UNIVERSITY OF CALIFORNIA,
IRVINE

Implicit Commitments of Theories of Arithmetic

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Philosophy

by

Thomas Colclough

Dissertation Committee:
Assistant Professor Toby Meadows, Chair
Chancellor's Professor Jeffrey Barrett
Associate Professor Jeremy Heis

2023

Dedication

To Catherine and Alan, for empowering me.

Table of Contents

Acknowledgments

I would like to express the deepest gratitude to my committee chair, Toby Meadows. His questions, feedback, and patience have been invaluable and informative to me throughout. I would also like to extend a sincere thank you to my committee members, Jeff Barrett, and Jeremy Heis, for their support and trust. Without all of these things, this dissertation would not have become possible.

All three of my committee members model different academic styles; all three styles share an infectious joy for philosophy. For inviting me to share in their different approaches during the last few years, my gratitude to all three also extends far beyond this work, from both a personal and a professional perspective.

Finally, I would like to thank my family and friends, for their unwavering support on the road to this Ph.D. I am indebted to Catherine, Alan, and Hannah; this degree would not have become a possibility without their confidence and encouragement. Thank you to Danny Mann and Katie Kearns; their mentorship taught me balance. And thank you to Mackenzie, for being my rock.

Vita

Thomas Colclough

2016	B.Sc. in Mathematics and Philosophy with specialism in Logic and Foundations, University of Warwick, U.K.
2017	Post-Graduate Certificate in Education (Mathematics), Staffordshire University, U.K.
2023	Ph.D. in Philosophy, University of California, Irvine, U.S.A.

Field of study

Logic and Philosophy of Science

Abstract of the Dissertation

Implicit Commitments of Theories of Arithmetic

by

Thomas Colclough

Doctor of Philosophy in Philosophy

University of California, Irvine, 2023

Assistant Professor Toby Meadows, Chair

The notion of *implicit commitments* of arithmetical theories has received some recent interest in the literature. However, current accounts lack a full understanding of: (1) what, mathematically, implicit commitments consist of, and (2) the epistemological force of the *commitment* of implicit commitments. This dissertation aims to provide a thorough account of (1) and (2).

Chapter 1

Introduction

Our central question of interest is: what are implicit commitments of theories of arithmetic? We aim to answer this question by answering two derivative questions in turn: *mathematically*, what are implicit commitments of theories of arithmetic? What is the *epistemological force* behind implicit commitments? Thus, our approach in this dissertation might be described as "mathematics first." By properly understanding the mathematics of implicit commitments, we can learn something about what the epistemological force of implicit commitments consists in; about what the *commitment* is, of implicit commitments.

We begin by addressing the first question above. The idea is simple: we cash out a set of implicit commitments of a theory of arithmetic S as axiomatized by a theory T extending S. We propose various different extensions of S, each corresponding to different sets of implicit commitments. Roughly, the thought is that in accepting S, one is implicitly committed to accept the extension T of S. Our mathematical work has interesting philosophical ramifications, and paves the way for our epistemological investigation. We approach the second question above by asking: what is the warrant we have for the mathematical principles which axiomatize theories of implicit commitments? We argue that traditional kinds of epistemic warrant cannot explain some of the consequence of our mathematical work. We formulate a kind of warrant which can. Finally, we give an account of the epistemological force behind implicit commitments, by examining the force underlying the warrant we have for them.

1.1 Motivation

What is the value in answering these questions? The mathematics of implicit commitments has received some recent attention (Nicolai & Piazza, 2019). This account proposes a *fixed* theory of implicit commitments, and claims to thereby offer a mathematical conception of implicit commitments compatible with a wide range of foundational positions in the philosophy of mathematics. The account in (Nicolai & Piazza, 2019) has merits. However, we think the idea of a fixed set of implicit commitments is a hasty move. We show that there are equally plausible alternatives to this fixed set of implicit commitments. We aim to improve on this account, and offer a more thorough understanding of the mathematics of implicit commitments.

Furthermore, one popular line of thought is that the implicit commitments of an arithmetical theory S include *reflection principles for* S, statements asserting that whatever is provable in S is true. The problem with this line is that it is untenable for certain foundational positions and their corresponding philosophies of mathematics. These foundational positions hold that in accepting certain theories S, one is *not* thereby implicitly committed to accept reflection principles for S. More generally, another line of thought is that the implicit commitments of an arithmetical theory S include arithmetical sentences not provable in S (for example, the canonical consistency sentence for S). Again, this is untenable for certain foundational positions and their corresponding philosophies of mathematics. In general, these foundational positions hold that in accepting certain theories S, one is not thereby implicitly committed to accept *any* sentences not provable in S.

