-order logic. Some have interpreted this phenomenon to indicate that standard theories of type-free truth in classical logic are implicitly incoherent.[25] In what follows, we will consider a different procedure to make the implicit acceptance of a theory explicit.

4.5 Reflection and Acceptance

Instead of taking for granted the idea that proof-theoretic reflection principles express trust or acceptance, one might investigate the notion of acceptance of a given theory **S** 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* **S** must be made explicit. An attempt to do this is provided by Galinon (2014). In his investigation of the acceptability of consistency, Galinon argues for two key principles. The first of these is the *Principle of (first-person) Responsibility*:

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

Second, he endorses the following principle:

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

Using these principles, Galinon (2014, p. 329) develops the following argument for the acceptance of consistency statements. Suppose a rational agent unconditionally

[25] The issue of the coherence of classical truth will be discussed in more detail in chapter 5.

accepts a mathematical theory **S**. Then, using the Principle of Responsibility, she must accept "**S** is acceptable". And from this, using the second principle, the agent is rationally obliged to accept that **S** is consistent. The Principle of Responsibility is a demanding requirement. Nevertheless, it seems to capture a correct intuition behind the notion of acceptability of theories.[26] Despite the intuitiveness of Galinon's principles, one might still wonder whether these are too strong. One might wonder whether reflecting on one's acceptance of **S** might not, in some cases, lead one to abandon rather than to accept one's acceptance of **S**. Of course this does not exclude that there are cases where we reflect on our acceptance of a theory **S** and *legitimately* conclude that **S** 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.[27]

Cieśliński provides an alternative analysis of reflection on one's mathematical beliefs and proposes the following informal understanding of acceptance of **S** :

> For any sentence φ, if I believed that φ has a proof in **S** and I had no independent reason to disbelieve φ, then I would be ready to accept φ. (Cieśliński, 2018, section 4)

Cieśliński (2017b, 2018) provides an axiomatic theory of believability that employs the informal notion of acceptance presented in the quote above. He makes this notion of acceptance of **S** explicit by extending **S** to a new theory $\mathbf{S}^+$, which captures

[26] Chapter 1 argued for a similar position and claimed that agents are *epistemically obligated* to believe in the consistency of the target foundational theory. To discuss the similarities and distinctions between Galinon's approach and mine would exceed the scope of this dissertation.

[27] This seems to be the conclusion drawn by Horsten (2021), as he argues that agents are only *rationally permitted* to accept, on the basis of reflecting on a theory **S** that they already accept, reflection principles for **S**. To adjudicate this dispute would exceed the scope of this dissertation.

the informal notion expressed above. He does this by presenting a theory of *believability*, which extends the theory **S** that we accept with a fresh predicate $\mathsf{B}(x)$ for believability and with axioms that govern its behaviour. The thought is that when a person reflects on the implicit commitments involved in her acceptance of a theory **S**, she comes to accept a theory of believability $\mathbf{Bel}(\mathbf{S})^-$ over **S**.[28] Suppose we start with a theory **S**, formulated in a language $\mathcal{L}_{\mathsf{S}}$. Let $\mathcal{L}_{\mathsf{SB}} = \mathcal{L}_{\mathsf{S}} \cup \{\mathsf{B}\}$. And let **SB** be the theory which is just like **S** except that its schemata range over all formulas of $\mathcal{L}_{\mathsf{SB}}$. The theory of believability $\mathbf{Bel}[\mathbf{S}]^-$ is an extension of **KB** with the following axioms and rules (Cieśliński, 2018, definition 13.4.1):[29]

(Ax_1) $\forall \psi \in \mathcal{L}_{\mathsf{SB}}[\mathsf{prov}_{\mathsf{SB}}(\psi) \to \mathsf{B}(\psi)]$,

(Ax_2) $\forall \varphi, \psi \in \mathcal{L}_{\mathsf{SB}}[(\mathsf{B}(\varphi) \wedge \mathsf{B}(\varphi \to \psi)) \to \mathsf{B}(\psi)]$,

(Ax_3) $\forall \varphi \in \mathcal{L}_{\mathsf{SB}}[\mathsf{B}(\forall x \mathsf{B}(\varphi(x))) \to \mathsf{B}(\forall x \varphi(x))]$

$$(NEC)\ \frac{\vdash \varphi}{\vdash \mathsf{B}(\varphi)}$$

This is the new formulation of the theory of believability, proposed and discussed in (Cieśliński, 2017a). As it is mentioned in (Cieśliński, 2017a), in the presented version of the believability theory, (Ax_3) is considered to be an improvement to the following rule:

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

[28] Cieśliński also considers a believability theory $\mathbf{Bel}(\mathbf{S})$ over **S**, which I do not discuss here.
[29] In the interest of readability we are sloppy with the Gödel coding in what follows.

which was used in Cieśliński (2017b, 2018) original presentation.[30]

Consider the "weak" typed disquotational truth theory $\mathbf{TB}^-$, which is like the disquotational theory $\mathbf{TB}$ except that the truth predicate is not allowed to occur in the induction schema. Suppose that we accept $\mathbf{TB}^-$. Then if we make the acceptance of $\mathbf{TB}^-$ explicit via $\mathbf{Bel}(\mathbf{TB}^-)$, we recover compositional principles for typed truth (Cieśliński, 2018, p. 264):

Theorem 11. $\mathbf{Bel}(\mathbf{TB}^-) \vdash \mathsf{B}(\mathsf{CT})$,

where $\mathsf{B}(\mathsf{CT})$ consists of all sentences $\mathsf{B}(\varphi)$ such that φ is an axiom of $\mathbf{CT}$. In particular we obtain the believability of mathematical induction for $\mathcal{L}_{\mathsf{T}}$ from a situation where we only accepted induction for $\mathcal{L}_{\mathsf{PA}}$. Analogous results hold in typefree settings. Consider the typed disquotational truth theory $\mathbf{TFB}^-$, which is like $\mathbf{TFB}$ except that the truth predicate is not allowed to occur in the induction schema. Suppose that we accept $\mathbf{TFB}^-$. Then if we make the acceptance of $\mathbf{TFB}^-$ explicit via $\mathbf{Bel}(\mathbf{TFB}^-)$, we recover compositional principles for typefree truth (Cieśliński, 2018, p. 266):

Theorem 12. $\mathbf{Bel}(\mathbf{TFB}^-) \vdash \mathsf{B}(\mathsf{KF})$

4.6 Conclusion

Let us take stock: if we are committed to disquotational truth and this commitment is made explicit via a theory of believability, then this theory proves that the compositional principles for truth are indeed believable. It is significant that

[30] As pointed out in (Cieśliński, 2017a), this newer version with (Ax_3) is preferred over the older version, which was also discussed by Horsten and myself in (Horsten and Zicchetti, 2021).

the believability theory does not contain a factivity principle or rule for the believability predicate B. Indeed, the inference from the believability of a statement to the statement itself is understood as a *defeasible* rule. Nonetheless, the theory of believability provides us with an argument for the following: in the absence of independent reasons to disbelieve compositional principles for truth, we should be ready to accept them. In this sense, this account provide an argument for the thesis that our commitment to compositional truth is not greater than the commitment to disquotational truth principles. It would take us too far to give a detailed evaluation of Cieśliński's position, so we limit ourselves to a few cautiously critical remarks. Cieśliński argues that processes of reflection on one's acceptance of a theory $\mathbf{S}$ can be described as proofs in a believability theory $\mathbf{Bel}(\mathbf{S}^{-})$ for $\mathbf{S}$. But it is not clear that all principles of $\mathbf{Bel}(\mathbf{S}^{-})$ are in all circumstances correct or acceptable. In particular, for the same reasons as why Galinon's Principle of Responsibility might be too strong, it is not clear how the axiom (Ax_1) of $\mathbf{Bel}(\mathbf{S}^{-})$ is precisely motivated. A deeper philosophical and epistemological analysis of reflection and of believability is crucial to make progress about these issues.

Chapter 5

Global Reflection, Trustworthiness and Positive Truth[1]

5.1 Introduction

Chapter 4 introduced and surveyed some relevant results and philosophical claims concerning reflection principles. The present chapter focuses on the *global reflection principles*' epistemological role in cognitive projects involving axiomatic theories of truth. Fischer et al. (2017, 2019) argued that these principles play a crucial role for the *trustworthiness* of (formal) theories. Their investigation shows that theories of full disquotational truth (to be introduced later in this chapter) are – in a sense to be explained later – *trustworthy* and therefore should be preferred over so-called scientific theories of truth in classical logic. Their results suggest that theories employing a fully disquotational concept of truth over non-classical logic

[1] As pointed out at the beginning of this dissertation, this chapter is a revised version of the article "Cognitive Projects and the Trustworthiness of Positive Truth", published in *Erkenntnis* (Zicchetti, 2022a). I am the sole author of this article. Revisions are minimal: phrasing and notation have been adapted to the present context. Moreover, sections 5.2 and 5.3 have not been modified.

are epistemically superior to the classical, rival theories of truth. This epistemological conclusion is supported by the results provided by Fischer et al. (2017). Once the trustworthiness of a theory is understood – and made formally precise – as its consistency and internal consistency with global reflection principles, one can prove that theories of full disquotational truth are trustworthy: they are consistent and internally consistent if closed under global reflection principles. On the other hand, most natural scientific theories of truth in classical logic fail to be trustworthy.

This chapter has two aims: to provide a cluster of theories of truth in classical logic that is trustworthy by the standard set in (Fischer et al., 2019). The second aim is to analyse the epistemological role played by such theories in cognitive projects. The chapter has the following structure: section 5.2 introduces the relevant notions and conventions concerning arithmetic, theories of truth, and reflection principles. After that, subsection 5.2.2 briefly surveys the theories of truth in non-classical logic presented by Fischer et al. (2017), and their result (Proposition 1) together with their observation that for some natural theories of truth – axiomatised in classical logic – **S**, **S** is internally inconsistent if closed under *global reflection*. The main section introduces and investigates theories of positive truth and falsity, and shows that these are consistent and internally consistent with the global reflection principle. Section 5.4 is devoted to the epistemological investigation of theories of positive truth: after presenting the relevant context of cognitive projects, the cornerstones, and epistemic norms, it introduces the distinction between *full disquotational* and *scientific* truth, and the notion of *trustworthiness*. This chapter argues that theories of positive are trustworthy. Moreover, one has good reasons to accept trustworthy theories over rival classical theories. The remainder of section 5.4 considers some

worries and problems for the proponent of theories of positive truth, and suggests some possible ways to respond to these worries.[2]

5.2 Notation and Conventions

This section presents the notation and conventions adopted in this chapter.[3] We focus on **PA** and theories of truth (and falsity) extending **PA**. We assume $=, \neg, \wedge, \forall$ as primitive logical symbols (and take $\vee, \rightarrow, \leftrightarrow, \exists$ as standardly defined). We call the base language of arithmetic $\mathcal{L}_0$. Terms of $\mathcal{L}_0$ are built in the usual way from variables, the constant 0, and by the application of successor, $+$ and $\times$. For a truth theory over **PA**, its language is called $\mathcal{L}_{\mathsf{T}}$, and expands the arithmetical vocabulary with the addition of a unary truth predicate T. Similarly, a theory of truth and falsity over **PA** is formulated in a language $\mathcal{L}_{\mathsf{TF}}$ expanding the arithmetical vocabulary with additional unary truth and falsity predicates T and F. This chapter focuses on theories of type-free truth (and falsity), that is, theories where the truth (resp. falsity) predicate can also apply to (codes of) formulas of the language containing both the truth and the falsity predicates. Since coding works perfectly fine in **PA**, we use standard conventions. For an expression e, $\#e$ is the Gödelnumber of e, and $\ulcorner e \urcorner$ is the code of e, i.e., the term in the language $\mathcal{L}_0$ representing $\#e$.

For the language $\mathcal{L}_0$, we have the usual formulas representing syntactic properties: we use $\mathsf{ter}_0(x)$ for the set of (Gödelnumbers of) terms of $\mathcal{L}_0$, $\mathsf{ct}_0(x)$ for the set of (Gödelnumbers of) closed terms of $\mathcal{L}_0$, $\mathsf{var}_0(x)$ for the set of (Gödelnumbers of)

[2]However, a thorough investigation of the possible responses would exceed the scope of this chapter.

[3]This section follows (Zicchetti, 2022a): the main definitions and theorems have not been revised. For a standard reference to notation and conventions, see (Halbach, 2014).

variables, $\mathsf{form}_0{}^n(x)$ for the set of (Gödelnumbers of) formulas with at most n free distinct variables, $\mathsf{sent}_0(x)$ for the set of (Gödelnumbers of) sentences of $\mathcal{L}_0$, where a sentence is a formula with at most 0 free distinct variables. One represents syntactic properties of the language $\mathcal{L}_\mathsf{T}$ (resp. $\mathcal{L}_\mathsf{TF}$) similarly. We use $\mathsf{sent}_\mathsf{TF}(x)$ for the set of (Gödelnumbers of) sentences of $\mathcal{L}_\mathsf{TF}$. We use $\mathsf{val}(x)$ to represent the evaluation function VAL, which for each Gödelnumber $\#t$ of a closed term t, it returns $t^\mathbb{N}$, that is, the value of t (in the standard model). Variables $s, t, ...$ range over closed terms. Moreover, we use $\forall s...$ as short for $\forall x(\mathsf{ct}(x) \rightarrow ...)$. The following conventions are adopted: $\varphi(\dot{x})$ as a shorthand for $\mathsf{sub}(\varphi(v), v, \mathsf{num}(x))$, informally standing for the result of substituting, in the formula $\varphi(v)$ the x^{th}-numeral for the free variable v. We use $\ulcorner\varphi\dot{x}\urcorner$ as a shorthand for $\mathsf{sub}(\ulcorner\varphi v\urcorner, \ulcorner v\urcorner, \mathsf{num}(x))$, informally standing for the result of substituting, in the code of the formula $\varphi(v)$, the code of the free variable v with the x^{th}-numeral. Moreover, we employ the dot notation for the representation of the respective syntactic functions, such as $\dot{\neg}$, $\dot{\wedge}$ and $\dot{\forall}$.

We have the usual Σ_1-formula $\mathsf{prov}_\mathsf{PA}(x)$, expressing formal provability in **PA**, reading informally as 'x is provable in **PA**' and is short for '$\exists x(\mathsf{prf}_\mathsf{PA}(x, y))$', where $\mathsf{prf}_\mathsf{PA}(x, y)$ is a Δ_0 formula expressing informally that x is a proof in **PA** of y. It is assumed that provability in **PA** is standard and satisfies the well-known derivability conditions hold. Moreover, we adopt the following understanding of consistency and internal consistency: for any theory of truth and falsity **S** extending **PA**, **S** is consistent just in case there is no sentence φ such that $\mathbf{S} \vdash \varphi \wedge \neg\varphi$. Moreover, **S** is internally consistent just in case there is no φ such that $\mathbf{S} \vdash \mathsf{T}\ulcorner\varphi \wedge \neg\varphi\urcorner$.[4]

[4]This formulation of internal consistency is called 'T-consistency' by Friedman and Sheard (1988). One can formulate the notion of F-consistency analogously. However, in this investigation we only focus on T-consistency.

5.2.1 Reflection Principles

For clarity, this section repeats some of the information about reflection principles – overlapping with chapter 4. For some first-order theory of truth (and falsity) **S** (containing **PA**) formulated in the expansion $\mathcal{L}_{\mathsf{TF}}$ of the arithmetical language $\mathcal{L}_0$, a proof-theoretic reflection principle for **S** is a 'soundness statement' for **S**, i.e., a statement expressing that everything provable in **S** is true. In addition to the reflection principles formulated in the language of **S** – presented in chapter 4 – we have the following *global reflection principle* formulated employing the truth predicate T:

$$\forall x(\mathsf{sent}_{\mathsf{TF}}(x) \wedge \mathsf{prov}_{\mathsf{S}}(x) \rightarrow \mathsf{T}(x))^5 \qquad (\mathsf{GRP}_{\mathsf{S}})$$

Since this chapter focuses on type-free truth and falsity, $\mathsf{GRP}_{\mathsf{S}}$ is formulated unrestrictedly, so that it does not only express that sentences of $\mathcal{L}_0$ – provable in **S** – are true, but that sentences containing both occurrences of the truth and falsity predicate – provable in **S** – are true.[6] Unless explicitly specified, with $\mathsf{GRP}_{\mathsf{S}}$ is intended the unrestricted version.

5.2.2 Reflection over Non-classical Truth

This section aims to briefly present the result by Fischer et al. (2017) that theories of full of disquotational and compositional type-free truth formulated in non-classical logic are consistent and internally consistent with the global reflection principle. Fischer et al. (2017) reason about the theory called $\mathbf{UTS_0}$, which is an extension of Elementary arithmetic **EA**, where **EA** is the same as **PA**, with the only distinction

[5] This principle was originally formulated by Kreisel and Lévy (1968).

[6] Clearly, one could also formulate $\mathsf{GRP}_{\mathsf{S}}$ for a typed truth predicate. One example of such formulation would be $\mathsf{GRP}_{\mathsf{S}}$ only for provable sentences of $\mathcal{L}_0$.

that in **EA** the induction schema is formulated for Δ_0 statements.[7] The theory $\mathbf{UTS_0}$ is formulated in a double-sided sequent calculus over the logic called *Basic De Morgan logic*. Roughly, a sequent is an expression of the form $\Gamma \Rightarrow \Delta$, where Γ, Δ are finite sets of formulas. Informally, one treats the formulas preceding the sequent arrow '$\Rightarrow$' as assumptions. The formulas in the succedent are disjunctively joined to form a single conclusion. Basic De Morgan logic is a sub-system of classical logic; it is obtained from classical logic by weakening the usual clauses for negation.[8] The axiom schemata for the truth predicate of $\mathbf{UTS_0}$ are the following unrestricted disquotational principles:

$$\varphi(x) \Rightarrow \mathsf{T}^{\ulcorner}\varphi\dot{x}^{\urcorner} \tag{T1}$$

$$\mathsf{T}^{\ulcorner}\varphi\dot{x}^{\urcorner} \Rightarrow \varphi(x) \tag{T2}$$

For any given theory of truth $\mathbf{S}$ containing $\mathbf{UTS_0}$, the uniform and global reflection principles are formulated in the following manner (the formulation is adapted to the weaker logic):

$$\frac{\Rightarrow \mathsf{prov}_{\mathsf{S}}{}^{\ulcorner}\varphi\dot{x}^{\urcorner}}{\Rightarrow \varphi(x)} \tag{$\mathsf{WRFN}_{\mathsf{S}}$}$$

$$\frac{\Rightarrow \mathsf{sent}_{\mathsf{T}}(x) \wedge \mathsf{prov}_{\mathsf{S}}(x)}{\Rightarrow \mathsf{T}(x)} \tag{$\mathsf{WGRP}_{\mathsf{S}}$}$$

Fischer et al. (2017) show that any such theory $\mathbf{S}$ containing $\mathbf{UTS_0}$ is consistent and internally consistent with $\mathsf{WGRP}_{\mathsf{S}}$. This follows from the fact that $\mathbf{S}$ is consistent and internally consistent with $\mathsf{WRFN}_{\mathsf{S}}$, together with the following proposition:

Proposition 1. (Fischer et al., 2017, Proposition 1) *Let a theory* $\mathbf{S}$ *contain* $\mathbf{UTS_0}$. *Then* $\mathbf{WRFN[S]}$ *and* $\mathbf{WGRP[S]}$ *are identical theories.*

[7] The fact that they reason about **EA** won't be relevant for my investigation.
[8] This logic is presented in detail by Fischer et al. (2017, Section 2.1, Table 1.).

Fischer et al. (2017) point out that the equivalence between uniform and global reflection is lost for many theories of type-free truth in classical logic:

> There is an intuitive connection between uniform and global reflection: both are intended to express the soundness of the base theory. It turns out, however, that this connection is lost in the classical axiomatizations of Kripke's fixed point construction considered by Horsten and Leigh (2017). For **S** an axiomatization of Kripke's fixed point construction in classical logic, in fact, the result of adding $\mathsf{GRP}_{\mathsf{S}}$ to it determines a severe restriction of the class of acceptable models: all consistent fixed points are excluded, i.e., if $(\mathbb{N}, S)$ models **GRP**[**S**] with S a fixed point, then S is inconsistent. In contrast, **RFN**[**S**] can have models of the form $(\mathbb{N}, S)$ for S a consistent fixed point (in fact all consistent fixed points).[9] (Fischer et al., 2017, p. 2638)

As Fischer et al. (2019) show, the argument for the internal inconsistency of such theories **S** is straightforward: for a liar sentence λ, **S** proves – by classical logic – the following statement: $(\lambda \wedge \neg \mathsf{T}^{\ulcorner}\lambda^{\urcorner}) \vee (\neg\lambda \wedge \mathsf{T}^{\ulcorner}\lambda^{\urcorner})$, and from this it is straightforward to prove, in **GRP**[**S**] that the liar sentence is both true and untrue.[10] In **KF**,[11] the equivalence between uniform and global reflection breaks: **KF** is internally inconsistent with $\mathsf{GRP}_{\mathsf{KF}}$ but consistent and internally consistent with $\mathsf{RFN}_{\mathsf{KF}}$.

[9] As pointed out in (Zicchetti, 2022a), this is slightly misleading. Fischer et al. (2017) *prima facie* claim that the theories investigated by Horsten and Leigh (2017) are internally inconsistent with global reflection. However, proof in Fischer et al. (2017, Footnote 13, p. 2638) employs truth-theoretic principles *unavailable* in the theories investigated by Horsten and Leigh (2017). From a charitable reading of the quote, the authors cannot mean that the theories investigated by Horsten and Leigh (2017) are internally inconsistent with the global reflection principle.

[10] Although this is folklore, Fischer et al. (2017, Footnote 13, p. 2638) show this.

[11] This theory has been presented by Feferman (1991) and Cantini (1989).

The next section investigates whether the theories of positive truth and falsity considered by Horsten and Leigh (2017) and Leigh (2016) are consistent and internally consistent with global reflection.

5.3 Reflection over Classical Positive Truth and Falsity[12]

This section aims to show that the theories of positive truth and falsity proposed by Leigh (2016) and Horsten and Leigh (2017) are consistent and internally consistent with global reflection. Section 5.3.1 introduces the relevant theories and their models, and section 5.3.2 argues for the claim that these theories are consistent and internally consistent with global reflection. First, it shows that these theories are consistent and internally consistent if closed under the rules of *Necessitation* and *Conecessitation* (Propositions 5 and 6). After that, it proves that standard models of these theories – closed under Necessitation and Conecessitation – are models of global reflection (Theorem 13).

5.3.1 Theories of Typefree Positive Truth and Falsity

The theory of positive truth and falsity biconditionals **TFB** extends **PA**, and expands the language $\mathcal{L}_0$ of **PA** to the language $\mathcal{L}_{\mathsf{TF}}$ with fresh truth and falsity predicates T and F. For any given φ in $\mathcal{L}_{\mathsf{TF}}$, $\overline{\varphi}$ denotes the dual of φ. Duals are introduced recursively:[13]

$$\overline{\varphi} = \neg\varphi \text{ (for } \varphi \text{ in } \mathcal{L}_0 \text{ and atomic)} \qquad \overline{\neg\overline{\varphi}} = \varphi$$

[12]The content of this section is as in (Zicchetti, 2022a).
[13]This the definition in (Leigh, 2016, p. 576).

$$\overline{\varphi \wedge \psi} = \overline{\varphi} \vee \overline{\psi} \quad \overline{\varphi \vee \psi} = \overline{\varphi} \wedge \overline{\psi}$$

$$\overline{\forall x \varphi} = \exists x \overline{\varphi} \quad \overline{\exists x \varphi} = \forall x \overline{\varphi}$$

$$\overline{\mathsf{T}s} = \mathsf{F}s \text{ and } \overline{\mathsf{F}s} = \mathsf{T}s$$

The language $\mathcal{L}^{+}_{\mathsf{TF}}$ is the *strictly positive* fragment of $\mathcal{L}_{\mathsf{TF}}$. For any φ, the expression φ is in $\mathcal{L}^{+}_{\mathsf{TF}}$ expresses that φ is strictly positive, i.e., that any occurrence of the truth and falsity predicates T and F in φ are under the scope of no negation symbols. φ is negative otherwise. $\mathsf{sent}^{+}_{\mathsf{TF}}$ denotes the set of (Gödelnumbers of) strictly positive sentences in the language $\mathcal{L}_{\mathsf{TF}}$. The theory **TFB** extends **PA** with the following axiom schemata:

$$\mathsf{T}\ulcorner\varphi\urcorner \leftrightarrow \varphi \tag{TFB1}$$

$$\mathsf{F}\ulcorner\varphi\urcorner \leftrightarrow \overline{\varphi}, \tag{TFB2}$$

for all sentences φ in $\mathcal{L}^{+}_{\mathsf{TF}}$. That is, the T-biconditionals and F-biconditionals are restricted to strictly positive sentences. **TFB** is a theory of *local* disquotation because its biconditionals are formulated for sentences only. The theory of *uniform* positive disquotational truth and falsity, **UTFB**, extends **PA** with similar axiom schemata, although formulated for open formulas:

$$\mathsf{T}\ulcorner\varphi\dot{x}\urcorner \leftrightarrow \varphi(x) \tag{UTFB3}$$

$$\mathsf{F}\ulcorner\varphi\dot{x}\urcorner \leftrightarrow \overline{\varphi(x)}, \tag{UTFB4}$$

for open formulas $\varphi(x)$ in $\mathcal{L}^{+}_{\mathsf{TF}}$.

$\mathbf{KF_{pos}}$ is the theory of positive compositional truth and falsity and extends **PA** with the following axioms:

$$\forall s \forall t((\mathsf{T}(s \doteq t) \leftrightarrow \mathsf{val}(s) = \mathsf{val}(t)) \wedge (\mathsf{T}\dot{\neg}(s \doteq t) \leftrightarrow \neg(\mathsf{val}(s) = \mathsf{val}(t)))) \tag{KF1}$$

$$\forall s \forall t((\mathsf{F}(s \doteq t) \leftrightarrow \neg(\mathsf{val}(s) = \mathsf{val}(t))) \wedge (\ F\dot{\neg}(s \doteq t) \leftrightarrow \mathsf{val}(s) = \mathsf{val}(t))) \tag{KF2}$$

$$\forall x \forall y(\mathsf{sent}^{+}_{\mathsf{TF}}(x \dot{\wedge} y) \rightarrow (\mathsf{T}(x \dot{\wedge} y) \leftrightarrow \mathsf{T}x \wedge \mathsf{T}y)) \tag{KF3}$$

$$\forall x \forall y(\mathsf{sent}^{+}_{\mathsf{TF}}(x \dot{\vee} y) \rightarrow (\mathsf{T}(x \dot{\vee} y) \leftrightarrow \mathsf{T}x \vee \mathsf{T}y)) \tag{KF4}$$

$$\forall x \forall y(\mathsf{sent}^{+}_{\mathsf{TF}}(x \dot{\wedge} y) \rightarrow (\mathsf{F}(x \dot{\wedge} y) \leftrightarrow \mathsf{F}x \,\overline{\wedge}\, \mathsf{F}y)) \tag{KF5}$$

$$\forall x \forall y(\mathsf{sent}^{+}_{\mathsf{TF}}(x \dot{\vee} y) \rightarrow (\mathsf{F}(x \dot{\vee} y) \leftrightarrow \mathsf{F}x \,\overline{\vee}\, \mathsf{F}y)) \tag{KF6}$$

$$\forall x \forall y(\mathsf{form}(x) \wedge \mathsf{var}(y) \wedge (\mathsf{sent}^{+}_{\mathsf{TF}}(\dot{\forall} yx)) \rightarrow (\mathsf{T}\dot{\forall} yx \leftrightarrow \forall z \mathsf{T}(x\dot{z}))) \tag{KF7}$$

$$\forall x \forall y(\mathsf{form}(x) \wedge \mathsf{var}(y) \wedge (\mathsf{sent}^{+}_{\mathsf{TF}}(\dot{\forall} yx)) \rightarrow (\mathsf{F}\dot{\forall} yx \leftrightarrow \overline{\forall} z \mathsf{F}(x\dot{z}))) \tag{KF8}$$

$$\forall x \forall y(\mathsf{form}(x) \wedge \mathsf{var}(y) \wedge (\mathsf{sent}^{+}_{\mathsf{TF}}(\dot{\exists} yx)) \rightarrow (\mathsf{T}\dot{\exists} yx \leftrightarrow \exists z \mathsf{T}(x\dot{z}))) \tag{KF9}$$

$$\forall x \forall y(\mathsf{form}(x) \wedge \mathsf{var}(y) \wedge (\mathsf{sent}^{+}_{\mathsf{TF}}(\dot{\exists} yx)) \rightarrow (\mathsf{F}\dot{\exists} yx \leftrightarrow \overline{\exists} z \mathsf{F}(x\dot{z}))) \tag{KF10}$$

$$\forall x(\mathsf{T}^{\ulcorner}\mathsf{T}\dot{x}^{\urcorner} \leftrightarrow \mathsf{T}(x) \wedge \mathsf{T}^{\ulcorner}\mathsf{F}\dot{x}^{\urcorner} \leftrightarrow \mathsf{F}(x)) \tag{KF11}$$

$$\forall x(\mathsf{F}^{\ulcorner}\mathsf{T}\dot{x}^{\urcorner} \leftrightarrow \mathsf{F}(x) \wedge \mathsf{F}^{\ulcorner}\mathsf{F}\dot{x}^{\urcorner} \leftrightarrow \mathsf{T}(x)), \tag{KF12}$$

where $\overline{\wedge} = \vee$, $\overline{\vee} = \wedge$, $\overline{\forall} = \exists$ and $\overline{\exists} = \forall$.

This investigation focuses on standard models of these theories, that is, models of these theories expanding the class of standard models of arithmetic, $\mathbb{N}$. Models of theories of positive truth and falsity are called $\mathcal{L}^{+}_{\mathsf{TF}}$-structures. An $\mathcal{L}^{+}_{\mathsf{TF}}$-structure

$\mathfrak{M} = (\mathbb{N}, S_1, S_2)$ is an expansion of $\mathbb{N}$ with a set S_1, interpreted as the extension of the truth predicate, and a set S_2, interpreted as the extension of the falsity predicate. We want to obtain the extensions of the truth and falsity predicates by starting from two sets S_1 and S_2 by iterating a positive inductive operation on the pair (S_1, S_2), denoted by $\Gamma(S_1, S_2) = [\Gamma^+(S_1, S_2), \Gamma^-(S_1, S_2)]$, such that

$$\Gamma^+(S_1, S_2) = \{\#\varphi | \varphi \in \mathsf{sent}^+_{\mathsf{TF}} \text{ and } (\mathbb{N}, S_1, S_2) \models \varphi\} \cup \{\#\varphi | \varphi \notin \mathsf{sent}^+_{\mathsf{TF}} \text{ and } \varphi \in S_1\}$$

$$\Gamma^-(S_1, S_2) = \{\#\varphi | \varphi \in \mathsf{sent}^+_{\mathsf{TF}} \text{ and } (\mathbb{N}, S_1, S_2) \models \overline{\varphi}\} \cup \{\#\varphi | \varphi \notin \mathsf{sent}^+_{\mathsf{TF}} \text{ and } \varphi \in S_2\}.$$

If the first expansion of $\mathbb{N}$ is $(\mathbb{N}, \emptyset, \emptyset)$, i.e., with S_1 and S_2 being empty, then $\Gamma^+(S_1, S_2)$ and $\Gamma^-(S_1, S_2)]$ are the following:

$$\Gamma^+(S_1, S_2) = \{\#\varphi | \varphi \in \mathsf{sent}^+_{\mathsf{TF}} \text{ and } (\mathbb{N}, S_1, S_2) \models \varphi\}$$

$$\Gamma^-(S_1, S_2) = \{\#\varphi | \varphi \in \mathsf{sent}^+_{\mathsf{TF}} \text{ and } (\mathbb{N}, S_1, S_2) \models \overline{\varphi}\}.$$

For S_1 and S_2 to be possible candidates for the extensions of the truth and falsity predicates, the operation $\Gamma(S_1, S_2)$ must reach fixed points, i.e., points such that $\Gamma(S_1, S_2) = (S_1, S_2)$. To show that Γ reaches fixed points it suffices to show that Γ is monotone, i.e., that if the pair (S_1', S_2') extends the pair (S_1, S_2), then $\Gamma(S_1', S_2')$ extends $\Gamma(S_1, S_2)$. We need to show that if $(S_1, S_2) \leq (S_1', S_2')$, then $\Gamma(S_1, S_2) \leq \Gamma(S_1', S_2')$.[14]

Proposition 2. Γ *is monotone.*

Monotonicity follows from the fact that Γ is a positive inductive operation. For any positive statement φ, one easily shows that by the definition of $\leq$ and the def-

[14]For clarity, for any two sets A, B we understand $(A, B) \leq (A', B')$ as $A \subseteq A'$ and $B \subseteq B'$.

inition of Γ, if the code of φ is in $\Gamma^+(S_1, S_2)$, then the code of φ is $\Gamma^+(S_1', S_2')$.[15] One treats positive statements in $\Gamma^-(S_1, S_2)$ analogously. Negative statements are trivially taken care of by the definition of Γ.

For this investigation, one still needs to show that fixed points of Γ are exactly the models of positive truth and falsity. We show that the fixed points of Γ are exactly the models of **TFB** and **UTFB**.

Proposition 3. *Assume that* $S_1, S_2 \subseteq \omega$. *Then the* $\mathcal{L}^+_{\mathsf{TF}}$*-structure* $(\mathbb{N}, S_1, S_2)$ *is a model of* **TFB** (*and also of* **UTFB**) *if and only if* $\Gamma(S_1, S_2) = (S_1, S_2)$.[16]

Sketch. For the left-to-right direction, we assume that $(\mathbb{N}, S_1, S_2) \models \mathbf{TFB}$. To show that $\#\varphi \in (S_1, S_2)$ if and only if $\#\varphi \in \Gamma(S_1, S_2)$ we have two cases to take care of: (i) φ is positive; (ii) φ is not positive.

(i) If φ is positive, we have the following equivalences:

1. $\#\varphi \in S_1$

   if and only if $(\mathbb{N}, S_1, S_2) \models \mathsf{T}^\ulcorner\varphi^\urcorner$

   if and only if $(\mathbb{N}, S_1, S_2) \models \varphi$ (by the assumption that $(\mathbb{N}, S_1, S_2) \models \mathbf{TFB}$)

   if and only if $\#\varphi \in \Gamma^+(S_1, S_2)$ (by the definition of Γ^+ and by the fact that φ is positive and that $(\mathbb{N}, S_1, S_2) \models \varphi$).

2. $\#\varphi \in S_2$

   if and only if $(\mathbb{N}, S_1, S_2) \models \mathsf{F}^\ulcorner\varphi^\urcorner$

[15] This is essentially the proof by Halbach (2014, Lemma 19.13.).

[16] This follows the proof by Halbach (2014, Theorem 19.15). As in (Zicchetti, 2022a), the proof is adapted to the context with duals and the falsity predicate.

if and only if $(\mathbb{N}, S_1, S_2) \models \overline{\varphi}$ (by the assumption that $(\mathbb{N}, S_1, S_2) \models \mathbf{TFB}$)

if and only if $\#\varphi \in \Gamma^-(S_1, S_2)$ (by the definition of Γ^- and by the fact that φ is a positive and that $(\mathbb{N}, S_1, S_2) \models \overline{\varphi}$).

(ii) If φ is not positive, then the claim that $\#\varphi \in (S_1, S_2)$ if and only if $\#\varphi \in \Gamma(S_1, S_2)$ follows trivially from the definition of Γ.

For the right-to-left direction, we assume that (S_1, S_2) is a fixed point of Γ and reason about some arbitrary φ in $\mathcal{L}^+_{\mathsf{TF}}$. We have the following equivalences:

3. $(\mathbb{N}, S_1, S_2) \models \mathsf{T}^\ulcorner\varphi^\urcorner$

   if and only if $\#\varphi \in S_1$

   if and only if $\#\varphi \in \Gamma^+(S_1, S_2)$ (by the assumption that $\Gamma(S_1, S_2) = (S_1, S_2)$)

   if and only if $(\mathbb{N}, S_1, S_2) \models \varphi$ (by the definition of Γ^+)

4. $(\mathbb{N}, S_1, S_2) \models \mathsf{F}^\ulcorner\varphi^\urcorner$

   if and only if $\#\varphi \in S_2$

   if and only if $\#\varphi \in \Gamma^-(S_1, S_2)$ (by the assumption that $\Gamma(S_1, S_2) = (S_1, S_2)$)

   if and only if $(\mathbb{N}, S_1, S_2) \models \overline{\varphi}$ (by the definition of Γ^-.)

Therefore, we conclude that the structures $(\mathbb{N}, S_1, S_2)$ verify the local disquotation axioms of **TFB** for all positive sentences. Moreover, as these structures $(\mathbb{N}, S_1, S_2)$ are standard, they also satisfy the axioms schemata of **UTFB**

$$\forall s_1 ... \forall s_n (\mathsf{T}^\ulcorner\varphi s_1, ..., s_n{}^\urcorner \leftrightarrow (\varphi(s_1, ..., s_n))$$

$$\forall s_1 ... \forall s_n (\mathsf{F}^{\ulcorner} \varphi s_1, ..., s_n {}^{\urcorner} \leftrightarrow (\overline{\varphi(s_1, ..., s_n)}))$$

for all positive positive formulas $\varphi(x_1, ..., x_n)$. □

Finally, we show that these $\mathcal{L}^{+}_{\mathsf{TF}}$-structures are also models of $\mathbf{KF_{pos}}$.

Proposition 4. [17] *Assume that* S_1, $S_2 \subseteq \omega$. *Then the following are equivalent:*

1. $(\mathbb{N}, S_1, S_2)$ *is a model of* **UTFB**
2. $\Gamma(S_1, S_2) = (S_1, S_2)$
3. $(\mathbb{N}, S_1, S_2)$ *is a model of* $\mathbf{KF_{pos}}$

Sketch. The equivalence between 1. and 2. is given by Proposition 3. Moreover, 3. implies 1. by the fact that **UTFB** is a sub-theory of $\mathbf{KF_{pos}}$.[18] It remains to be shown that 2. implies 3. To argue for this one follows the strategy adopted for proposition 3. We assume that $\Gamma(S_1, S_2) = (S_1, S_2)$ and reason about some positive φ. Axioms KF1, KF2, KF11, KF12 are instances of the biconditionals of **UTFB**, so we don't need to consider them. For axioms KF3 – KF6, we argue informally that these axioms are, when considered schematically, instances of the biconditionals of **UTFB** and therefore each instance of them is satisfied by the $\mathcal{L}^{+}_{\mathsf{TF}}$-structures $(\mathbb{N}, S_1, S_2)$ given the equivalence with 1. The quantified versions of the axioms are satisfied by induction on the complexity of φ, ψ. For axioms KF7 – KF10, we take KF7 as an example: for some positive $\dot{\forall} v\varphi$ we assume that $(\mathbb{N}, S_1, S_2) \models \mathsf{T}^{\ulcorner} \dot{\forall} v\varphi {}^{\urcorner}$ and see that we have the following equivalences: $(\mathbb{N}, S_1, S_2) \models \mathsf{T}^{\ulcorner} \dot{\forall} v\varphi {}^{\urcorner}$, if and only if $(\mathbb{N}, S_1, S_2) \models \forall x \varphi(x)$ (because $\dot{\forall} v\varphi$ is a positive sentence and we have the

[17] This is (Leigh, 2016, proposition 1.2).

[18] This is (Leigh, 2016, Lemma 5.2). A similar result has been shown by Cantini (1989, Lemma 3.2 (ii)), for the versions of disquotational and compositional truth without the falsity predicate.

biconditionals of **UTFB**), if and only if for all n, $(\mathbb{N}, S_1, S_2) \models \varphi(n)$, if and only if $(\mathbb{N}, S_1, S_2) \models \mathsf{T}^{\ulcorner}\varphi n^{\urcorner}$. One reason analogously about the axioms KF8 – KF10. □

The following section reasons about $\mathbf{KF_{pos}}$ and shows that $\mathbf{KF_{pos}}$ is consistent and internally consistent with the rules of *Necessitation* and *Conecessitation* for the truth predicate. Moreover, it shows that $\mathbf{KF^*_{pos}}$, i.e., the theory resulting by closing $\mathbf{KF_{pos}}$ under *Necessitation* and *Conecessitation* for the truth predicate, is consistent and internally consistent with global reflection.

5.3.2 Internal Consistency with Global Reflection

This section aims to show that $\mathbf{KF_{pos}}$ – and therefore also **TFB** and **UTFB** – is consistent and internally consistent with global reflection. To do so we extend $\mathbf{KF_{pos}}$ to the theory $\mathbf{KF^*_{pos}}$, by closing $\mathbf{KF_{pos}}$ under the following rules for the truth predicate T:[19]

$$\frac{\varphi}{\mathsf{T}^{\ulcorner}\varphi^{\urcorner}}\ \text{NEC}; \quad \frac{\mathsf{T}^{\ulcorner}\varphi^{\urcorner}}{\varphi}\ \text{CONEC}$$

One might ask why we do not investigate analogous rules for the falsity predicate F. We avoid doing so because our aim is to show that the $\mathbf{KF_{pos}}$ is consistent and internally consistent with global reflection, and the closure under NEC and CONEC is here only technically useful, and investigate rules for the falsity predicate is not necessary for this purpose. Informally, this section aims to show that $\mathbf{KF^*_{pos}}$ has standard models, and that $\mathbf{KF^*_{pos}}$ is internally consistent.[20] We will prove the following:

[19] These rules are not allowed in proofs from premises. They should be understood as closure conditions on theories. Looking at NEC, for instance, one understands the rule in the following manner: if $\mathbf{KF^*_{pos}}$ proves φ, then it also proves $\mathsf{T}^{\ulcorner}\varphi^{\urcorner}$.

[20] This section follows Halbach's idea of the proof in (Halbach, 2014, Theorem 19.21, p. 271), that closing the theory of positive truth **PUTB** under NEC and CONEC results in a consistent theory.

Proposition 5. *There are standard models of* $\mathbf{KF}^{\mathbf{nec}}_{\mathbf{pos}}$*, where* $\mathbf{KF}^{\mathbf{nec}}_{\mathbf{pos}}$ *is* $\mathbf{KF}_{\mathbf{pos}}$ *together with* NEC*.*

Proposition 6. *Any application of* CONEC *in* $\mathbf{KF}^{*}_{\mathbf{pos}}$ *is admissible in the theory. In other words,* $\mathbf{KF}^{*}_{\mathbf{pos}}$ *proves the same theorems as* $\mathbf{KF}^{\mathbf{nec}}_{\mathbf{pos}}$*.*

Proof of Proposition 5. [21] We want to construct a standard model $\mathfrak{M}^*$ of $\mathbf{KF}^{\mathbf{nec}}_{\mathbf{pos}}$. $\mathfrak{M}^*$ is supposed to be the model obtained by closing $(\mathbb{N}, S_1, S_2)$ under Γ, starting with $S_2 = \emptyset$ and with S_1 being the following set A of (codes of) non-positive statements. We define A as the set of codes of non-positive sentences provable in $\mathbf{KF}^{\mathbf{nec}}_{\mathbf{pos}}$:

$$A := \{\#\varphi | \varphi \notin \mathsf{sent}^{+}_{\mathsf{TF}} \text{ and } \mathbf{KF}^{\mathbf{nec}}_{\mathbf{pos}} \vdash \varphi\}$$

Proposition 2 shows that closing $(\mathbb{N}, A, S_2)$ under Γ reaches fixed points. We observe from Propositions 3 and 4 that $\mathfrak{M}^* \models \mathbf{KF}_{\mathbf{pos}}$. We show that NEC is valid in $\mathfrak{M}^*$. We only focus on applications of NEC to φ that are not positive, because for any φ in $\mathsf{sent}^{+}_{\mathsf{TF}}$, NEC is derivable from the biconditionals for the truth predicate.[22] We reason by standard induction on the number of applications of NEC. One reasons about some derivation in $\mathbf{KF}^{\mathbf{nec}}_{\mathbf{pos}}$ and lets some application of NEC to a non-positive sentence be given. We focus on a sub-proof Q of this derivation, such that Q ends with an application of NEC

$$\frac{\varphi}{\mathsf{T}^{\ulcorner}\varphi^{\urcorner}}.$$

If the above application is the first application of NEC, then we can conclude that everything up to and including φ is provable in $\mathbf{KF}_{\mathbf{pos}}$. By the fact that $\mathfrak{M}^*$ is a model of $\mathbf{KF}_{\mathbf{pos}}$ we have that $\mathfrak{M}^* \models \varphi$. Moreover, since we assumed that φ is

[21] This is the proof in (Zicchetti, 2022a).

[22] This employs the fact, mentioned earlier in the sketch of Proposition 4 that **UTFB** is a sub-theory of $\mathbf{KF}_{\mathbf{pos}}$.

not positive and provable in $\mathbf{KF}^{\mathbf{nec}}_{\mathbf{pos}}$,[23] we conclude by definition of A that the code of φ is in A. By the fact the pair (A, S_2) is a fixed point of Γ we conclude that $\mathfrak{M}^* \models \mathsf{T}^{\ulcorner}\varphi^{\urcorner}$. Now, assume that in $\mathbf{KF}^{\mathbf{nec}}_{\mathbf{pos}}$, n applications of NEC are satisfied in $\mathfrak{M}^*$. We reason about some sub-derivation Q' ending with the $n+1$ application of NEC:

$$\frac{\varphi}{\mathsf{T}^{\ulcorner}\varphi^{\urcorner}}$$

By our induction hypothesis, $\mathfrak{M}^* \models \varphi$. Moreover, by the fact that φ is non-positive by assumption and provable in $\mathbf{KF}^{\mathbf{nec}}_{\mathbf{pos}}$, we reason analogously to the case of the first application of NEC and argue that $\mathfrak{M}^* \models \mathsf{T}^{\ulcorner}\varphi^{\urcorner}$. Therefore, we conclude that $\mathfrak{M}^*$ satisfies NEC. □

Proof of Proposition 6. [24] We want to show that any application of CONEC is admissible, i.e., that for any φ proved in $\mathbf{KF}^{*}_{\mathbf{pos}}$ with any number of applications of CONEC, φ is also provable in $\mathbf{KF}^{\mathbf{nec}}_{\mathbf{pos}}$. We do so by induction of the number of applications of CONEC. We reason about some arbitrary derivation R in $\mathbf{KF}^{*}_{\mathbf{pos}}$, and let some applications of CONEC be given. We focus on applications of CONEC to non-positive sentences, since CONEC is derivable from the biconditionals for the truth predicate for positive sentences. We focus on some sub-derivation P of R in $\mathbf{KF}^{*}_{\mathbf{pos}}$ ending with an application of CONEC

$$\frac{\mathsf{T}^{\ulcorner}\varphi^{\urcorner}}{\varphi}$$

If the above application is the first application of CONEC in R, then we conclude that everything up to and including $\mathsf{T}^{\ulcorner}\varphi^{\urcorner}$ is provable in $\mathbf{KF}^{\mathbf{nec}}_{\mathbf{pos}}$. Therefore, by Proposition 5 $\mathfrak{M}^* \models \mathsf{T}^{\ulcorner}\varphi^{\urcorner}$. From this we conclude that the code of φ is in $\Gamma(A, S_2)$,

[23] Trivially, since φ is provable in $\mathbf{KF}_{\mathbf{pos}}$ and $\mathbf{KF}^{\mathbf{nec}}_{\mathbf{pos}}$ extends $\mathbf{KF}_{\mathbf{pos}}$.

[24] This is the proof in (Zicchetti, 2022a).

and by the fact that φ is not positive by assumption we conclude that the code of φ has to be in A. By the definition of A we conclude that $\mathbf{KF}^{\mathbf{nec}}_{\mathbf{pos}} \vdash \varphi$. Therefore, there is a derivation in $\mathbf{KF}^{\mathbf{nec}}_{\mathbf{pos}}$ such that φ is provable without the application of CONEC in the sub-derivation P of R. Now, we assume that in $\mathbf{KF}^{*}_{\mathbf{pos}}$, n applications of CONEC are admissible. We reason about some sub-derivation of P' ending with the $n+1$ application of CONEC:

$$\frac{\mathsf{T}^{\ulcorner}\varphi^{\urcorner}}{\varphi}$$

By induction hypothesis $\mathsf{T}^{\ulcorner}\varphi^{\urcorner}$ is provable in $\mathbf{KF}^{\mathbf{nec}}_{\mathbf{pos}}$, and by Proposition 5 we have that $\mathfrak{M}^{*} \models \mathsf{T}^{\ulcorner}\varphi^{\urcorner}$. Therefore, the code of φ is in $\Gamma(A, S_2)$ and by the assumption that φ is not positive, we conclude that the code of φ has to be in A. By the definition of A we conclude that φ is provable in $\mathbf{KF}^{\mathbf{nec}}_{\mathbf{pos}}$, so the $n+1$ application of CONEC is also admissible. □

We observe from Propositions 5 and 6 that we have standard models of $\mathbf{KF}^{\mathbf{nec}}_{\mathbf{pos}}$, which are also models of $\mathbf{KF}^{*}_{\mathbf{pos}}$. Moreover, $\mathbf{KF}^{*}_{\mathbf{pos}}$ is consistent and internally consistent; it is consistent because it has models. The internal consistency follows from the fact that $\mathbf{KF}^{*}_{\mathbf{pos}}$ is closed under CONEC; if $\mathbf{KF}^{*}_{\mathbf{pos}}$ were to be internally inconsistent, then we would have a sentence φ, such that $\mathbf{KF}^{*}_{\mathbf{pos}}$ proves $\mathsf{T}^{\ulcorner}\varphi \wedge \neg\varphi^{\urcorner}$. The closure under CONEC would imply that $\mathbf{KF}^{*}_{\mathbf{pos}}$ proves $\varphi \wedge \neg\varphi$. However, this contradicts the fact that $\mathbf{KF}^{*}_{\mathbf{pos}}$ has models. From Propositions 5 and 6 it is straightforward to show that $\mathbf{KF}^{*}_{\mathbf{pos}}$ is consistent and internally consistent with global reflection. We formulate global reflection unrestrictedly:

$$\forall x(\mathsf{sent}_{\mathsf{TF}}(x) \wedge \mathsf{prov}_{\mathsf{KF}^{*}_{\mathsf{pos}}}(\mathsf{x}) \rightarrow \mathsf{T}(x)) \qquad (\mathsf{GRP}_{\mathsf{KF}^{*}_{\mathsf{pos}}})$$

Theorem 13. $\mathfrak{M}^* \models \mathbf{GRP}[\mathbf{KF}^*_{\mathbf{pos}}]$.[25]

Proof. We reason about $\mathfrak{M}^*$ and take some sentence φ, such that $\mathfrak{M}^* \models \mathsf{prov}_{\mathsf{KF}^*_{\mathsf{pos}}}\ulcorner\varphi\urcorner$ and reason in the following manner: since $\mathsf{prov}_{\mathsf{KF}^*_{\mathsf{pos}}}\ulcorner\varphi\urcorner$ is true in $\mathfrak{M}^*$ and is an arithmetical sentence, $\mathsf{prov}_{\mathsf{KF}^*_{\mathsf{pos}}}\ulcorner\varphi\urcorner$ is true in $\mathbb{N}$. This shows – by the meaning of the provability predicate – that $\mathbf{KF}^*_{\mathbf{pos}} \vdash \varphi$. By the fact that $\mathbf{KF}^*_{\mathbf{pos}}$ is closed under NEC we conclude that $\mathbf{KF}^*_{\mathbf{pos}} \vdash \mathsf{T}\ulcorner\varphi\urcorner$. By the fact that NEC is satisfied in $\mathfrak{M}^*$ we have that $\mathfrak{M}^* \models \mathsf{T}\ulcorner\varphi\urcorner$. That is, $\mathfrak{M}^*$ is a model of $\mathbf{GRP}[\mathbf{KF}^*_{\mathbf{pos}}]$. □

Moreover, we have the following Corollary:

Corollary 1. $\mathbf{GRP}[\mathbf{KF}^*_{\mathbf{pos}}]$ *is consistent and internally consistent.*

5.4 Philosophical Discussion

This section discusses the philosophical and epistemological ramifications of the previous section's results for discussing *trustworthiness*. This section employs some of the notions introduced in chapter 1. To avoid confusion and enhance clarity we repeat the relevant epistemological notions. The remainder of the section considers some worries and issues for the proponents of theories of positive truth and aims to provide preliminary responses to them.

5.4.1 Projects and Cornerstones

Agents constantly engage in cognitive inquiries. Using the terminology introduced by Wright (2004), they engage in cognitive projects. Informally, a cognitive project consists of a (collection of) question(s) and a (collection of) procedure(s) agents

[25] This is the proof in (Zicchetti, 2022a).

might competently execute to answer the project's question. As pointed out in chapter 1, agents engage in cognitive projects in the empirical sciences, philosophy, and mathematics. Projects have *cornerstones*: propositions essential for the significance and integrity of the inquiry. Consider some cognitive project aiming to make some cognitive achievement about some arithmetical subject matter and suppose this project employs some theory **S** to prove theorems about this subject matter. As pointed out also in chapter 1, a cornerstone of this project is that **S** is non-trivial; if **S** is trivial, then **S** is not a reliable source of evidence to believe propositions about the subject matter or to claim cognitive achievement about the subject matter.[26] Following Wright's characterisation of cornerstones, we have that the soundness of **S** and the consistency of the concepts employed by **S** are also cornerstones.

5.4.2 Trustworthiness of Truth

Fischer et al. (2019) distinguish between two concepts of truth in the context of cognitive projects: a concept of *scientific* truth and a concept of *full disquotational* truth.[27] Fischer et al. (2019) describe scientific truth as a theoretical concept, employed in scientific theories to *explain non-semantic facts.* They argue that the scientific concept of truth is not different from other scientific, theoretical concepts employed in science. A fundamental characteristic of the scientific concept of truth is that the logic of truth should inherit the logic of the non-semantic language. In the context of the previous sections, where the non-semantic theory is first-order **PA** formulated in classical logic, a theory of scientific truth should also be formulated in

[26]This is essentially what has been said in chapter 1. See also (Pedersen, 2021).

[27]The authors remain open about whether this distinction amounts to the distinction by Field (994a) between *inflationary* and *deflationary* truth, or to the distinction by McGee (005b) between *disquotational* and *causally explanatory* - other times called *correspondence* - truth. To investigate this issue would exceed the scope of this chapter.

classical logic. On the other hand, disquotational truth only intends to be a device of quotation and disquotation. This informal concept follows the intuition that it should be unproblematic to assert that if some state of affairs is so and so, then it is true that some state of affairs is so and so, and *vice versa*.[28] Fischer et al. (2019) argue that, since the full disquotational concept of truth intends to be a device of naive, i.e. unrestricted, quotation and disquotation, this concept should be governed by some non-classical logic (to avoid triviality). Fischer et al. (2019) identify $\mathbf{TS_0}$ and $\mathbf{UTS_0}$ – presented in section 5.2.2 – as theories of full disquotational truth.

It has been pointed out in the previous paragraph that the soundness of the relevant theory **S** employed in the project, and the consistency of **S**'s concepts are cornerstones. In the context of cognitive projects employing a theory of truth, this amounts in the following: the soundness of the theory of truth and the consistency of the concept of truth employed are cornerstones. *Trustworthiness* is an adequacy condition on theories of truth arising from the following reflection on the importance of the cornerstones mentioned above: if the soundness of **S** is made explicit by the addition of $\mathsf{GRP_S}$ to **S**, then **S**'s concept of truth should remain consistent. Conversely, if **S** is either inconsistent or internally inconsistent with $\mathsf{GRP_S}$, then **S** is *untrustworthy*. To put this succinctly, the slogan could be that truth must be trustworthy: any concept of truth theory employed in cognitive projects should remain consistent if the soundness of the relevant theory of truth employed is made explicit. The desirability of trustworthiness – or some form of coherence – of truth

[28]This is the informal intuition that Tarski has at the beginning of (Tarski, 1936), which goes back to Aristotle.

has been already pointed out by Horsten and Halbach (2015) and Leitgeb (2007).[29]

Interpreting the results of section 5.2.2, we have that the theories $\mathbf{TS_0}$ and $\mathbf{UTS_0}$ investigated in (Fischer et al., 2017) are *trustworthy*. This is supported by Proposition 1: these theories of full disquotational truth in non-classical logic are (internally) consistent with global reflection. On the other hand, many theories of scientific truth in classical logic are not trustworthy. Fischer et al. (2019) claim the following:

> Theories of [scientific] truth do not sit well with statements of their own soundness. [...] Scientific notions of truth, are inadequate if such a requirement is adopted. [...] In theories of classical truth we cannot consistently hold that what they prove is true, and not false. This entails that scientific theories of truth suffer the same fate, by our assumption that only theories of classical truth can be considered theories of scientific truth. (Fischer et al., 2019, pp. 7 - 8)

As we saw in section 5.2.2, **KF** is internally inconsistent with global reflection and therefore not trustworthy by the standards set by Fischer et al. (2019). There are other theories of scientific truth that are not trustworthy by the same standards: **FS** is an example of such theories.[30] It is well-known that **FS** is inconsistent with global reflection:

Proposition 7. (Horsten et al., 2012, Proposition 4.6) $\mathbf{GRP}[\mathbf{FS}]$ *is inconsistent.*[31]

[29] Fischer et al. (2019) formulate the soundness of the relevant **S** with global reflection. In doing so, they follow the idea already expressed by Kreisel and Lévy (1968). We agree with Fischer et al. (2019) and Kreisel and Lévy (1968) that global reflection is the most natural, intended formulation of soundness.

[30] See (Halbach, 2014) and (Friedman and Sheard, 1987) for two presentations of **FS**. Another example is the theory **VF**. For a presentation of **VF** see for instance (Cantini, 1990).

[31] This is so because **FS** is ω-inconsistent. A (recursively axiomatisable) theory **S** is ω-inconsistent just in case there is a φ, such that $\mathbf{S} \vdash \varphi(n)$ for all n and $\mathbf{S} \vdash \neg\forall x\varphi(x)$.

For clarity, let us make the following observation explicitly:

Observation 1. $\mathbf{TS_0}$ *and* $\mathbf{UTS_0}$ *are trustworthy, whereas* **KF** *and* **FS** *are not.*

The results of section 5.3 support the following observation concerning theories of positive truth (and falsity) in classical logic:

Observation 2. *The theories* **TFB**, **UTFB** *and* $\mathbf{KF^*_{pos}}$ *are trustworthy.*

The result that theories of positive truth and falsity are consistent and internally consistent with global reflection supports this latter observation. The following section makes some remarks on the value of the trustworthiness of truth in the relation to epistemic norms of cognitive projects. The following section overlaps with parts of chapter 1.

5.4.3 Norms and Trustworthiness

Cognitive projects are *epistemic practices*: they have aims and goals, which are pursued by the agents engaging in them. Practices have norms, where norms can be informally seen as 'rules' that regulate the practice (to some degree). Epistemic norms can be informally understood as 'rules' regulating epistemic dimensions of the practice:[32] Epistemic norms have normative force. In an epistemic practice X with epistemic norms $R_1, ...R_n$, agents engaging with X *ought to* follow the epistemic norms. Moreover, as pointed out in chapter 1, epistemic norms describe when it is epistemically permissible (resp not epistemically permissible) to hold various epistemic attitudes. As pointed out in chapter 1, the following norm about agents' beliefs seems quite intuitive (No Dogmatism):

[32]This follows my presentation in chapter 1.

(ND) Agents should refrain from believing p in the presence of compelling evidence against p.

Using the terminology of chapter 1, (ND) can be understood as claiming that agents should be responsive to *overriding* defeaters, where an overriding defeater is a compelling counter-evidence against a target proposition. It seems intuitive to judge agents as blameworthy if they are not responsive to defeaters for not conforming to the epistemic norm (ND). As we saw in chapter 1, this notion of blameworthiness follows Brown (2018, p. 389) and Boult (2021b).[33] When thinking about the (ND) norm, it is intuitive to think that "we might judge a subject blameworthy for dogmatically continuing to believe a claim even after receiving evidence which undermines it." (Brown, 2018, p. 389) In chapter 1, we also considered the following norm (Epistemic Responsibility):

(ER) Agents should be in a position to rationally claim to be warranted to believe propositions, for which they have evidential warrant.

This norm sees also quite intuitive. Agents engaging in projects aiming to make cognitive achievement about some subject matter should be in a position to rationally claim to be warranted in their beliefs. Moreover, they should be in a position to rationally claim their cognitive achievements.

In the light of the proposed understanding of blame and epistemic norms, we can reflect on Observations 1 and 2 about trustworthiness. Proponents of untrustworthy theories, such as **KF** or **FS**, can be evaluated as epistemically blameworthy for their

[33] As pointed out in chapter 1, one might worry that there is no genuine epistemic kind of blame. Such worries might be motivated by the idea that blame is a *moral* concept ((Kauppinen, 2018)). This dissertation assumes that the epistemic account of blame is coherent enough.

commitments to those theories because their commitment to untrustworthy theories *prima facie* violates (ND): there is compelling evidence that such theories are untrustworthy. Indeed, both **KF** and **FS** are (internally) inconsistent with global reflection. Moreover, commitment to such theories seems to violate (ER). Since there is compelling evidence that such theories are not trustworthy, agents committed to an untrustworthy theory **S** are hardly in a position to rationally claim to be warranted in their beliefs supported by **S** precisely because **S** is not trustworthy. Moreover, they do not seem to be in a position to rationally claim cognitive achievement supported by **S**. Therefore, from the perspective of these epistemic norms, trustworthy theories – in either classical or non-classical logic – seem preferable. When trustworthiness is considered, it seems that theories of positive truth are on a par with theories of full disquotational truth. The following section considers some significant worries against the proponent of positive truth.

5.4.4 Positive Truth and Cognitive Projects: the Worries[34]

As we saw in section 5.2.2, **GRP**[**KF**] proves that the liar is both true and false. The reason for this result is that global reflection provides a bridge between the *external* and the *internal* logic of **KF**, where the former is the logic outside the scope of T and the latter is the logic inside the scope of T. **KF**'s external logic is classical, whereas its internal logic is non-classical. Global reflection is problematic because it pushes **KF**'s classical external negation inside the scope of **KF**'s non-classical truth predicate.

[34]The same worries are discussed in (Zicchetti, 2022a).

In contrast, the proof of internal inconsistency is blocked in the case of $\mathbf{GRP}[\mathbf{KF}^*_{\mathbf{pos}}]$; the proof of the internal inconsistency in $\mathbf{GRP}[\mathbf{KF}]$ employs *unrestricted* compositional principles. Compositionality together with KF11 and KF12 are enough to derive the internal inconsistency. However, $\mathbf{GRP}[\mathbf{KF}^*_{\mathbf{pos}}]$ has compositional principles for strictly positive statements only. In $\mathbf{KF}^*_{\mathbf{pos}}$ the external negation does not interact with the internal negation – the falsity predicate –, and this is essential to block the proof of the internal inconsistency. However, Nicolai (2021, p. 736) pointed out one can define a translation $(*)$ from the language $\mathcal{L}_{\mathsf{T}}$ into the strictly positive language $\mathcal{L}^+$ that essentially replaces negative occurrences of truth with the falsity predicate F of $\mathcal{L}^+$.[35] If the proponent of positive truth were to accept the translation function $(*) : \mathcal{L}_{\mathsf{T}} \to \mathcal{L}^+$, then via global reflection the internal inconsistency would arise again.[36] This result seems to threaten the philosophical importance of the internal consistency of $\mathbf{GRP}[\mathbf{KF}^*_{\mathbf{pos}}]$: one can argue that – via the translation $(*)$ – the proponent of positive truth finds herself in the same position as the proponent of theories such as **KF**. This worry is pressing because via the translation $(*)$ positive truth would also result in an untrustworthy theory. As a result, the proponent of positive truth would be epistemically blameworthy for her commitments. Let us put this worry explicitly:

(translation) Can the proponent of positive truth have a warrant for her acceptance of **S** and nevertheless be warranted in rejecting the translation $(*)$?[37]

Additionally, one might express an even deeper worry concerning the question of warrant to accept theories of positive truth to start with:

[35]The details of the translation are not important for our purposes. What is crucial is that, employing the translation, F is understood as $\neg\mathsf{T}$. The details of the translation can be found in Nicolai (Nicolai, 2021, p. 751).

[36]This is the proof in (Nicolai, 2021, Proposition 1).

[37]This is the worry presented in Zicchetti (2022a).

(warrant) Is there any warrant to accept theories of positive truth to start with? If there is such a warrant, what is its force?[38]

The issue of (warrant) is not new. Indeed, it addresses the well-known problem of providing a principled argument for theories of positive truth. Theories of positive truth are usually conceived as a response to the paradoxes, insofar as the restriction to positive biconditionals straightforwardly retains consistency without loss of generality and proof-theoretic strength for the arithmetical language.[39] However, such restriction is taken to be artificial. Halbach (2009), Horsten and Leigh (2017) and Cieśliński (2017b, 2015) independently argue that the theories of positive truth and falsity are well-motivated, via a careful analysis and diagnosis of the paradoxes of truth. The intuition behind this idea is that, by analysing the paradoxes of truth, one formulates the hypothesis that paradoxes necessarily involve occurrences of the truth predicate that are not strictly positive.[40] However, the force of the argument for positive truth needs to be spelled out: one might (and should!) ask how good the warrant stemming from the analysis of the paradoxes is. To my best knowledge, the only place where the question about the warrant's goodness is explicitly discussed by Cieśliński (2015, Section 5.5.):

> Restoration of the consistency of disquotational theory is a natural aim. Naive, unrestricted T-schema generates a contradiction – that's a fact to which all truth theorists must react and the disquotationalist is no exception. Restoring the consistency of a theory of truth should be

[38] This is the worry presented in Zicchetti (2022a).

[39] This is well-known. See (Halbach, 2014, Corollary 19.18) or (Cieśliński, 2015, Theorem 11).

[40] Curry's paradox fits into this hypothesis, if implication is not taken as a primitive: if '$\rightarrow$' is taken to be defined as usually, then Curry's paradox also involves a negative occurrence of the truth predicate. See for instance Cieśliński (2017b, pp. 53 – 54) for a discussion of this issue.

treated as a permissible motivation for the disquotationalist to proceed. The question is only how far it can take us.

Although it seems acceptable to motivate positive truth to restore consistency, the question concerning the force of such motivation is still open. How far can this take us? A way of addressing this question involves explaining the relations and dependencies between the respective answers to (warrant) and (translation). Finally, a third worry involves the informal claim that positive truth is a scientific concept of truth. This worry is related to the issues that theories of positive truth might be too restrictive and thereby inadequate for scientific cognitive projects. To address this worry, we can state the following question explicitly:

(project) What cognitive project can a theory of positive truth be associated with, so that positive truth embodies the concept of scientific truth?[41]

The remainder of this section aims to address these worries and to suggest possible responses to them.[42]

5.4.5 Responses

We start with (project). It is well-known that theories of positive truth are restrictive with respect to their truth-theoretic principles. On the other hand, the truth-theoretic principles of theories of full disquotational truth are fully unrestricted. And even the classical theory **KF** has fully general compositional principles. To successfully address (project) the proponent of positive truth must provide a cognitive project, where positive truth plays the theoretical role embodied by scientific truth. The proponent of positive truth seems to have a good, preliminary response

[41] This worry is presented in (Zicchetti, 2022a).

[42] However, a thorough defence of each response would exceed the scope of this chapter.

to (project): positive truth is employed in cognitive projects as a tool, or a device, to investigate mathematical, non-semantic subject matters. When considering inquiries into purely non-semantic subject matters, positive truth is general enough: $\mathbf{KF}^*_{\mathbf{pos}}$ is proof-theoretically equivalent to its unrestricted formulation, **KF**, with respect to the language without the truth (and falsity) predicate. Within these cognitive projects – so the proponent of positive truth can argue – positive truth is as general as the scientific truth predicate of **KF**.

However, one should recognise that positive truth can hardly be a scientific concept of truth in cognitive projects aiming to investigate *semantic facts involving truth*. When investigating some fully general notion of truth, the choice of positive truth needs an independent motivation. Additionally, from a model theoretic perspective positive truth is not as general as the non-classical theories: theories such as $\mathbf{UTS_0}$ do not exclude natural models of truth. On the other hand, theories such as $\mathbf{KF}^*_{\mathbf{pos}}$ do exclude natural models, such as the minimal model.[43]

Concerning (warrant), we can see that, *if* one focuses on the proposed cognitive projects, where positive truth is a theoretical tool to investigate purely mathematical subject matters, then the proponent of positive truth should be able to employ Cieśliński's argument from the analysis of the paradoxes to motivate her choice of positive truth: the restriction of the unrestricted biconditionals to some subsets thereof is motivated by the argument from the paradoxes, and this seems to be enough to warrant the *instrumental* acceptance of positive truth. After all, positive

[43] Although this issue somewhat limits the generality of positive truth, it does not immediately threaten the theoretical role played by positive truth to investigate purely mathematical subject matters. After all, for this purpose the theory of truth must allow for standard models.

truth is just a useful tool. Cieśliński's reasoning suggests that the careful analysis of the paradoxes provides the agent with a warrant to accept *some* restrictions of the T-biconditionals. However, the proponent of positive truth still has to motivate positive truth explicitly.[44] Fortunately, the proponent of positive truth does not need a philosophical motivation because for the target cognitive projects a warrant for instrumental acceptance in sufficient, and for such a warrant simple pragmatic considerations about the virtues of positive truth for the success of the project should be enough. Concerning (translation): given the instrumental acceptance of positive truth, the proponent of positive truth in such purely mathematical cognitive projects has good reasons to reject the translation $(*)$ given by pragmatic considerations; the translation would bring the inconsistency back, threatening the success of the cognitive enquiry.

5.4.6 Conclusion

Let us clarify: these preliminary responses seem to provide a good initial defence of positive truth. However, it should be added that by focusing on instrumental acceptance the proponent of positive truth may have too easy answers to (translation) and (warrant): the proponent of positive truth can respond to these challenges 'only' via pragmatic considerations. In contrast, the proponent of full disquotational truth seems to have a *philosophical* argument for her warrant to accept full disquotational truth: the truth predicate embodies, or captures, some informal concept of full disquotational truth. The challenge to provide a philosophical argument for the choice of positive truth and falsity is still open. This is a possible direction to investigate: she could aim to understand the concept of truth as expressing some epistemic

[44]This is so because Cieśliński's reasoning would also motivate the choice of some typed notion of truth, with no need to opt for a type-free positive truth predicate.

notions similar to *warranted assertibility*.[45] Alternatively, one could understand truth as the stronger notion of *super-assertibility*.[46] Under these interpretations, the proponent of positive truth (and falsity) should have good *philosophical reasons* to reject the translation (*) and would therefore have a philosophical answer to (translation).[47] However, one would have to explain how the notion of warranted assertibility (resp. super-assertibility) motivates the choice of the positive biconditionals. Moreover, the proponent of positive truth would also have to address the usual objections against the thesis that truth can be understood as an epistemic concept.[48] After that, the proponent of positive truth would have to *at least* assess whether this philosophical argument for positive truth provides good answers to (warrant), (translation) *and* (project).

[45] This has been investigated for instance by Kvanvig (1999) and Tennant (1995).

[46] This has been proposed by Wright (1996).

[47] The argument against $(*)$ follows from the so-called problem of neutral states of information, discussed for instance by Kvanvig (1999).

[48] To provide such a fully fleshed-out philosophical interpretation of positive truth and falsity in terms of assertibility is left open for another occasion.

Part III

Internal Categoricity and Determinacy

Summary of Part III

This dissertation's third and final part focuses on so-called categoricity statements. As pointed out in the introduction of this dissertation, the property of categoricity played an essential role in discriminating between two types of mathematical structures: *particular* and *general* structures (Isaacson, 2011, p. 18). Chapter 6 investigates so-called Parsons-style, *internal categoricity*, in the context of *first-order* theories. Chapter 6 is a revised version of (Fischer and Zicchetti, 2022).

Chapter 6

This chapter investigates Parsons-style categoricity theorems, called 'internal', in contrast to Dedekind-style, 'external' categoricity. This chapter introduces internal categoricity theorems and focuses on the following issue: it has been argued in (Button and Walsh, 2016) and (Button and Walsh, 2018) that such theorems are inadequate *qua* internal categoricity theorems because they are not general enough. The main reason for this loss of generality is the choice of first-order logic instead of second-order resources. Chapter 6 focuses on providing a Parsons-style categoricity theorem that is (as I will argue) general enough and therefore good enough *qua* internal categoricity theorem. Chapter 6 provides a *truth-theoretic*, Parsons-style categoricity theorem employing a primitive, axiomatic notion of truth. This work relates to and expands on the work done by Button and Walsh (2016, 2018), Mount and Waxman (2021), Simpson and Yokoyama (2013) and Feferman and Hellman (1995).

The second part of chapter 6 focuses on whether internal categoricity theorems provide *determinacy* of truth. We will argue that internal categoricity theorems

do provide a form of *internal determinacy* of truth. This is connected to work done by Field (2001), Button and Walsh (2016, 2018), Button (2022), Väänänen (2012); Väänänen (2020), Väänänen and Wang (2015), Hamkins and Yang (2013) and Maddy and Väänänen (2022). This internal determinacy is going to be spelled out in terms of so-called *intolerance theorems*. Finally, we will show how this internal notion of determinacy does not contradict the result by Hamkins and Yang (2013) that (to put it briefly) *satisfaction is not absolute*.

As pointed out in the introduction of this dissertation, chapter 6 focuses on the question of internal categoricity of arithmetical theories.[49]

Epistemological Issues

Whilst this dissertation will not discuss this further, there is a significant connection to be mentioned between categoricity and epistemological issues. Parsons (1990, 2008) investigates the epistemological issue of determining whether internal categoricity theorems imply or support *agreement* between agents accepting *prima facie* different arithmetical theories. Parsons argues that internal categoricity theorems *force* agreement between agents, so long as each agent can perform, or go through, the proof of the internal categoricity theorem. Moreover, since internal categoricity implies – the theorem to be called later – *intolerance*, which is an internalised version of elementary equivalence, the two agents will agree on the truth value of all sentences of the pure arithmetical vocabulary.[50] The claim that agents must agree

[49]For an extensive discussion of internal categoricity in the case of set theory see for instance (Button and Walsh, 2016). Investigating this topic would exceed the scope of this dissertation.

[50]This has been discussed and acknowledged in many places. See for instance (Walsh and Ebels-Duggan, 2015) and (Fischer, 2021).

that their respective theories are in some sense equivalent is the so-called Parsons' (Equivalence Claim):

(EC) Any two agents accepting schematic arithmetic *must* regard each other's theory as equivalent.[51]

With (EC) Parsons aims to argue that any two agents accepting arithmetic are obligated to accept that their theories are in some sense equivalent *because they are internally categorical*.

Although the issue of determining whether (EC) is correct is significant, we will not investigate it in this dissertation. However, the work provided in chapter 6 is essential to evaluate the status of (EC) for at least two reasons: first, the acceptable versions of internal categoricity provided in chapter 6 are needed for (EC) to be general enough in the case of first-order theories. The second reason is the following: the remarks and comments on the notion of *internal determinacy* of arithmetical truth implied by the internal categoricity theorems will be relevant to determine the *scope* of the possible meaning of the equivalence mentioned in (EC). Therefore, this work will be crucial to determine the extent of the (possible) agreement between agents.

[51] This thesis has been attributed to Parsons by Field (2001) with the name 'Intermediate Claim'.

Chapter 6

Internal Categoricity and Determinacy[1]

6.1 Introduction

The property of categoricity plays a significant role in both mathematics and philosophy of mathematics. Informally, categoricity allows us to distinguish between two types of theories: theories about a *particular* subject matter and theories about a somewhat *general* subject matter. As Isaacson (2011, p. 18) points out, mathematicians study two sorts of structures: *particular structures* and *general structures*. The distinction between them is informally marked by the natural language use of a definite article for theories of particular structures. One usually speaks of *the* natural numbers. When considering arithmetic, traditionally philosophers have strong

[1]This chapter is a revised version of the article "Truth, Categoricity and Determinateness" (Fischer and Zicchetti, 2022), which is currently under review. The original article is co-authored with Martin Fischer. The ideas in that article – and hence in chapter 6 – originated from the many discussions with Martin. Both he and I equally contributed to the ideas present in the article. Sections 6.1, 6.2 and 6.3 have been revised. However, the technical work presented in section 6.4 remains unchanged. Finally, section 6.5 has been minimally revised. For additional information about the origin of (Fischer and Zicchetti, 2022), see Publications.

intuitions about it being about a particular subject matter. This distinction is well-known, but sometimes acknowledged and explained in different terms. Shapiro (1997, pp. 40-41) considers the same distinction between theories about particular structures such as arithmetic, and theories about general structures such as group theory. In this terminology, the former are called *non-algebraic* theories. In contrast, the latter are called *algebraic*.[2] Additionally, we have *prima facie* good mathematical evidence to support the claim that arithmetic is a about a particular subject matter; Dedekind's categoricity theorem shows that second-order arithmetic is categorical in the sense that all full models of second-order arithmetic are isomorphic. In other words, second-order arithmetic pins down *a unique structure, up to isomorphism*. Moreover, one could claim that *arithmetical truth* is *determinate* at least in the following sense: that for any arithmetical statement φ, isomorphic models agree on the truth value of φ. This *elementary equivalence* is a consequence of Dedekind's categoricity.[3] For this investigation we follow the terminology adopted by Button and Walsh (2016, 2018) and call Dedekind's categoricity theorem *external*.

However, it has been argued that this external approach makes some problematic philosophical assumptions – to be discussed later. In order to circumvent these issues, some philosophers proposed an *internal* version of arithmetical categoricity. For the moment quite informally, this chapter focuses exactly on this internal approach to the categoricity of arithmetic. It does so by focusing on two aims: to generalise the Parsons-style version of internal categoricity for first-order arithmetic – to be presented and discussed in section 6.3 – and to investigate whether and

[2]Similar comments have been made also by many others. See for instance Grzegorczyk (1962), Kline (1982), Button and Walsh (2018).

[3]Here we do not focus on investigating this form of elementary equivalence. For more details about this, see Button and Walsh (2016, Corollary 2.5).

to what extent Parsons-style internal categoricity theorems for arithmetic provide determinacy of arithmetical truth. The chapter has the following structure: section 6.2 presents the relevant distinctions between external and internal approaches to categoricity. Moreover, it briefly discusses the issues with Dedekind's theorem and introduces the position called *internalism*. Section 6.3 surveys the relevant versions of internal categoricity found in the literature and discusses them mainly from the perspective of generality (in a sense to be made explicit later in the chapter) and determinacy. In particular, this section discusses some of the advantages and limitations of the different versions of internal categoricity. After that, section 6.4 provides a generalisation of the internal categoricity theorem for first-order arithmetic presented by Parsons (1990, 2008) and discussed also by Feferman (2013), Button and Walsh (2016) and Maddy and Väänänen (2022). This section aims to circumvent the expressive limitations of a first-order internal categoricity theorem for arithmetic by introducing a primitive truth predicate with an axiomatic theory of truth. Finally, section 6.5 draws some philosophical conclusions, supported by section 6.4, concerning the issue of whether internal categoricity provides some form of determinacy. We will argue that a notion of internal determinacy is obtained, as a form of *intolerance theorem*. However, neither *semantic* (or *external*) determinacy nor *standardness* is obtained.

A final remark is in order: this chapter does not want to adjudicate any debate between external and internal approaches. It takes some issues with the external approach at face value and investigates whether and to what extent internal approaches can circumvent some of the issues and still support our informal, initial

intuition of the uniqueness of arithmetical structures and determinacy of arithmetical truth.

6.2 Externalism and Internalism

Dedekind's categoricity theorem provides a piece of mathematical evidence for the categoricity of second-order arithmetic. It claims that all *full models* of second-order arithmetic are isomorphic.[4] Without doubting the mathematical importance and correctness of this theorem, it has been argued that Dedekind's approach nevertheless relies on problematic philosophical assumptions. The philosophical issues with Dedekinds approach are connected to the restriction on *full models* of arithmetic. As Button and Walsh (2018) notice:

> The 'second-order' component of 'full second-order logic' is surely unobjectionable. No one can prevent mathematicians from speaking a certain way, or from formalising their theories using any symbolism they like. The qualifying expression 'full', however, is rather more delicate. Here, it describes a particular semantics for second-order logic: one in which the second-order quantifiers must range over the entire powerset of the first-order domain of the structure. (Button and Walsh, 2018, p. 290)

Roughly, the issue at hand is that the notion of *full model of second-order arithmetic* cannot be characterised solely by means provided by second-order arithmetic. For this reason, this notion must be provided by other, external means, where 'external' is supposed to signalise that the means employed to characterise the target notion are meta-theoretic, and not provided by second-order arithmetic alone. However,

[4]For two slightly differently presented proofs of this theorem see for instance (Shapiro, 1991, p. 84, Theorem 4.10) and (Button and Walsh, 2016, p. 155, Theorem 7.3).

employing external notions, which are not given by the means available in second-order arithmetic, is problematic for several reasons. Button and Walsh (2018, p. 291) argue that this approach is vulnerable to a sort of Putnam-style 'just more theory' objection.[5] The idea behind the worry is that the new resources to characterise the target notion of full model of second-order arithmetic are just more theory: they are up to reinterpretation. Similarly, Parsons argues that employing external concepts to characterise the target notion of full model of second-order arithmetic is vulnerable to a form of *relativism*. An example discussed by Parsons (2008) is the case of set-theoretic relativism, where the notion of "full model of second-order arithmetic" is defined is some set theory:

> The observation about models of set theory would suggest a kind of relativism, akin to the relativism about cardinality that Skolem argued for on the basis of similar logical considerations: What we have in mind by "natural number" is relative to the underlying set theory. The language of set theory admits different interpretations, which give rise to different sets of natural numbers that are not even isomorphic structures. Parsons (2008, p. 275)

On the other hand, *internalism* tries to avoid these issues by avoiding the use of external notions not characterisable by means of the target theory.[6] When focusing on the case of second-order arithmetic, *internalism wants to avoid the talk of full models of arithmetic altogether*. Of course, to have any hope to do so, an internalist approach has to provide a different understanding of an arithmetical structure. Let

[5]Button (2022) argued for an analogous point.

[6]Core ideas of *internalism* have already been discussed by Parsons (1990), Parsons (2008, p. 112). Väänänen (2012) uses the term 'internal categoricity', but it can already be found earlier in Walmsley (2002). The philosophical position of internalism is introduced and discussed in (Button and Walsh, 2016, 2018).

us be more explicit about this issue: according to the external approach, an arithmetical structure is understood model-theoretically, for instance as a triple $(N, 0, s)$, where N is a set of objects, the domain of the model, 0 is the specific zero element in the domain, and s is a function on N that behaves as the successor function is supposed to behave. According to the external approach, full models of second-order arithmetic can be understood as quadruples $(N, \mathcal{P}(N), 0, s)$, where now $\mathcal{P}(N)$, is the full power set of N and is the domain of the second-order quantifiers of second-order arithmetic. To be a full model of second-order arithmetic can be externally understood as being such a quadruple that *satisfies* $\mathbf{PA^2}$, where $\mathbf{PA^2}$ is some suitable axiomatisation of second-order arithmetic:[7]

$$(N, \mathcal{P}(N), s, 0) \models \mathbf{PA^2} \tag{1}$$

In contrast to this approach, *internalism* wants to understand the talk about arithmetical structures and models of arithmetic without invoking these external, metatheoretic notions of a full model or satisfaction. An internal approach aims to understand arithmetical structures object-theoretically, as a triple with a one-place predicate, a first-order element, and a one-place function symbol, such that

$$\mathsf{PA}(\mathsf{N}, \mathsf{s}, 0), \tag{2}$$

holds of these elements, where $\mathsf{PA}(\mathsf{N}, \mathsf{s}, 0)$ is a shorthand for the conjunction of the axioms of second-order arithmetic relativised to N, s and 0. Analogously, when

[7] We will present this axiomatisation later. For the moment, it suffices to say that $\mathbf{PA^2}$ is just like first-order arithmetic, with the exception that instead of an induction schema $\mathbf{PA^2}$ employs and induction axiom, i.e., a single sentence, where – rougly – the second-order quantifier ranges over subsets of the natural numbers. Where the domain of the second-order quantifier is $\mathcal{P}(N)$, then the second-order quantifier in the induction axiom is thought of ranging over all subsets of the natural numbers.

considering an arithmetical sentence φ, the external approach spells out the claim that, for an arithmetical φ, φ is satisfied in a model of arithmetic similarly to (1):

$$(N, s, 0) \models \varphi \tag{3}$$

In contrast to this, internalism understand (3) object-linguistically:

$$\mathsf{PA}(\mathsf{N}, \mathsf{s}, 0) \rightarrow \varphi(\mathsf{N}, \mathsf{s}, 0) \tag{4}$$

where $\varphi(\mathsf{N}, \mathsf{s}, 0)$ is the relativisation of φ to the parameters N, s, and 0. It is crucial to see that the pair (1), (3) is quite different from (2), (4): in (2), (4) there is no reference to external, meta-theoretic notions, as it is completely expressed object-linguistically. As pointed out by Button and Walsh (2016), this position is inspired by Parsons' understanding of structures and "language as used is prior to semantic reflection on it." (Parsons, 2008). As Button and Walsh point out:

> We cannot emphasise enough the difference between the [external] expression $(N, s, 0) \models \mathbf{PA^2}$ and Parsons's expression $\mathsf{PA}(\mathsf{N}, \mathsf{s}, 0)$. Parsons expression does not mention any (sets of) sentences, and all suggestion of a satisfaction relation has vanished. So, when Parsons speaks of a 'model of second-order Peano arithmetic', there is no longer any hint of any language-object relation, and hence no hint of a model in the model-theorist's sense. Button and Walsh (2018, pp. 297-8)

However, although Button and Walsh (2016) are inspired by Parsons, they deviate from Parsons in a relevant respect. As we will point out later, they spell out the internal categoricity of second-order arithmetic – and a related result called *intolerance* – in the framework of pure second-order logic (with unrestricted comprehension). In

contrast to this, our investigation wants to be faithful to Parsons original approach to the categoricity of arithmetic in first-order logic.[8] The following section continues with a brief survey and with a discussion of two versions of internal categoricity for arithmetic: the version presented by Button and Walsh (2016) employing – as already mentioned – pure second-order logic, and the version presented by Parsons (1990, 2008), Feferman (2013) and Väänänen (2020).[9]

6.3 Internal Categoricity: Theorems and Discussion

Button and Walsh (2018, p. 228) prove a version of internal categoricity in second-order logic: they take as the base theory the system **CA** of pure second order logic with unrestricted second-order comprehension as a logical principle:

$$\exists X \forall x (x \in X \leftrightarrow \varphi(x)), \qquad (\mathsf{COMP})$$

i.e. (COMP) holds for any φ in which X does not occur free, formulated in the empty signature. That is, φ only contains connectives, existential and universal quantifiers, variables, second-order parameters, and the identity sign.[10] As pointed out in section 6.2, to be an arithmetical structure was spelled out as the following:

[8] For completenesse, it should be mentioned that internalism is not restricted to arithmetical theories and structures. Button and Walsh (2016) and Button (2022) investigate internalism concerning set-theory and even model-theory. To investigate these positions would exceed the scope of this dissertation.

[9] For completeness, it should be noted that we will not investigate the following two alternative versions of categoricity: a schematic version based on a language expansion by suitable expressions for two internal structures, exemplified by Väänänen and Wang (2015), and the variant formulated in a second-order arithmetical language with one primitive vocabulary proposed by Simpson and Yokoyama (2013).

[10] See (Button and Walsh, 2018, p. 224).

$\mathsf{PA}(\mathsf{N}, \mathsf{s}, 0)$, where $\mathsf{PA}(\mathsf{N}, \mathsf{s}, 0)$ was understood as a shorthand for the conjunction of the axioms of second-order arithmetic relativised to N, s and 0. Normally, one would think of N, s and 0 as constant symbols in an arithmetical language. However, in the context of pure second-order logic, this is expressed by the second-order formula $\mathsf{PA}(X, f, z)$, where X is a second-order variable, f is a function variable, and z is an individual variable. $\mathsf{PA}(X, f, z)$ is an abbreviation for the following:

$$z \in X \wedge \forall x(x \in X \rightarrow (\exists! y(y \in X \wedge f(x) = y))) \wedge \quad \text{(PA:1)}$$

$$\forall x(x \in X \rightarrow f(x) \neq z) \wedge \quad \text{(PA:2)}$$

$$\forall x, y(x \in X \wedge y \in X \rightarrow (f(x) = f(y) \rightarrow x = y)) \wedge \quad \text{(PA:3)}$$

$$\forall Z \subseteq X(z \in Z \wedge \forall y(y \in X \rightarrow (y \in Z \rightarrow f(y) \in Z))) \rightarrow X = Z \quad \text{(PA:4)}$$

$\mathsf{PA}(X, f, z)$ should remind us of a finite axiomatisation of second-order arithmetic. However, this formulation of $\mathsf{PA}(X, f, z)$ does not use or mention any fixed arithmetical vocabulary; instead of being a sentence of the fixed second-order arithmetical language, it is a formula in the deductive system of pure second-order logic with the parameters X, f, z.[11] Button and Walsh use $\mathsf{ISO}(Y, N, f, z, M, g, w)$, to state that 'Y is an isomorphism between the structures N, f, z and M, g, w'. ISO is short for

$$\forall x \forall y((x, y) \in Y \rightarrow (x \in N \wedge y \in M)) \wedge \quad \text{(iso:1)}$$

$$\forall x \in N \, \exists! y \in M((x, y) \in Y) \wedge \quad \text{(iso:2)}$$

$$\forall y \in M \, \exists! x \in N((x, y) \in Y) \wedge \quad \text{(iso:3)}$$

$$(z, w) \in Y \wedge \forall x, y((x, y) \in Y \rightarrow (f(x), g(y)) \in Y) \quad \text{(iso:4)}$$

[11] See (Button and Walsh, 2018, p. 228).

Roughly, (iso:1) says that Y is a map from N to M, (iso:2) and (iso:3) say that Y is a bijection and (iso:4) says – informally – that Y satisfies the structures under the respective 'successor functions' f and g. With this notion of internal isomorphism at hand, Button and Walsh (2016, p. 228, Theorem 10.2) prove the following theorem:

Theorem 14 (Button and Walsh)**. CA** *proves*

$$\forall N \forall f \forall z \, \forall M \forall g \forall w (\mathsf{PA}(N, f, z) \wedge \mathsf{PA}(M, g, w) \rightarrow \exists X \; \mathsf{ISO}(X, N, f, z, M, g, w))$$

The proof is basically an internalisation of the proof given by Shapiro (1991, Theorem 4.10) and is quite similar to the proof of the internal categoricity of second-order arithmetic provided by Väänänen and Wang (2015, Theorem 1).[12] The gist of the idea of the proof is to use the following second-order formula $H(Y)$, expressing that Y is hereditary:

$$H(Y) \leftrightarrow ((z, w) \in Y \wedge (\forall x \in N)(\forall y \in M)[(x, y) \in X \rightarrow (f(x), g(y)) \in Y]),$$

By the schema of comprehension in **CA** then Button and Walsh prove the existence of the set of all objects that are in the universal closure of all Z, for hereditary Z:

$$\exists X \forall x \forall y ((x, y) \in X \leftrightarrow \forall Z[H(Z) \rightarrow (x, y) \in Z]), \qquad (\#)$$

They prove the internal isomorphism by verifying that, for an arbitrary R instantiating the second-order existential quantifier in (#), R satisfies (iso:1), (iso:2), (iso:3) and (iso:4). Button and Walsh (2016, p. 228) interpret this theorem as saying that all internal arithmetical structures are internally isomorphic.

[12]For the proof provided by Button and Walsh see (Button and Walsh, 2016, pp. 243-244).

Although the existence of the isomorphism in (#) is proved by impredicative comprehension, Parsons (2008, p. 281) pointed out that the proof of categoricity is *essentially first-order*. For the construction of the isomorphism it is possible to rely on suitable recursion to define the isomorphism from below. A justification of such a move and the problem of the expressive weakness of first-order theories are often handled by reference to an open-ended conception of arithmetic. Such an open-ended conception of arithmetic does not settle on a fixed first-order arithmetical theory, but allows for suitable language expansions by 'definite' predicates, that then are allowed in induction. An explicit proof of such a first-order version has been given recently in (Väänänen, 2020) bearing some resemblance to the schematic second-order version. Since first-order categoricity results are less familiar than their second-order counterparts and we will build upon it later we sketch the proof here. We work in a first order language $\mathcal{L}^{ij}$ with a signature containing two primitive arithmetical vocabularies $(\mathsf{N}^i, \mathsf{S}^i, 0^i, +^i, \times^i, \mathsf{N}^j, \mathsf{S}^j, 0^j, +^j, \times^j)$. The theory $\mathbf{PA}^{ij}$ is then Peano arithmetic in the language $\mathcal{L}^{ij}$ with the axioms of $\mathbf{Q}$ for both vocabularies and full induction for formulas φ of the mixed language $\mathcal{L}^{ij}$ for both number properties N^i as well as N^j, i.e.

$$\varphi(0^i) \wedge \forall x(\mathsf{N}^i(x) \to (\varphi(x) \to \varphi(\mathsf{S}^i(x)))) \to \forall x(\mathsf{N}^i(x) \to \varphi(x)) \qquad (\mathsf{IND}^i)$$

$$\varphi(0^j) \wedge \forall x(\mathsf{N}^j(x) \to (\varphi(x) \to \varphi(\mathsf{S}^j(x)))) \to \forall x(\mathsf{N}^j(x) \to \varphi(x)) \qquad (\mathsf{IND}^j)$$

$\mathsf{ISO}(\chi)$ is short for:

$$\forall x \forall y(\chi(x,y) \to (\mathsf{N}^i(x) \wedge \mathsf{N}^j(y))) \wedge \text{(iso:1)}$$
$$\forall x(\mathsf{N}^i(x) \to \exists! y(\mathsf{N}^j(y) \wedge \chi(x,y)) \wedge \text{(iso:2)}$$
$$\forall y(\mathsf{N}^j(y) \to \exists! x(\mathsf{N}^i(x) \wedge \chi(x,y) \wedge \text{(iso:3)}$$

$$\chi(0^i, 0^j) \wedge \forall x, y(\chi(x, y) \to \chi(\mathsf{S}^i(x), \mathsf{S}^j(y))) \wedge \text{(iso:4)}$$

$$\forall x, y, z, x', y'(\chi(x, x') \wedge \chi(y, y') \wedge \chi(x +^i y, z) \to z = x' +^j y') \wedge \text{(iso:5)}$$

$$\forall x, y, z, x', y'(\chi(x, x') \wedge \chi(y, y') \wedge \chi(x \times^i y, z) \to z = x' \times^j y') \text{ (iso:6)}$$

$\mathsf{ISO}(\chi)$ is supposed to be an approximation of the internal isomorphism proved by Button and Walsh in pure second-order logic. There are some relevant distinctions: we cannot prove the existence of the isomorphism explicitly, but one can only witness the existence of a formula expressing the isomorphism. Moreover, this isomorphism is not proved for all structures, but only for two given copies of the language of arithmetic. With this notion of internal isomorphism at hand, one can prove a first-order, schematic version of the theorem proved by Button and Walsh.

Theorem 15 (Parsons, Feferman, Väänänen)**.** *There is a $\Sigma_1[\mathcal{L}^{ij}]$-formula χ, such that*

$$\mathbf{PA}^{ij} \vdash \mathsf{ISO}(\chi)$$

Sketch. In the arithmetical language $\mathcal{L}^{ij}$ containing two copies of the arithmetical language we have two formulas χ, χ' in $\Sigma_1[\mathcal{L}^{ij}]$ representing the primitive recursive functions $f : N^i \to N^j$ and $g : N^j \to N^i$, such that:

$$f(0^i) = 0^j; \quad \forall x(\mathsf{N}^i(x) \to [f(\mathsf{S}^i(x)) = \mathsf{S}^j(f(x))])$$

$$g(0^j) = 0^i; \quad \forall y(\mathsf{N}^j(y) \to [g(\mathsf{S}^j(y) = \mathsf{S}^i(g(y))])$$

One proves $\mathsf{ISO}(\chi)$ employing these functions:

1. $\mathbf{PA}^{ij}$ proves the following: $\forall x(\mathsf{N}^i(x) \to g(f(x)) = x)$ by (IND^i). $g(f(0^i) = g(0^j) = 0^i$. Assume that $g(f(x)) = x$ and show that $g(f(\mathsf{S}^i(x)) = \mathsf{S}^i(x)$. We have the following: $g(f(S^i(x)) = g(S^j(f(x)) = S^i(g(f(x)) = S^i(x)$. Officially

we have induction on a formula containing the formulas χ, χ', so the instance of (IND^i) used, contains both vocabularies.

2. $\mathbf{PA}^{ij}$ proves that $\forall y(\mathsf{N}^j(y) \to f(g(y)) = y$ similarly by (IND^j).

3. $\mathbf{PA}^{ij}$ proves that $f : N^i \to N^j$. Show $\forall x(\mathsf{N}^i(x) \to \exists! y(\mathsf{N}^j(y) \wedge \chi(x, y)))$ by induction on N^i. The existential claim for 0^i is obvious. The uniqueness follows by the axiom $\forall x(\mathsf{N}^i(x) \to x = 0^i \vee \exists y(\mathsf{N}^i(y) \wedge x = \mathsf{S}^i(y)))$ and the axiom $\forall x(\mathsf{N}^j(x) \to \mathsf{S}^j(x) \neq 0^j)$. For the induction step we assume $\mathsf{N}^i(x) \to \exists! y(\mathsf{N}^j(y) \wedge \chi(x, y))$ and $\mathsf{N}^i(\mathsf{S}^i(x))$. The existence follows by the IH and the axiom $\forall x(\mathsf{N}^j(x) \to \exists y(\mathsf{N}^j(y) \wedge y = \mathsf{S}^j(x)))$. The uniqueness follows by the injectivity of S^j and the IH.

4. $\mathbf{PA}^{ij}$ proves that $g : N^j \to N^i$ similarly.

5. $\mathbf{PA}^{ij}$ proves that f is injective, i.e. $\forall xy(\mathsf{N}^i(x) \wedge \mathsf{N}^i(y) \to (f(x) = f(y) \to x = y))$. We assume that for a, b, $N^i(a) \wedge N^i(b)$, $f(a) = f(b)$ and that $a \neq b$. Using the fact that $\forall x(N^i(x) \to g(f(x)) = x)$, we have that $g(f(a)) \neq g(f(b))$. However, this contradicts the functionality of g in 4 and of f in 3. Therefore, f is 1-1.

6. $\mathbf{PA}^{ij}$ proves that f is onto, i.e. that $\forall y(\mathsf{N}^j(y) \to \exists x(\mathsf{N}^i(x) \wedge f(x) = y))$. By induction on N^j. For 0 obvious. Assume that $\mathsf{N}^j(\mathsf{S}^j(y))$. By IH $\exists x(\mathsf{N}^i(x) \wedge f(x) = y))$. If $\mathsf{N}^i(x)$, then also $\mathsf{N}^i(\mathsf{S}^i x)$ and $f(\mathsf{S}^i x) = S^j(y)$.

   The bijectiveness of f establishes (iso:2) and (iso:3). (iso:1) and (iso:4) are given by definition.

7. $\mathbf{PA}^{ij}$ proves (iso:5) by induction on y in N^i:

   If $y = 0^i$, then $f(x +^i 0^i) = f(x) = x' = x' +^j 0^j$.

If $y = \mathsf{S}^i(n)$, then

$$f(x +^i y) = f(x +^i \mathsf{S}^i(n)) = f(\mathsf{S}^i(x +^i n)) = \mathsf{S}^j(f(x +^i n)) =$$
$$f(x) +^j \mathsf{S}^j(f(n)) = f(x) +^j f(\mathsf{S}^i(n)) = f(x) +^j f(y)$$

8. $\mathbf{PA}^{ij}$ proves (iso :6) analogously.

□

One can gloss on this theorem as saying that the two given internal structures are internally isomorphic.

One great advantage of the pure second-order logic version of categoricity is its generality. As I pointed out, Button and Walsh prove the internal isomorphism for all arithmetical structures. Moreover, they can prove the existence of the isomorphism directly. In contrast, as (Button and Walsh, 2016, p. 240) point out, the first-order version of the isomorphism faces some *expressive limitations*. We cannot prove the existence of the isomorphism explicitly, but one can only witness the existence of a formula expressing the isomorphism. Moreover, this isomorphism is not proved for all structures, but only for the two given copies of the language of arithmetic. Section 6.4 aims to overcome these expressive limitations concerning the first-order version of internal categoricity, by introducing a primitive truth predicate. Before going into the details of this approach, the following section discusses the question of determinacy of arithmetical truth.

6.3.1 The Issue of Determinacy

There is a sense in which internal categoricity implies determinacy. Button and Walsh argue that a corollary of the categoricity theorem, called *intolerance*, implies that arithmetical truth is determinate.[13] Intolerance establishes that for any arithmetical sentence φ, all internally isomorphic structures agree on the truth value of φ: deviations are not tolerated so to say. In this sense, intolerance is an internal version of elementary equivalence. Although we basically follow Button and Walsh (2018, p. 245) I have renamed their intermediate step as intolerance. The reason is that we think that this already contains the interesting fact that deviations are not tolerated.

Theorem 16 (Button and Walsh)**.** *In* **CA** *we can prove for any* $\varphi \in \mathcal{L}_A^2$*:*[14]

$$\mathsf{ISO}(R, N, f, z, M, g, w) \rightarrow [\varphi(N, f, z) \leftrightarrow \varphi(M, g, w,)]$$

From the internalist's perspective, this is a theorem saying that 'no object-language deviation between internal-structures is tolerated' (Button and Walsh, 2018, p. 232). In other words, two internal structures must evaluate all arithmetical sentences uniformly. Since internalism avoids employing or mentioning any semantic notion, evaluation in an internal structure is stated in this form.

The following section tackles the main issue concerning the first-order version of categoricity and aims to provide a *generalised* version of Theorem 15. As we will

[13] Button and Walsh also discuss how internal categoricity obviously fails to 'pin down the natural numbers' (Button and Walsh, 2018, p. 231), when structures are understood externally. However, this is not a problem for the present investigation, since we focus on internalism.

[14] For a proof of Theorem 16 see (Button and Walsh, 2018, p. 245). For easier readability we did not mention the additional parameters.

show, this generalised version of Theorem 15 – introduced in subsection 6.4.2 as Proposition 9 – is an approximation of the theorem provided by Button and Walsh. For the moment informally, this approximation is going to be achieved by quantifying over arbitrary internal structures by means of the truth predicate.[15]

6.4 Internal categoricity and truth

This section introduces a primitive notion of truth to extend our first-order arithmetical theory. The aim is to provide a result that overcomes some of the shortcomings of the previous internal categoricity results. Subsection 6.4.1 expands the internal isomorphism of Theorem 15 to an isomorphism between internal structures *with their respective notions of truth.* We interpret this result as establishing an internalised version of intolerance, similar to Theorem 16, which partially overcomes the expressive limitations of the first-order categoricity result. Analogously to Button and Walsh's interpretation of intolerance as a justification for a *univocal* arithmetical theory, we understand the truth theoretic intolerance as a sufficient reason for a univocal theory of truth. This result will be exemplary for our discussion of determinacy in section 6.5. Finally, section 6.4.2 focuses on generalising Theorem 15 to an internal isomorphism between 'all' internal arithmetical structures.

6.4.1 Unique truth

Button and Walsh (2018) showed that internal categoricity can be used to argue for a canonical theory of arithmetic. In this section, we expand this observation to theories of arithmetical truth. We sketch an argument for an extension of the

[15] As we will point out later, this follows an argument by Feferman and Hellman (1995). However, instead of using a primitive notion of truth, they rely on a theory of finite sets and classes.

internal categoricity claim including a notion of truth, providing a truth-theoretic version of intolerance (Proposition 8). To do this, we expand our mixed arithmetical theory of arithmetic $\mathbf{PA}^{ij}$ by two compositional theories of truth. We expand our two languages $\mathcal{L}^i$ and $\mathcal{L}^j$ by truth predicates so that for our mixed language $\mathcal{L}^{ij}_{\mathsf{T}}$ the signature is $(0^i, 0^j, \mathsf{N}^i, \mathsf{N}^j, \mathsf{S}^i, \mathsf{S}^j, +^i, +^j, \times^i, \times^j, \mathsf{T}^i, \mathsf{T}^j)$. The intended range of the truth predicate T^i (respectively T^j) are the sentences of the arithmetical sublanguage $\mathcal{L}^i_A$ (respectively $\mathcal{L}^j_A$). We sometimes use φ^i as the result of substituting in $\varphi \in \mathcal{L}_A$ the respective vocabulary by its i-counterparts.

We will establish the internal categoricity in a theory $\mathbf{CT}^{ij}$ which is basically two versions of the compositional axioms expanding the basic arithmetical axioms for k. So for $k \in \{i, j\}$ we have:

$$(\mathsf{CT1}^k) \qquad \mathsf{ct}^k(x) \wedge \mathsf{ct}^k(y) \rightarrow (\mathsf{T}^k(x \doteq^k y) \leftrightarrow \mathsf{val}^k(x) = \mathsf{val}^k(y))$$

$$(\mathsf{CT2}^k) \qquad \mathsf{sent}^k(x) \rightarrow (\mathsf{T}^k(\dot{\neg}^k x) \leftrightarrow \neg \mathsf{T}^k(x))$$

$$(\mathsf{CT3}^k) \qquad \mathsf{sent}^k(x) \wedge \mathsf{sent}^k(y) \rightarrow (\mathsf{T}^k(x \mathbin{\dot{\wedge}^k} y) \leftrightarrow \mathsf{T}^k(x) \wedge \mathsf{T}^k(y))$$

$$(\mathsf{CT4}^k) \qquad \mathsf{form}^k(x) \wedge \mathsf{var}^k(v) \rightarrow (\mathsf{T}^k(\dot{\forall}^k v\, x) \leftrightarrow \forall y \in \mathsf{N}^k\, \mathsf{T}^k(\mathsf{sub}^k(x, \mathsf{num}^k(y))))$$

We also allow for formulas of our mixed language $\mathcal{L}^{ij}_{\mathsf{T}}$ (including both truth predicates) to appear in both our induction principles IND^i and IND^j.

We want to expand the isomorphism f of Theorem 15 in a natural way by assuming that it is possible to identify the relevant syntactical primitives, i.e. we assume that for all syntactical constants e, $f(e^i) = e^j$. Moreover, we assume that f commutes with the syntactical application functions, so that term and formula

building operations are preserved under our isomorphism f. So for the applications $f(\mathsf{appl}^i(x,y,z) = \mathsf{appl}^j(f(x),f(y),f(z))$.[16] This suffices to expand our isomorphism to all the syntactical predicates and functions since the first-order version of the isomorphism establishes that for all $\varphi(x) \in \mathcal{L}_A$:

$$\mathbf{PA}^{ij} \vdash \forall x(\varphi^i(x) \leftrightarrow \varphi^j(f(x))) \tag{‡}$$

We add to our previous $\mathsf{ISO}_{i \rhd j}(\chi)$ (Section 3.3) the additional requirement:

$$\forall x, x'(\chi(x,x') \wedge \mathsf{T}^i(x) \leftrightarrow \mathsf{T}^j(x')) \tag{iso:7}$$

Proposition 8. *There is some $\chi \in \mathcal{L}_{\mathsf{T}}^{ij}$, such that*

$$\mathbf{CT}^{ij} \vdash \mathsf{ISO}(\chi)$$

Proof. We use the same construction of χ as in the proof of Theorem 15 witnessing the isomorphism f. Since (iso:1)-(iso:6) are already established we show the additional (iso:7) arguing in $\mathbf{CT}^{ij}$:

By (‡) we know that if $\neg\mathsf{sent}^i(x)$, then $\neg\mathsf{sent}^j(f(x))$.

Now if $\mathsf{sent}^i(x)$ we argue by (internal) induction on $|x|$, which is the grade of x, that if $\mathsf{T}^i(x)$ then $\mathsf{T}^j(f(x))$.

1. If $\mathsf{sent}^i(x)$ and $|x| = 0$, then x is $s \doteq^i t$ for some closed terms s, t, with $\mathsf{term}^i(s)$ and $\mathsf{term}^i(t)$. By assumption $\mathsf{T}^i(s \doteq^i t)$ and so by $(\mathsf{CT1}^i)$ we have $\mathsf{val}^i(s) =$

[16] A possibly more elaborate version would employ some disentangled setting as in (Leigh and Nicolai, 2013). In contrast to that setting, in which bridging principles (so-called 'coding axioms') are postulated, in (Mount and Waxman, 2021) a categoricity argument to establish a form of intolerance in a second-order setting is provided, in which the bridging principles are derivable.

$\mathsf{val}^i(t)$. By (‡) we have $\mathsf{val}^i(s) = \mathsf{val}^i(t) \Leftrightarrow \mathsf{val}^j(f(s)) = \mathsf{val}^j(f(t))$. By ($\mathsf{CT1}^j$) we get $\mathsf{T}^j(f(s) \doteq^j f(t))$.

2. If x is $\dot{\neg}^i y$ for some $\mathsf{form}^i(y)$ with $|y| < |x|$. Then by assumption and ($\mathsf{CT2}^i$) we have $\neg\mathsf{T}^i(y)$. Since $|y| < |x|$ we can use the IH to argue that $\neg\mathsf{T}^j(f(y))$. By ($\mathsf{CT2}^j$) we have $\mathsf{T}^j(\dot{\neg}^j f(y))$ and by (7) we have that $\mathsf{T}^j(f(\dot{\neg}^i y))$.

3. If x is $y \mathbin{\dot{\wedge}^i} z$ for some $\mathsf{form}^i(y)$, $\mathsf{form}^i(z)$ with $|y|, |z| < |x|$, we argue similarly.

4. If x is $\dot{\forall}^i v y$ for some $\mathsf{form}^i(y)$ and $\mathsf{var}^i(v)$ with $|y| < |x|$, then by $\mathsf{T}^i(x)$ and by ($\mathsf{CT4}^i$) we have that $\forall z \in \mathsf{N}^i\mathsf{T}^i(\mathsf{sub}^i(y, \mathsf{num}^i(z))$. Then we see by IH that $\forall z \in \mathsf{N}^j\mathsf{T}^j(f(\mathsf{sub}^i(y, \mathsf{num}^i(z)))$. By (‡) we have that $\forall z \in \mathsf{N}^j\mathsf{T}^j(\mathsf{sub}^j(f(y), \mathsf{num}^j(f(z))))$. And by ($\mathsf{CT4}^j$) we have that $\mathsf{T}^j \dot{\forall}^j f(v) f(y)$.

□

Proposition 8 supports the idea that the relativization of the two truth predicates is superfluous and that one can adopt a univocal theory of truth. This is not unplausible from an internalist perspective. Button and Walsh argue for an analogous conclusion in (Button and Walsh, 2018, Section 10.5.) concerning the natural number predicates; the proof of the internal isomorphism provides a warrant to drop the indices and accept a univocal theory of arithmetic. Therefore, we can remove the indices and work with a unified theory of truth, i.e. a single truth predicate for both languages.

6.4.2 General categoricity and intolerance with truth

So far we have been arguing schematically, i.e., for two arbitrarily given internal arithmetical structures. We will now generalize the result to get a general version

of categoricity, in which we quantify over arbitrary arithmetical interpretations.

To do this, we generalize the reverse mathematics version of the argument following a strategy due to Feferman and Hellman (1995). Similarly to the reverse mathematics case, we use a primitive set of notions only for one natural number structure, breaking the symmetry with the previous –schematic– case, where we worked with two given arithmetical structures and two sets of primitive vocabularies. Whereas we use our mathematical vocabulary for our usual mathematical discourse, to talk about *our* natural numbers, we interpret the 'alternative' internal arithmetical structures as possible language expansions, so that we can only indirectly talk about them by mentioning the expanded linguistic resources. We expand our linguistic repertoire in a suitable way to talk about the alternative interpretation, mainly by using the truth predicate. Under these assumptions we can show that any alternative arithmetical interpretation is equivalent to 'our' interpretation.

We will support this claim by providing an internal categoricity result that is based on the categoricity result due to Feferman and Hellman (1995). For the bridging principles they use a theory of finite sets and classes called **EFSC**.[17] Our approach replaces the finite set theory with our expanded theory of syntax and truth.[18] In the following we make the assumptions on the expanded theory of syntax explicit. We work within the language of first-order **PA**, with a predicate $\mathbb{N}$ for 'our' natural numbers and an additional truth predicate T. The arithmetical axioms are formu-

[17]For the presentation of **EFSC** see (Feferman and Hellman, 1995, p. 3 -4). Although it is interpretable in $\mathbf{ACA_0}$, the interpreted version is not as interesting as it could be, due to the fact that by interpreting all objects as elements of $\mathbb{N}$ also the internal models will be subsets of $\mathbb{N}$. This is a quite similar assumption as in the reverse version that we gave up.

[18]It is well known that there is a close connection between adjunctive set theory and concatenation theory, see for example Damjanovic (Damnjanovic, 2017).

lated relativized to this $\mathbb{N}$ and the theory of syntax is standard for our language. In order to talk about alternative arithmetical interpretations, we employ a language expansion $\mathcal{L}_{\mathsf{T}}(P)$ of $\mathcal{L}_{\mathsf{T}}$ by an arbitrary one-place predicate P, whose intended interpretation is an alternative natural number property.

To expand our theory of truth in a suitable way several assumptions must be in place. First of all the quantifiers are not restricted to $\mathbb{N}$ so that $\forall x\varphi$ is not the same as $\forall x(\mathbb{N}(x) \rightarrow \varphi)$. So also $\forall x(P(x) \rightarrow \mathbb{N}(x))$ is not assumed as in the case of (Simpson and Yokoyama, 2013). Additionally, we make some assumptions with respect to our language and interpretations: **CT** is based on a general theory of syntax. We allow the syntactical predicates to include expressions built up from vocabularies different from the basic ones. We assume that the language expansions are based on the usual syntactical assumptions that are captured by an application function. We allow for an expansion of the class of predicates, constants and function symbols. The additional resources (relative to P) are captured by syntactical predicates $\mathsf{pred}^P, \mathsf{const}^P, \mathsf{func}^P$. In order to do so, we adjust the syntactical vocabulary, so that the general syntactical principles hold. For instance the class of terms should now be closed under the expanded function symbol application. Our syntactic repertoire includes syntactical predicates representing the syntactical categories of the expanded language. So there is a formula $\mathsf{ct}(x)$ that represents the closure of the set of individual constants under function application, and similarly a formula form representing the set of formulas for an expanded language.

It is now possible to characterise an arbitrary arithmetical interpretation, which we call a Peano system.[19] A P-Peano system (a Peano system relative to P) is a triple $(\overline{P}, h, a)$, such that $\mathsf{pred}^P(\overline{P})$, $\mathsf{func}^P(h)$ and $\mathsf{const}^P(a)$. The first component is intended to be a sortal predicate for the range of the quantifiers of the Peano system, the second component is a one-place function symbol, representing the successor function of the Peano system, and the third a constant denoting the zero-element of the Peano system. In order to have a suitable form of quantification over terms we assume that the constant and the function symbol form a systematic naming device for the Peano system. For our general application function we also assume a representation appl, that will also operate on the terms of the expanded language. More specifically, we assume that besides the standard one-place term predicate term^A, we have an additional term^P predicate that is intended to range over the 'terms' of the expanded language. In the following we list the linguistic part of the assumptions on a P-Peano system:

$$\mathsf{const}^P(x) \to \mathsf{term}^P(x) \tag{L_1}$$

$$\mathsf{var}(x) \to \mathsf{term}^P(x) \tag{L_2}$$

$$\mathsf{term}^P(x) \wedge \mathsf{func}^P(y) \to \mathsf{term}^P(\mathsf{appl}(y, x)). \tag{L_3}$$

We take closed terms ct^P as usual without free variables. We assume that we only have standard names for elements of P. We also have a generalized form of the term valuation function val. We use $n, m, ...$ for variables ranging over $\mathbb{N}$, $v, w, ...$ for variables ranging over P, and $x, y, ...$ for unrestricted quantification.

[19] Such a terminology is used in Simpson's and Yokoyama's (Simpson and Yokoyama, 2013).

$$\forall n \exists y(\mathsf{ct}^A(y) \wedge \mathsf{val}(y) = n) \tag{L_4}$$

$$\forall v \exists! y(\mathsf{ct}^P(y) \wedge \mathsf{val}(y) = v) \tag{L_5}$$

We use again s, t as abbreviations for closed terms. In line with our expansion, we can expand the function $\mathsf{num}(x)$. Whereas in the case of elements of $\mathbb{N}$ it assigns the x-th numeral, in the case of elements of P it assigns the unique x-th standard term. In the following we will use $x^\bullet$ as shorthand for $\mathsf{num}(x)$, expanding the usual $\dot{x}$ notation.

$$\forall x \in P(\mathsf{num}(x) = y \leftrightarrow \mathsf{ct}^P(y) \wedge \mathsf{val}(y) = x) \tag{L_6}$$

$$\mathsf{val}(a) \in P \wedge \forall x(P(x) \rightarrow P(\mathsf{val}(\mathsf{appl}(h, x^\bullet)))) \tag{L_7}$$

Analogously, we expand the notion of a P-formula, form^P, for formulas only containing notions of the expanded language and closed under application of conjunction, negation and $\mathbb{N}$-restricted quantification and P-restricted quantification.

$$\mathsf{term}^P(s) \wedge \mathsf{term}^P(t) \rightarrow \mathsf{form}^P(\mathsf{appl}(\doteq, s, t)) \tag{L_8}$$

$$\mathsf{term}^P(t) \wedge \mathsf{pred}^P(\overline{P}) \rightarrow \mathsf{form}^P(\mathsf{appl}(\overline{P}, t)) \tag{L_9}$$

The usual closure conditions for $\neg, \wedge, \forall$ are labelled as (L_{10}) - (L_{12}). We use $\ulcorner\varphi\urcorner$ as a shorthand for formulas and specifically $\ulcorner\varphi(x)\urcorner$ for a formula with x the only free variable. Also the syntactic substitution function is generalized to sub. For the

substitution of terms in formulas we use [ ] to indicate that these are quantified from the outside, so for example $\ulcorner\varphi[t/v]\urcorner$ is short for $(\mathsf{sub}(\ulcorner\varphi\urcorner, v, t)))$ and if an indicated free variable is used then $\ulcorner\varphi[t]\urcorner$ is short for $\mathsf{sub}(\ulcorner\varphi\urcorner, t)$.

Let $\mathsf{PPS}^P(\overline{P}, h, a)$ be short for $\mathsf{pred}^P(\overline{P}) \wedge \mathsf{func}^P(h) \wedge \mathsf{const}^P(a) \wedge \bigwedge_{i \leq 12}(L_i)$. With this we formulate the system $\mathbf{CT}[P]$, the $\mathcal{L}_{\mathsf{T}}(P)$-theory containing the following axioms relativised to P, intended to be the sortal predicate of an arbitrary Peano system:

$$\mathsf{PPS}^P(\overline{P}, h, a) \Rightarrow \forall s \forall t \mathsf{T}(s \doteq t) \leftrightarrow \mathsf{val}(s) = \mathsf{val}(t) \tag{$\mathsf{CT1}^P$}$$

$$\mathsf{PPS}^P(\overline{P}, h, a) \Rightarrow \forall t(\mathsf{T}(\mathsf{appl}(\overline{P}, t)) \leftrightarrow P(\mathsf{val}(t))) \tag{$\mathsf{CT2}^P$}$$

$$\mathsf{PPS}^P(\overline{P}, h, a) \Rightarrow \mathsf{T}\ulcorner \dot{\neg} \varphi\urcorner \leftrightarrow \neg \mathsf{T}\ulcorner\varphi\urcorner \tag{$\mathsf{CT3}^P$}$$

$$\mathsf{PPS}^P(\overline{P}, h, a) \Rightarrow \mathsf{T}\ulcorner\varphi \mathbin{\dot{\wedge}} \psi\urcorner \leftrightarrow \mathsf{T}\ulcorner\varphi\urcorner \wedge \mathsf{T}\ulcorner\psi\urcorner \tag{$\mathsf{CT4}^P$}$$

$$\mathsf{PPS}^P(\overline{P}, h, a) \Rightarrow \forall t \in \mathsf{ct}^A \mathsf{T}\ulcorner\varphi[t]\urcorner \leftrightarrow \mathsf{T}\ulcorner\dot{\forall} x(\dot{\mathbb{N}}(x) \mathbin{\dot{\rightarrow}} \varphi)\urcorner \tag{$\mathsf{CT5}^P$}$$

$$\mathsf{PPS}^P(\overline{P}, h, a) \Rightarrow \forall t \in \mathsf{ct}^P \mathsf{T}\ulcorner\varphi[t]\urcorner \leftrightarrow \mathsf{T}\ulcorner\dot{\forall} x(\mathsf{appl}(\overline{P}, x) \mathbin{\dot{\rightarrow}} \varphi)\urcorner \tag{$\mathsf{CT6}^P$}$$

With the truth predicate and $\mathbf{CT}[P]$ in the background we can formulate the usual properties of a Peano system $\mathsf{PS}^P(\overline{P}, h, a)$ i.e. $\mathsf{PPS}^P(\overline{P}, h, a)$ plus

$$\forall v(\neg \mathsf{T}(\mathsf{appl}(h, v^\bullet) \doteq a)) \tag{N-1}$$

$$\forall v, w(\mathsf{T}(\mathsf{appl}(h, v^\bullet) \doteq \mathsf{appl}(h, w^\bullet))) \rightarrow \mathsf{T}(v^\bullet \doteq w^\bullet)) \tag{N-2}$$

$$\forall \ulcorner\varphi(x)\urcorner(\mathsf{T}\ulcorner\varphi[a]\urcorner \wedge \forall v(\mathsf{T}\ulcorner\varphi[v^\bullet]\urcorner \rightarrow \mathsf{T}\ulcorner\varphi[\mathsf{appl}(h, v^\bullet)]\urcorner) \rightarrow \mathsf{T}\ulcorner\dot{\forall} v\varphi\urcorner) \tag{N-3}$$

With our naming machinery we can easily build singleton sets of elements of either $\mathbb{N}$ or P using our num-function,

$$\forall x \exists \ulcorner \varphi \urcorner \forall y (\mathsf{T} \ulcorner \varphi [y^\bullet] \urcorner \leftrightarrow y = x).$$

just by taking $\ulcorner \varphi \urcorner$ to be $\mathsf{num}(y) \doteq \mathsf{num}(x)$.

Important for our case is that we can also build singletons of ordered pairs of elements of $\mathbb{N}$ and P simultaneously. This is due to our naming machinery and the fact that our syntax theory and truth work for the wider range. Additionally, we can use disjunctions to mimic the talk about 'finite' subsets by adjunction.

$$\forall \ulcorner \varphi \urcorner \exists \ulcorner \psi \urcorner \forall x, y (\mathsf{T} \ulcorner \psi [y^\bullet] \urcorner \leftrightarrow \mathsf{T} \ulcorner \varphi [y^\bullet] \urcorner \vee y = x)$$

With this in place, we can carry out a proof in $\mathbf{CT}[P]$ by following Feferman and Hellman's strategy. The basic task is to define a function from our natural numbers into the domain of P, s.t. the following holds

$$\begin{aligned} f(0) &= a \qquad (*) \\ f(\mathsf{S}(n)) &= \mathsf{appl}(h, f(n)) \end{aligned}$$

The proof strategy is to approximate this function using suitable formulas for 'finite' sets of elements of both 'domains'. We let $\mathsf{Rec}(\ulcorner \varphi \urcorner, h, a, n)$ be the conjunction of (R_1) - (R_3) with

$$\forall m \forall w (\mathsf{T} \ulcorner \varphi [m^\bullet, w^\bullet] \urcorner \rightarrow m \leq n) \qquad (R_1)$$

$$\forall w (\mathsf{T} \ulcorner \varphi (\overline{0}, w^\bullet) \urcorner \leftrightarrow \mathsf{T} \ulcorner w^\bullet \doteq a \urcorner) \qquad (R_2)$$

$$\forall m < n \forall w (\mathsf{T} \ulcorner \varphi [\dot{\mathsf{S}}\, m^\bullet, w^\bullet] \urcorner \leftrightarrow \exists u (\mathsf{T} \ulcorner \varphi [m^\bullet, u^\bullet] \urcorner \wedge \mathsf{T} \ulcorner \mathsf{appl}(h, u^\bullet) \doteq w^\bullet \urcorner)) \qquad (R_3)$$

Since our compositional theory works for the expanded language we can as usual use the commutation of truth with the connectives to show the following Lemma:

Lemma 1.

$$\exists \ulcorner\psi\urcorner \forall n(\mathsf{T}\ulcorner\psi[\ulcorner\varphi\urcorner, h, a, n^\bullet]\urcorner \leftrightarrow \mathsf{Rec}(\ulcorner\varphi\urcorner, h, a, n))$$

We show the following (Theorem 3. in (Feferman and Hellman, 1995)):

Lemma 2. *(i)* $\mathsf{Rec}(\ulcorner\varphi\urcorner, h, a, n) \rightarrow \forall m \leq n \exists! w(\mathsf{T}\ulcorner\varphi[m^\bullet, w^\bullet]\urcorner)$*;*

(ii) $\mathsf{Rec}(\ulcorner\varphi\urcorner, h, a, n) \wedge \mathsf{Rec}(\ulcorner\psi\urcorner, h, a, n) \rightarrow \forall m \forall v(\mathsf{T}\ulcorner\varphi[m^\bullet, v^\bullet] \dot{\leftrightarrow} \psi[m^\bullet, v^\bullet]\urcorner)$*;*

(iii) $\forall n \, \exists \ulcorner\varphi\urcorner \, \mathsf{Rec}(\ulcorner\varphi\urcorner, h, a, n)$*;*

(iv) $\mathsf{Rec}(\ulcorner\psi\urcorner, h, a, \mathsf{S}(n)) \wedge \forall m \forall w(\mathsf{T}\ulcorner\varphi[m^\bullet, w^\bullet]\urcorner \leftrightarrow \mathsf{T}\ulcorner\psi[m^\bullet, w^\bullet]\urcorner \wedge m \leq n) \rightarrow \mathsf{Rec}(\ulcorner\varphi\urcorner, h, a, n)$*;*

(v) $(\mathsf{Rec}(\ulcorner\varphi\urcorner, h, a, n) \wedge \mathsf{Rec}(\ulcorner\psi\urcorner, h, a, m) \wedge n \leq m) \rightarrow \forall l, w(\mathsf{T}\ulcorner\varphi[l^\bullet, w^\bullet]\urcorner \rightarrow \mathsf{T}\ulcorner\psi[l^\bullet, w^\bullet]\urcorner)$.

Sketch. The proof follows Feferman and Hellman (Feferman and Hellman, 1995, p. 8).

(*i*) We argue by induction on $m \in \mathbb{N}$. For $m = 0$ we use (R_2). For the induction step we use (R_3) and the functionality of appl.

(*ii*) By induction on $m \in \mathbb{N}$ we show $\forall v \mathsf{T}\ulcorner\varphi[m^\bullet, v^\bullet] \dot{\leftrightarrow} \psi[m^\bullet, v^\bullet]\urcorner$.

If $m = 0$ $\mathsf{T}\ulcorner\varphi[\overline{0}, v^\bullet]\urcorner \leftrightarrow \mathsf{T}\ulcorner v^\bullet \doteq a\urcorner \leftrightarrow \mathsf{T}\ulcorner\psi[\overline{0}, v^\bullet]\urcorner$.

For $m = \mathsf{S}(k)$ $\mathsf{T}\ulcorner\varphi[(\mathsf{S}(k))^\bullet, v^\bullet]\urcorner \leftrightarrow \exists w(\mathsf{T}\ulcorner v^\bullet \doteq \mathsf{appl}(h, w)\urcorner \wedge \mathsf{T}\ulcorner\varphi[m^\bullet, w]\urcorner) \overset{IH,(i)}{\leftrightarrow}$
$\exists w(\mathsf{T}\ulcorner v^\bullet \doteq \mathsf{appl}(h, w)\urcorner \wedge \mathsf{T}\ulcorner\psi[m^\bullet, w]\urcorner) \leftrightarrow \mathsf{T}\ulcorner\psi[(\mathsf{S}(k))^\bullet, v^\bullet]\urcorner$.

(*iii*) We argue by induction on n in $\mathbb{N}$.

For $n = 0$ we use $\ulcorner\varphi\urcorner = n^\bullet \doteq \overline{0} \mathbin{\dot{\wedge}} v^\bullet \doteq a$.

For $n = \mathsf{S}(k)$ we argue by adjunction. By induction hypothesis and (i)

$\exists \ulcorner\psi\urcorner(\mathsf{Rec}(\psi, h, a, k) \wedge \exists! w \mathsf{T}\ulcorner\psi[k^\bullet, w^\bullet]\urcorner$. Then for

$\varphi(x, y) \leftrightarrow \psi(x, y) \vee (x = \mathsf{S}(k) \wedge y = \mathsf{val}(\mathsf{appl}(h, w^\bullet))$, then

$\mathsf{Rec}(\ulcorner\psi\urcorner, h, a, \mathsf{S}(k))$.

(iv) A direct argument suffices;

(v) We argue by induction on n. $n = 0$ is simple.

For $n = \mathsf{S}(k)$ let

$\mathsf{Rec}(\ulcorner\varphi\urcorner, h, a, \mathsf{S}(k)) \wedge \mathsf{Rec}(\ulcorner\psi\urcorner, h, a, m) \wedge \mathsf{S}(k) \leq m$.

Then we can define $\ulcorner\varphi^-\urcorner$, such that

$\mathsf{T}\ulcorner\varphi^-[l^\bullet, w^\bullet]\urcorner \leftrightarrow \mathsf{T}\ulcorner\varphi[l^\bullet, w^\bullet] \mathbin{\dot\wedge} \dot\neg\, (l^\bullet \mathbin{\dot=} \dot{\mathsf{S}}\, k^\bullet)\urcorner$, then by (iv) $\mathsf{Rec}(\ulcorner\varphi^-\urcorner, h, a, k)$ and $k < m$, so we can use the induction hypothesis to argue that

$\mathsf{T}\ulcorner\varphi^-\urcorner \to \mathsf{T}\ulcorner\psi\urcorner$. Now if $\mathsf{T}\ulcorner\varphi[\dot{\mathsf{S}}\, k^\bullet, w^\bullet]\urcorner$ for some $w \in P$, then by the assumption that $\mathsf{Rec}(\ulcorner\psi\urcorner, h, a, m)$ and $\mathsf{S}(k) \leq m$ and the uniqueness in (i) we have $\mathsf{T}\ulcorner\psi[\dot{\mathsf{S}}\, k^\bullet, w^\bullet]\urcorner$.

□

Using the properties of Lemma 2 we can show that

$$\forall n \exists! u \exists \ulcorner\varphi\urcorner\, (\mathsf{Rec}(\ulcorner\varphi\urcorner, h, a, n) \wedge \mathsf{T}\ulcorner\varphi[n^\bullet, u^\bullet]\urcorner) \tag{!}$$

We follow Feferman and Hellman and define our function f satisfying (*) by the following formula

$$f(n, u) :\leftrightarrow \exists \ulcorner\varphi\urcorner\, (\mathsf{Rec}(\ulcorner\varphi\urcorner, h, a, n) \wedge \mathsf{T}\ulcorner\varphi[n^\bullet, u^\bullet]\urcorner)$$

With this we have established the existence of a function $f : \mathbb{N} \to P$, such that f satisfies (*), namely $f(0) = a$ and $\forall n(f(\mathsf{S}(n)) = \mathsf{appl}(h, f(n)))$.

Now we can establish categoricity for the relevant notion of isomorphism. We let $\mathsf{ISO}_{\mathbb{N} \triangleright P}(f, (\overline{P}, h, a))$ be short for

$$\forall n \forall v(f(n, v) \to (\mathbb{N}(n) \wedge P(v))) \wedge \qquad \text{(iso:1)}$$

$$\forall n \exists! u\, f(n, u) \wedge \qquad \text{(iso:2)}$$

$$\forall v \exists! m\, f(v, m) \wedge \qquad \text{(iso:3)}$$

$$f(0, \mathsf{val}(a)) \wedge \forall n, u(f(n, u) \to f(\mathsf{S}(n), \mathsf{val}(\mathsf{appl}(h, u^\bullet)))) \qquad \text{(iso:4)}$$

Then we can establish the isomorphism in a related fashion to Feferman and Hellman.[20]

Proposition 9.

$$\mathbf{CT}[P] \vdash (\mathsf{PS}^P(\overline{P}, h, a) \to \mathsf{ISO}_{\mathbb{N} \triangleright P}(f, (\overline{P}, h, a)))$$

Proof. We have (iso:1) by definition of Rec. By (!) we have (iso:2) and therefore, to simplify the presentation, we work in the following with f as a function symbol. By (R_3) we also have (iso:4).

What remains to be shown is that f is one-to-one and onto, (iso:3). For the former we use induction on n in $\mathbb{N}$ to show $f(n) = f(m) \to n = m$. For $n = 0$ we use (R_2) to establish that $v = \mathsf{val}(a))$. If $m \neq 0$, then there is some $k \in \mathbb{N}$ with $m = \mathsf{S}(k)$ and $f(m) = f(\mathsf{S}(k))$. But then for some $u \in P$, by (N-1) $v = \mathsf{val}(\mathsf{appl}(h, u^\bullet)) \neq \mathsf{val}(a)$.

For $n = \mathsf{S}(k)$ we show that $\forall m(f(\mathsf{S}k) = f(m) \to \mathsf{S}(k) = m)$. But by (*) $f(\mathsf{S}(k)) = \mathsf{appl}(h, f(k))$ and so we show that $\forall m(\mathsf{appl}(h, f(k)) = f(m) \to \mathsf{S}(k) =$

[20] Compare (Feferman and Hellman, 1995, Theorem 5.).

m). For $m = 0$ we directly get a contradiction $\mathsf{S}(k) = 0$. For $m = \mathsf{S}(l)$ we have $\mathsf{appl}(h, f(k)) = \mathsf{appl}(h, f(l))$ and so $f(k) = f(l)$ by (N-2) and by induction hypothesis $k = l$ and so $\mathsf{S}(k) = \mathsf{S}(l) = m$.

For the surjectivity we make use of (N-3) on P. The idea is to use it on the subset X of P, such that $X = \{u \in P \mid \exists n f(n) = u\}$. In order to apply (N-3) we must make sure that there is a formula φ, such that $\mathsf{T}^{\ulcorner}\varphi^{\urcorner}$ defines X. We have $v \in X$ iff $P(v) \wedge \exists n(f(n) = v)$, which is by definition of f equivalent to $P(v) \wedge \exists n \exists^{\ulcorner}\varphi^{\urcorner}(\mathsf{Rec}(^{\ulcorner}\varphi^{\urcorner}, h, a, n) \wedge \mathsf{T}^{\ulcorner}\varphi[n^{\bullet}, v^{\bullet}]^{\urcorner})$. Then we can use the $^{\ulcorner}\psi^{\urcorner}$ from Lemma 1 to reformulate it as $P(v) \wedge \exists n \exists^{\ulcorner}\varphi^{\urcorner}(\mathsf{T}^{\ulcorner}\psi[^{\ulcorner}\varphi^{\urcorner}, h, a, n^{\bullet}]^{\urcorner} \wedge \mathsf{T}^{\ulcorner}\varphi[n^{\bullet}, v^{\bullet}]^{\urcorner}))$. Then we can use the expanded T-biconditionals for P and the commutation axioms to see that there is a term χ, such that $\mathsf{T}^{\ulcorner}\chi[v^{\bullet}]^{\urcorner}$ is extensionally equivalent to the previous (with the parameters h, a hidden).

Now we use (N-3) to show: $\mathsf{T}^{\ulcorner}\chi[a]^{\urcorner} \wedge \forall w(\mathsf{T}^{\ulcorner}\chi[w^{\bullet}]^{\urcorner} \rightarrow \mathsf{T}^{\ulcorner}\chi[\mathsf{appl}(h, w^{\bullet})]^{\urcorner})$. The a case is obvious. In the successor case we assume that for some w there is some n such that $\mathsf{T}^{\ulcorner}\chi[n^{\bullet}, w^{\bullet}]^{\urcorner}$. By definition of f, $f(\mathsf{S}n) = \mathsf{appl}(h, f(n))$ and so $\mathsf{T}^{\ulcorner}\chi[\dot{\mathsf{S}}\, n^{\bullet}, \mathsf{appl}(h, w^{\bullet})]^{\urcorner}$. □

6.5 Reconsidering Determinacy

This section reconsiders the issue of how and to what degree our results provide some form of determinacy of arithmetical truth. First, it discusses Theorem 8 and its relevance for the claim that arithmetical truth is determinate and a possible objection.

Theorem 8 establishes that the internal isomorphism between arithmetical structures can be lifted to our primitive truth predicates. Button and Walsh argue that

arithmetical intolerance motivates the acceptance of a *univocal* theory of arithmetic, and we follow this line of argument by claiming that Theorem 8, as a form of truth-theoretic intolerance, motivates the acceptance of a univocal theory of truth. This is *prima facie* in tension with a result by Hamkins and Yang (2013). They claim that "the definiteness of the theory of truth for a structure does not follow as a consequence of the definiteness of the structure in which that truth resides." (Hamkins and Yang, 2013), p.26. So even if the natural number structure $\langle \mathbb{N}, +, \cdot, 0, 1, < \rangle$ is definite in the sense (exemplified in their theorem) that two **ZFC**-models agree on the interpretation of the arithmetical vocabulary, this does not imply that arithmetical truth is determinate, insofar as there is 'arithmetical' statement σ, such that the two models evaluate it by assigning two different truth values to σ. This philosophical conclusion is based on their Theorem in (Hamkins and Yang, 2013), p.5:

Theorem 17 (Hamkins and Yang)**.** *Every consistent extension of* **ZFC** *has two models M_1 and M_2, which agree on the natural numbers and on the structure $\langle \mathbb{N}, +, \cdot, 0, 1, < \rangle^{M_1} = \langle \mathbb{N}, +, \cdot, 0, 1, < \rangle^{M_2}$, but which disagree on their theories of arithmetic truth, in the sense that there is in M_1 and M_2 an arithmetic sentence σ, such that M_1 thinks σ is true, but M_2 thinks it is false.*

The main strategy of their first proof exploits an observation due Krajewski (1974) that shows that for a nonstandard model of arithmetic that admits a full satisfaction class there is an elementary extension that admits two incompatible full satisfaction classes. Hamkins and Yang show that these arithmetical models can be understood as **ZFC**-standard models.[21] The full inductive satisfaction classes closely correspond to the theory **CT**, which allows a more or less direct interpretation. Arithmetical

[21] For the terminology see also (Enayat, 2014).

models of **PA** admitting a full inductive satisfaction class can be characterised as models of **CT**.

From a model-theoretic perspective the theorem shows that we can expand a **PA**-model $\mathfrak{A}$ with two satisfaction classes S and S' in an incompatible way, i.e., such that $(\mathfrak{A}, S)$ and $(\mathfrak{A}, S')$ are models of **CT** and disagree about the truth of an arbitrary arithmetical statement σ, such that $\sigma \in S$ and $\sigma \notin S'$. They draw the following philosophical conclusion:

> [T]he definiteness of the theory of truth for a structure does not follow as a consequence of the definiteness of the structure in which that truth resides. Even in the case of arithmetic truth and the standard model of arithmetic $\mathbb{N}$, we claim, it is a *philosophical error* to deduce that arithmetic truth is definite just on the basis that the natural numbers themselves and the natural number structure $\langle \mathbb{N}, +, \times, 0, 1, < \rangle$ is definite. At bottom, our claim is that one does not get definiteness-of-truth *for free* from definiteness-of objects and definiteness-of-structure. [our emphasis] (Hamkins and Yang, 2013, p. 26)

The first pointed – emphasised in the previous quote – concerns the claim that it is a philosophical erroneous inference to conclude the determinacy of truth from the definiteness of the structure. Theorem 17 is beyond doubt and supports Hamkins and Yang's philosophical conclusion at least for an external understanding of determinatecy and of structures. However, from an internalist understanding of determinacy, their technical result does not support an equivalent philosophical conclusion. As our result shows, from an internalist perspective, the acceptance of axiomatic truth-theoretic principles, that allow for some form of bridging, does imply determi-

nacy of truth – in the form of internalised intolerance exemplified by Proposition 8. In our case, the scenario sketched by Hamkins and Yang is excluded by expanding induction to include statements of the mixed vocabulary. By doing so we make sure that the two **CT**-models agree on their respective arithmetical truths.[22]

Hamkins' and Yang's philosophical argument presupposes expressive resources and distinctions that are not directly available in an internalist conception. For example the relevant counterexample σ is a nonstandard sentence and also the models are non-standard models of arithmetic and set theory. For the internalist these fine-grained distinctions are not within the range of interpretations that she is able to discriminate. Parsons and others have convincingly argued that such distinctions are not within the reach of internalism.[23] For the internalist the models are not given as something external existing independently. The internalist can only make sense of these models via descriptions of the model, as given in the object language. But once the nonstandard model is given by a description, via a suitable language expansion for example, the internalist cannot only recognize the interpretation, but also conceive its inadequacy as a relevant alternative anymore. Parsons recognises that no strong notion of semantic determinacy, nor a notion of standardness are consequences of internal categoricity:

[22]We should point out that Hamkins and Yang do not directly argue against internalism. Their argument is employed in the context of Feferman's philosophy of mathematics and within his claim that the definiteness of the natural numbers implies the determinacy of arithmetical truth. For an analysis of Feferman's argument the consequences are dependent on an interpretation of Feferman's conceptual structuralism. It would be interesting to investigate whether Feferman's conceptual structuralism can be understood as a form of internalism. In that case, Theorem 17 might not undermine Feferman's claim.

[23]Compare Parsons' position spelled out in (Parsons, 2008, p. 288) but also the discussion in (Button and Walsh, 2018), p.279 f.

> No proof has been given that the "intended model" [...] is the standard one. Although [...] the [target arithmetical structures] are isomorphic, it does not follow that they are standard. [...] We can see that two purported number sequences are isomorphic [...] but we cannot in the end get away from the fact that the result obtained is one "within mathematics" (in Wittgenstein's phrase). [...] [Internal categoricity] does not protect the language of arithmetic from an interpretation completely from outside, that takes quantifiers over numbers as ranging over a nonstandard model. One might imagine a God who constructs such an interpretation, and with whom dialogue is impossible. But so far the interpretation is, in the Kantian phrase, "nothing to us." (Parsons, 2008, pp. 287-288)

Concerning the second point, we admit that internal determinacy is not 'for free'. We do assume that a suitable expansion of the range of a truth predicate to an arbitrary property P is possible. However, we think that it is possible to argue for the plausibility of our assumptions as background assumptions for suitable agents. At least our assumptions appear less problematic than the assumptions of the reverse mathematics case that P is always a subset of our natural numbers.

6.6 Conclusion

Let us take stock: with the expressive resources provided by the primitive truth predicate we argued for a univocal theory of truth. Proposition 9 circumvents some of the expressive limitations of the first-order categoricity theorem. In contrast to the second-order version we take one internal structure as given and consider possi-

ble alternatives. Our approach is asymmetric and closer to the reverse mathematics version developed by Simpson and Yokoyama (2013). We think that our asymmetric and schematic versions have some advantages. Moreover, this approach should also be of interest to anyone sceptical about some of the assumptions or presuppositions involved in the pure second-order version. For instance, one could be reluctant to accept full, impredicative **CA** as a logical principle. But even if one accepts **CA** as logic, there is still the question of whether and to what degree such principles are acceptable. For instance, one might be interested in a general, predicatively provable, version of categoricity. Proposition 9 provides one such version. Moreover, abandoning the version in pure second-order logic might be advantageous when considering the epistemological issue of agreement mentioned in the introduction of Part III. Parsons' Equivalence Thesis claimed that any two agents accepting arithmetic must agree that their respective theories are equivalence. However, employing **CA** as the common ground between the agents might trivialise the Equivalence Thesis thereby undermining its epistemological significance. Let us conclude with a further issue for future work: one might wonder whether introducing a truth predicate is compatible with internalism. For the moment, there is *prima facie* no reason to reject the use of a primitive truth predicate: this seems to be not only in line with a Davidsonian conception, but also with a deflationary conception of truth. On a deflationary conception the main purpose of the truth predicate is expressive and this seems to be in line with Parsons' internalist understanding of language as used prior to semantic reflection on it. With this, the conception of truth employed here is supposed to be different in spirit from a model-theoretic conception of truth, as truth-in-a-structure. Although a full evaluation of the truth-theoretic approach is

not possible at this early stage we hope to have provided some first steps towards a more attractive and plausible picture of internalism.

Summary and Conclusions

This dissertation investigated the epistemology of different meta-theoretic properties of (mathematical) theories. These properties play an essential role in mathematical inquiries and are of great importance for the epistemology and philosophy of mathematics. What follows provides a summary of this dissertation's claims and results. After that, we draw some exploratory and speculative conclusions and morals concerning this dissertation's approach.

Part I of this dissertation investigated the epistemology of consistency of mathematical theories within the context of Wright's cognitive projects. Chapter 1 investigated the question of what justification agents have – if they have any – to believe that the target theories employed in *epistemic foundational projects* are consistent and argued that agents are *entitled* to believe in the consistency of such theories. Moreover, we argued that entitlement is not simply a permission but an *obligation* to believe that the target theories are consistent. Chapter 2 focused on whether, for some entitlement-based epistemology, it is coherent to endorse the thesis that entitled beliefs do not constitute knowledge. This chapter argued that such thesis is coherent and acceptable in its own right. Chapter 3 continued this investigation into the epistemology of consistency and analysed so-called soundness arguments

for consistency. We provided a diagnosis of why and to what extent such arguments can be evaluated as *epistemically defective*. Part II investigated the epistemological role played by reflection principles in the context of cognitive projects. After an introduction to the main results and issues concerning reflection principles in chapter 4, chapter 5 focused on the role of *global reflection principles* in expressing the *trustworthiness* of theories (of truth). More precisely, the chapter provided a cluster of theories of truth in classical logic that are trustworthy. Moreover, we argued that trustworthy theories are should be preferred. Finally, Part III focused on the notion of internal categoricity, particularly the connection between internal categoricity theorems and determinacy of truth. Chapter 6 provided a truth-theoretic version of internal categoricity and internal determinacy, improving on Parsons-style approaches to categoricity. Although these results are important in their own right, the investigation in chapter 6 constituted essential, preliminary work needed to discuss epistemological issues related to internal categoricity. This dissertation's results and claims helped us make significant progress concerning important issues in the epistemology of mathematics.

Nevertheless, additional, important philosophical issues arising from this dissertation need investigation in the future. As pointed out in the summary of chapter 6, one immediate issue to be investigated in the future is Parsons' equivalence thesis. Concerning the non-evidentialist epistemology investigated in chapters 1 and 2: one future investigation is the interaction and dependence between entitled and evidentially warranted beliefs. This will amount to investigating – *inter alia* – Conservativism and Liberalism (presented in chapter 3) in more depth. Building on considerations in chapters 1 and 5, an additional focus for future work is the role

played by epistemic norms in regulating our mathematical inquiries. Investigating these and other issues arising from this thesis is going to be essential to make substantial progress in (the) epistemology (of mathematics).

The progress made by this dissertation is a result of – *inter alia* – its methodological approach to the epistemology of mathematics. Roughly, that approach consisted in – to put it succinctly – taking issues in the epistemology of mathematics *seriously*: this approach has been committed (and still is!) to the idea that concepts and frameworks from epistemology are needed to make progress concerning fundamental issues in the epistemology of mathematics. This methodological view takes a different stance on the issue of the *authority* of mathematics and epistemology. It aims to vindicate the authority of epistemology when it comes to epistemological issues in the philosophy of mathematics. However, this approach does not want epistemology to ignore mathematics or mathematical results. Mathematical results and considerations should *inform* the epistemology of mathematics. Epistemologists should consider and weigh evidence provided by mathematics and acknowledge the *authority* of mathematics when it comes to mathematical issues. Finally, this approach wishes to develop a new and more fruitful dialogue between epistemology and mathematics, retaining the authority of both disciplines. The investigation of this methodology and how it relates and connects to others – possibly rival – methodological positions[24] constitutes a further, important, *metaphilosophical* question for future research.

[24]One alternative methodological position that naturally comes to mind is *mathematical naturalism* or *Second Philosophy*, as introduced by Maddy (2007).

Bibliography

Alspector-Kelly, M. (2019). *Against knowledge closure*. Cambridge University Press.

Boolos, G. S. (1993). *The Logic of Provability*. Cambridge University Press.

Boult, C. (2020). There is a distinctively epistemic kind of blame. *Philosophy and Phenomenological Research*, 103(3):518–534.

Boult, C. (2021a). Epistemic blame. *Philosophy Compass*, 16(8).

Boult, C. (2021b). The significance of epistemic blame. *Erkenntnis*.

Brown, J. (2018). What is epistemic blame? *Noûs*, 54(2):389–407.

Brown, J. (2019). Epistemically blameworthy belief. *Philosophical Studies*, 177(12):3595–3614.

Burge, T. (2003). Perceptual entitlement*. *Philosophy and Phenomenological Research*, 67(3):503–548.

Burge, T. (2020). Entitlement: The basis for empirical epistemic warrant. In Pedersen, N. and Graham, P., editors, *Epistemic Entitlement*, pages 37–142. Oxford University Press.

Button, T. (2022). Mathematical internal realism. In Conant, J. and Chakraborty, S., editors, *Engaging Putnam*, pages 157–182. De Gruyter.

Button, T. and Walsh, S. (2016). Structure and Categoricity: Determinacy of reference and truth value in the philosophy of mathematics. *Philosophia Mathematica*, 24(3):283 – 307.

Button, T. and Walsh, S. (2018). *Philosophy and Model Theory*. Oxford University Press.

Cantini, A. (1989). Notes on formal theories of truth. *Zeitschrift für mathematische Logik und Grundlagen der Mathematik*, 35:97–130.

Cantini, A. (1990). A theory of formal truth arithmetically equivalent to ID_1. *The Journal of Symbolic Logic*, 55:244–259.

Carter, J. A. (2016). Meta-epistemic defeat. *Synthese*, 195(7):2877–2896.

Chandler, J. (2009). The transmission of support: a bayesian re-analysis. *Synthese*, 176(3):333–343.

Cieśliński, C. (2010). Truth, conservativeness, and provability. *Mind*, 119:409–422.

Cieśliński, C. (2015). Typed and untyped disquotational truth. In Achourioti, T., Galinon, H., Fernández, J. M., and Fujimoto, K., editors, *Unifying the Philosophy of Truth. Logic, Epistemology, and the Unity of Science*. Springer.

Cieśliński, C. (2017a). *Believability theories. Corrigendum to: The Epistemic Lightness of Truth, Deflationism and its Logic*. Cambridge University Press.

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

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

Cohen, L. J. (1992). *An Essay on Belief and Acceptance*. New York: Clarendon Press.

Coliva, A. (2008). Moore's proof, liberals, conservatives–is there a (Wittgensteinian) third way? In Coliva, A., editor, *Mind, Meaning and Knowledge. Themes from the philosophy of Crispin Wright*. Oxford University Press.

Coliva, A. (2015). *Extended Rationality*. Palgrave Macmillan UK.

Damnjanovic, Z. (2017). Mutual interpretability of robinson arithmetic and adjunctive set theory with extensionality. *The Bulletin of Symbolic Logic*, 23(4):381–404.

Davis, J. K. (2014). Faultless disagreement, cognitive command, and epistemic peers. *Synthese*, 192(1):1–24.

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

Dretske, F. (2005). Is knowledge closed under known entailment? the case against closure. In Steup, M. and Sosa, E., editors, *Contemporary Debates in Epistemology*, pages 13–26. Blackwell.

Dretske, F. (2006). Information and closure. *Erkenntnis*, 64(3):409–413.

Dummett, M. (1978). *Truth and other Enigmas*. London.

Enayat, A. (2014). Standard models of arithmetic. In Kaså, M., editor, *Idées Fixes: A Festschrift Dedicated to Christian Bennet on the Occasion of His 60th Birth-*

day, pages 55–64. Department of Philosophy, Linguistics and Theory of Science, University of Gothenburg.

Engel, P. (1998). Believing, holding true, and accepting. *Philosophical Explorations*, 1(2):140–151.

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

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

Feferman, S. (1991). Reflecting on incompleteness. *The Journal of Symbolic Logic*, 56:1–47.

Feferman, S. (2013). Categoricity and open-ended axiom systems. slides (online).

Feferman, S. (2014). Logic, mathematics and conceptual structuralism. In Rush, P., editor, *The Metaphysics of Logic*, pages 72–92. Cambridge University Press.

Feferman, S. and Hellman, G. (1995). Predicative foundations of arithmetic. *Journal of philosophical logic*.

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

Feldman, R. (1995). In defence of closure. *The Philosophical Quarterly*, 45(181):487.

Field, H. (1994a). Deflationist views of meaning and content. *Mind*, 103(411):249–285.

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

Field, H. (2001). Correspondence truth, disquotational truth, and deflationism. In Lynch, M. P., editor, *The Nature of Truth: Classical and Contemporary Perspectives*, pages 483–503. Cambridge.

Field, H. (2022). The power of naive truth. *The Review of Symbolic Logic*, 15(1):225–258.

Fischer, M. (2021). Another look at reflection. *Erkenntnis*.

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

Fischer, M., Horsten, L., and Nicolai, C. (2019). Hypatia's silence truth justification and entitlement. *Noûs*. online first doi: 10.1111/nous.12292.

Fischer, M., Nicolai, C., and Dopico, P. (2021). Nonclassical truth with classical strength. A proof-theoretic analysis of compositional truth over hype. *The Review of Symbolic Logic*, pages 1–24.

Fischer, M. and Zicchetti, M. (2022). Truth, categoricity and determinateness. *under review*.

Franzén, T. (2004a). *Inexhaustibility*. Lecture Notes in Logic. CRC Press.

Franzén, T. (2004b). Transfinite progressions: a second look at completeness. *The Bulletin of Symbolic Logic*, 10(3):367 –89.

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

Friedman, H. and Sheard, M. (1988). The disjunction and existence properties for axiomatic systems of truth. *Annals of Pure and Applied Logic*, 40(1):1–10.

Fujimoto, K. (2017). Deflationism beyond arithmetic. *Synthese*, 196(3):1045–1069.

Fujimoto, K. (2021). The function of truth and the conservativeness argument. *Mind*, 131(521):129–157.

Galinon, H. (2014). Acceptation, cohérence et responsabilité. In *Liber Amicorum Pascal Engel*. Geneva: University of Geneva.

Girard, J.-Y. (1987). *Proof Theory and Logical Complexity*. Neapel.

Grzegorczyk, A. (1962). On the concept of categoricity. *Studia Logica: An International Journal for Symbolic Logic*, 13:39–66.

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

Halbach, V. (2009). Reducing compositional to disquotationl truth. *The Review of Symbolic Logic*, 2(4):786–798.

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

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

Halbach, V. and Nicolai, C. (2017). On the costs of non-classical logic. *Journal of Philosophical Logic*. Online first https://link.springer.com/article/10.1007/s10992-017-9424- 3.

Hale, B. and Wright, C. (2005). Logicism in the twenty-first century. In Shapiro, S., editor, *The Oxford Handbook of Philosophy of Mathematics and Logic*, pages 125–134. New York.

Hamkins, J. D. and Yang, R. (2013). Satisfaction is not absolute. *arXiv e-prints*, page arXiv:1312.0670.

Hawthorne, J. (2004). *Knowledge and Lotteries*. Oxford University Press.

Heck, Jr, R. G. (2015). Consistency and the theory of truth. *Rev. Symb. Log.*, 8(3):424–466.

Henderson, D. (2020). Are epistemic norms fundamentally social norms? *Episteme*, 17(3):281–300.

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

Horsten, L. (2021). On reflection. *The Philosophical Quarterly*, 71(4).

Horsten, L. and Halbach, V. (2015). Norms for theories of reflexive truth. In Fujimoto, K., Fernandez, J. M., Galinon, H., and Achourioti, T., editors, *Unifying the Philosphy of Truth*, volume Unifying the Philosophy of Truth. Springer Verlag.

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

Horsten, L., Leigh, G. E., Leitgeb, H., and Welch, P. (2012). Revision revisited. *The Review of Symbolic Logic*, 5(4):642–664.

Horsten, L. and Zicchetti, M. (2021). Truth, reflection and commitment. In Stern, J. and Nicolai, C., editors, *Modes of Truth. The Unified Approach to Truth, Modalities, and Paradox*. Routledge.

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

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

Isaacson, D. (2011). The reality of mathematics and the case of set theory. In Novák, Z. and Simonyi, A., editors, *Truth, Reference, and Realism*, pages 1–75. Central European University Press.

Jenkins, C. (2007). Entitlement and rationality. *Synthese*, 157(1):25–45.

Kauppinen, A. (2018). Epistemic norms and epistemic accountability. *Philosophers' Imprint*, 18.

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

Ketland, J. (2005). Deflationism and the Gödel phenomena: Reply to Tennant. *Mind*, 114(453):75–88.

Ketland, J. (2010). Truth, Conservativeness, and Provability: Reply to Cieśliński. *Mind*, 119(474):423–436.

Kline, M. (1982). *Mathematics: The loss of certainty*, volume 686. Galaxy Books.

Kölbel, M. (2004). III-faultless disagreement. *Proceedings of the Aristotelian Society (Hardback)*, 104(1):53–73.

Krajewski, S. (1974). Mutually inconsistent satisfaction classes. *Bull. Acad. Polon. Sci. Sér. Sci. Math. Astronom. Phys*, 22:983–987.

Kreisel, G. (1960). La prédicativité. *Bulletin de la Société Mathématique de France*, pages 371–391.

Kreisel, G. and Lévy, A. (1968). Reflection principles and their use for establishing the complexity of axiomatic systems. *Zeitschrift für mathematische Logik und Grundlagen der Mathematik*, 14:97–142.

Kvanvig, J. L. (1999). Truth and superassertibility. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, 93(1):1–19.

Leigh, G. E. (2016). Reflecting on truth. *IfCoLog Journal of Logics and their Applications*, 3(4):557–594.

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

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

Łełyk, M. (2022). Model theory and proof theory of the global reflection principle. *The Journal of Symbolic Logic*, pages 1–42.

Łełyk, M. and Nicolai, C. (2022). A theory of implicit commitment. *Synthese*, 200(4).

Luper, S. (2020). Epistemic Closure. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2020 edition.

Lynch, M. (2010). Epistemic circularity and epistemic incommensurability. In A. Haddock, A. Millar, D. P., editor, *Social Epistemology*, pages 262 – 77. Oxford University Press.

Maddy, P. (2007). *Second Philosophy: A Naturalistic Method*. Oxford University Press.

Maddy, P. and Väänänen, J. (2022). Philosophical uses of categoricity arguments.

Martin, B. (2019). Searching for deep disagreement in logic: The case of dialetheism. *Topoi*, 40(5):1127–1138.

McGee, V. (2005b). Afterword: Trying (with limited success) to demarcate the disquotation-correspondence intuition. In Beall, J. and Armour-Garb, B., editors, *Deflationary Truth*. Open Court.

Moore, G. E. (1939). Proof of an external world. *Proceedings of the British Academy*, 25(5):273 – 300.

Moretti, L. (2010). Wright, okasha and chandler on transmission failure. *Synthese*, 184(3):217–234.

Moretti, L. and Piazza, T. (2017). Defeaters in current epistemology: introduction to the special issue. *Synthese*, 195(7):2845–2854.

Moretti, L. and Piazza, T. (2018). Transmission of Justification and Warrant. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2018 edition.

Mount, B. M. and Waxman, D. (2021). Stable and unstable theories of truth and syntax. *Mind*.

Murzi, J. and Rossi, L. (2018). Conservative deflationism? *Philosophical Studies*, 177(2):535–549.

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

Neta, R. (2007). Fixing the transmission: the new Mooreans. In Nuccetelli, S. and Seay, G., editors, *Themes From G. E. Moore: New Essays in Epistemology and Ethics*. Clarendon Press.

Neta, R. (2010). Liberalism and Conservatism in the epistemology of perceptual belief. *Australasian Journal of Philosophy*, 88(4):685–705.

Nicolai, C. (2015). Deflationary truth and the ontology of expressions. *Synthese*, 192(12):4031–4055.

Nicolai, C. (2021). Fix, express, quantify: Disquotation after its logic. *Mind*, 130(519):727–757.

Nicolai, C. and Piazza, M. (2018). The implicit commitment thesis of arithmetical theories and its semantic core. *Erkenntnis*.

Nottelmann, N. (2007). *Blameworthy Belief: A Study in Epistemic Deontologism*. Dordrecht: Springer.

Okasha, S. (2004). Wright on the transmission of support: A bayesian analysis. *Analysis*, 64(2):139–146.

Parsons, C. (1990). The uniqueness of the natural numbers. *Iyyun: The Jerusalem Philosophical Quarterly*, 39:13 – 44.

Parsons, C. (2008). *Mathematical Thought and its Objects*. Cambridge University Press.

Pedersen, N. and Graham, P. (2020). Recent work on epistemic entitlement. *American Philosophical Quarterly*, 57(2):193–214.

Pedersen, N., Graham, P., Bachman, Z., and Rosa, L. (2020). Introduction and overview: Two entitlement projects. In Pedersen, N. and Graham, P., editors, *Epistemic Entitlement*, pages 1–34. Oxford University Press.

Pedersen, N. J. (2008). Entitlement, value and rationality. *Synthese*, 171(3):443–457.

Pedersen, N. J. L. L. (2016). Hume's principle and entitlement: On the epistemology of the neo-fregean programme. In Ebert, P. and Rossberg, M., editors, *Abstractionism*. Oxford University Press.

Pedersen, N. J. L. L. (2021). Cornerstone epistemology: scepticism, mathematics, non-evidentialism, consequentialism, pluralism. In Pedersen, N. J. L. L. and Moretti, L., editors, *Non-Evidentialist epistemology*. Brill.

Pedersen, N. J. L. L. (2022). Entitlement, generosity, relativism, and structure-internal goods. *Metaphilosophy*.

Piazza, M. and Pulcini, G. (2013). Strange case of Dr. soundness and Mr. consistency. In *In Logica yearbook*, pages 161–172. College Publications.

Picollo, L. and Schindler, T. (2021). Is deflationism compatible with compositional and tarskian truth theories? In Stern, J. and Nicolai, C., editors, *Modes of Truth. The Unified Approach to Truth, Modalities, and Paradox*, pages 41 – 68. Routledge.

Pollock, J. (1970). The structure of epistemic justification. *American Philosophical Quarterly*, 4:62–78.

Pollock, J. L. (1987). Epistemic norms. *Synthese*, 71(1):61–95.

Priest, G. (2008). *An Introduction to Non-classical Logic From if to is*. Cambridge, 2nd edition.

Pritchard, D. (2016). Anti-luck virtue epistemology and epistemic defeat. *Synthese*, 195(7):3065–3077.

Pryor, J. (2000). The skeptic and the dogmatist. *Noûs*, 34(4):517–549.

Pryor, J. (2004). What's wrong with Moore's argument. *Philosophical Issues*, 14:349 – 77.

Ranalli, C. (2018). Deep disagreement and hinge epistemology. *Synthese*, 197(11):4975–5007.

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

Rettler, L. (2017). In defense of doxastic blame. *Synthese*, 195(5):2205–2226.

Roland, J. W. (2007). Maddy and mathematics: Naturalism or not. *The British Journal for the Philosophy of Science*, 58(3):423–450.

Schmidt, S. (2021). Epistemic blame and the normativity of evidence. *Erkenntnis*.

Schütte, K. (1964). Eine grenze für die beweisbarkeit der transfiniten induktion in der verzweigten typenlogik. *Archiv für Mathematische Logik und Grundlagenforschung*, 7:45–60.

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

Shapiro, S. (1991). *Foundations without Foundationalism*. Clarendon Press, Oxford.

Shapiro, S. (1997). *Philosophy of Mathematics. Structure and Ontology*. Oxford University Press.

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

Shapiro, S. (2004). Foundations of mathematics: Metaphysics, epistemology, structure. *The Philosophical Quarterly*, 54(214):16–37.

Shapiro, S. (2011). Epistemology of Mathematics: What are the questions? What count as answers? *The Philosophical Quarterly*, 61(242):130–150.

Silins, N. (2005). Transmission failure failure. *Philosophical Studies*, 126(1):71–102.

Silins, N. (2007). Basic justification and the moorean response to the skeptic. In Gendler, T. S. and Hawthorne, J., editors, *Oxford Studies in Epistemology Volume 2*, pages 108–142. Oxford University Press.

Simpson, S. G. and Yokoyama, K. (2013). Reverse mathematics and Peano categoricity. *Annals of Pure and Applied Logic*, 164:284–293.

Smith, M. (2013). Entitlement and evidence. *Australasian Journal of Philosophy*, 91(4):735–753.

Smith, M. (2020). Full blooded entitlement. In Pedersen, N. and Graham, P., editors, *Epistemic Entitlement*, pages 281–296. Oxford University Press.

Smith, P. S. and Lynch, M. P. (2020). Varieties of deep epistemic disagreement. *Topoi*, 40(5):971–982.

Tarski, A. (1936). Über den Begriff der logischen Folgerung. *Actes du Congrès International de Philosophie Scientifique*, 7:1–11.

Tennant, N. (1995). On negation, truth and warranted assertibility. *Analysis*, 55(2):98–104.

Tennant, N. (2002). Deflationism and the Gödel Phenomena. *Mind*, 111(443):551–582.

Tennant, N. (2005). Deflationism and the Gödel phenomena: Reply to Ketland. *Mind*, 114(453):89–96.

Tennant, N. (2010). Deflationism and the Gödel Phenomena: Reply to Cieśliński. *Mind*, 119(474):437–450.

Tucker, C. (2010). When transmission fails. *Philosophical Review*, 119(4):497–529.

Turing, A. (1939). Systems of logic based on ordinals. *Proceedings of the London Mathematical Society*, 2:161–228.

Turri, J. (2010). On the relationship between propositional and doxastic justification. *Philosophy and Phenomenological Research*, 80(2):312–326.

Väänänen, J. (2012). Second-order logic or set theory. *Bulletin of Symbolic Logic*, 18(1):91–121.

Väänänen, J. (2020). Tracing internal categoricity. *Theoria*, 0:1–10.

Väänänen, J. and Wang, T. (2015). Internal categoricity in arithmetic and set theory. *Notre Dame Journal of Formal Logic*, 56(1):121–134.

Van Fraassen, B. (1980). *The Scientific Image*. Oxford University Press.

Volpe, G. (2011). Cornerstones: You'd better believe them. *Synthese*, 189(2):317–336.

Walmsley, J. (2002). Categoricity and indefinite extensibility. *Proceedings of the Aristotelian Society*, 102(1):239–257.

Walsh, S. and Ebels-Duggan, S. (2015). Relative categoricity and abstraction principles. *The Review of Symbolic Logic*, 8(3):572–606.

Warren, J. (2020). Infinite reasoning. *Philosophy and Phenomenological Research*, 103(2):385–407.

Waxman, D. (2017). Deflationism, Arithmetic, and the Argument from Conservativeness. *Mind*, 126(502):429–463.

Waxman, D. (b). Transmission failure and the semantic argument for consistency. Manuscript.

Williams, M. (1988). Epistemological realism and the basis of scepticism. *Mind*, 97(387):415–439.

Williams, M. (2013). Skepticism, evidence and entitlement. *Philosophy and Phenomenological Research*, 87(1):36–71.

Williamson, T. (2000). *Knowledge and its limits*. Oxford University Press.

Wright, C. (1994). About "the philosophical significance of Gödel's theorem": Some Issues. In McGuinness, B. and Oliveri, G., editors, *The Philosophy of Michael Dummett*, pages 167–204. Springer.

Wright, C. (1996). Precis of truth and objectivity. *Philosophy and Phenomenological Research*, 56(4):863.

Wright, C. (2002). (Anti-)sceptics simple and subtle: G.E. Moore and John McDowell*. *Philosophy and Phenomenological Research*, 65(2):330–348.

Wright, C. (2003). Some reflections on the acquisition of warrant by inference. In Nuccetelli, S., editor, *New Essays on Semantic Externalism and Self-Knowledge*, pages 57–78. MIT Press.

Wright, C. (2004). Warrant for nothing (and foundations for free)? *Proceedings of the Aristotelian Society, Supplementary Volumes*, 78:167–245.

Wright, C. (2012). Replies part iv: Warrant, transmission and entitlement. In Coliva, A., editor, *Mind, Meaning, and Knowledge. Themes from the Philosophy of Crispin Wright*, pages 451–486. Oxford University Press.

Wright, C. (2014). On epistemic entitlement (ii): Welfare state epistemology. In Dodd, D. and Zardini, E., editors, *Scepticism and Perceptual Justification*, pages 213–247. Oxford University Press.

Wright, C. (2016). Abstraction and epistemic entitlement: On the epistemological status of hume's principle. In Ebert, P. A. and Rossberg, M., editors, *Abstractionism: Essays in Philosophy of Mathematics*, pages 161–185. Oxford University Press.

Zicchetti, M. (2022a). Cognitive projects and the trustworthiness of positive truth. *Erkenntnis*.

Zicchetti, M. (2022b). The moderate conception, other minds and knowledge closure. *under review*.