Recent commentaries on implicit commitments seem to favor one or more of these lines of thought at the expense of another. For example, the fixed theory of implicit commitments proposed in (Nicolai & Piazza, 2019) excludes reflection principles. Łełyk and Nicolai (2022) propose an axiomatization of the minimal commitments implicit in the acceptance of a theory S which includes reflection principles for S, and reject the idea that one may accept S, but not be implicitly committed to accept the corresponding reflection principle. Horsten (2021) argues that there is nothing wrong with the idea that one may accept S but not be implicitly committed to accept any other sentences not provable in S, and thus rejects the idea that the implicit commitments of an arithmetical theory S can include arithmetical sentences not provable in S. There is no account of implicit commitments in the literature on which these seemingly incompatible lines of thought can coexist. We approach things differently, and claim to offer such an account.

Finally, we do not think there is a clear epistemological account in the literature of the warrant we have for implicit commitments (much less an epistemological account of the warrant we have for implicit commitments which can reconcile the three lines of thought above). For example, suppose implicit commitments are understood to include reflection principles. Some argue that our trust in reflection principles is as warranted as our trust in S (Fischer, 2021). Some argue that accepting S in conjunction with a disquotational conception of truth provides sufficient warrant to accept reflection principles for S (Fischer, Horsten, & Nicolai, 2021; Horsten & Leigh, 2016). Others argue that accepting S in conjunction with a fully compositional conception of truth provides sufficient warrant to accept reflection principles for S (Ketland, 2005, 2010; Shapiro, 1998). Yet others argue that the warrant for reflection principles comes from a "process of reflection" on the accepted theory S (Cieśliński, 2010; Tennant, 2002, 2005). Some have proposed strategies to extend S by reflection principles warranted by acceptance of S (Cieśliński, 2017; Feferman, 1962, 1991; Franzén, 2004; Turing, 1939). Some view the warrant for reflection through the lens of the distinction between the epistemic notions of entitlement and justification (Burge, 1993; Wright, 2004). Fischer et al. (2021) adopt the view that in accepting a theory S, one is entitled to accept a reflection principle for S. Horsten and Leigh (2016) argue that when we are justified in believing a theory, we are thereby entitled to adopt a corresponding reflection principle. In other accounts, Horsten (2021) proposes an epistemological analysis of the process of reflection which assumes the consistency of the accepted theory. Łełyk and Nicolai (2022) argue that justified belief in the axioms of a theory is preserved to a corresponding reflection principle. All sorts of epistemic notions have appeared here: warrant, trust, acceptance, justified belief. It is not clear how to (if we can at all) understand these terms uniformly across this literature. We hope to offer a clearer story about the warrant we have for implicit commitments in general. (This includes a story about the warrant we have for reflection principles.) Along the way, we also fix an understanding of all these epistemic notions, and say how they feature in our story.

Thus, overall, we aim to provide a more thorough understanding of the implicit commitments of theories of arithmetic, whereby the epistemology of implicit commitments is responsive to the mathematics of implicit commitments.

1.2 Outline

Let us describe our argumentative route. We begin in chapter 2 by formulating an answer to the question: mathematically, what are implicit commitments of theories of arithmetic? Our approach in chapter 2 is motivated by three existing ideas in the literature: (1) the idea that in accepting a given theory S, one is somehow rationally obliged to accept sentences not provable in S, (2) the idea that in accepting S, one is not thereby implicitly committed to accept sentences not provable in S, and (3) the idea that in accepting S, one is thereby implicitly committed to accept reflection principles for S. We postpone saying anything about the epistemological notion of acceptance occurring in (1), (2), and (3), until chapter 3. Rather, our goal in chapter 2 is to cash out implicit commitments simply as theories, which extend suitable theories of arithmetic S. We offer a mathematical framework for analysis of theories of implicit commitments, according to which these three seemingly incompatible ideas can coexist.

A consequence of our mathematical work in chapter 2 is that there are theories of arithmetic S, theories of implicit commitments T_1, $\mathsf{T}_2 \supseteq \mathsf{S}$, and foundational positions in the philosophy of mathematics which hold that:

- in accepting S, one is implicitly committed to accept T_1, and
- in accepting S, one is implicitly committed to accept T_2, but
- in accepting S, one is not implicitly committed to accept $\mathsf{T}_1 \cup \mathsf{T}_2$.

This consequence of our work in chapter 2 motivates our approach to answering the second question above: what is the epistemological force behind implicit commitments?

The intuitive idea is that in claiming that I am implicitly committed to a principle (P), then I *should* believe (P). What is the force of this *should*? To get a handle on the epistemological force underlying implicit commitments, we try to put our finger on the warrant we have for implicit commitments, in a way that is responsive to our mathematical observations from chapter 2. Along the way, we also narrow down on what the epistemological notion of acceptance involves. Identifying the warrant we have for implicit commitments takes us through chapters 3–6. In chapter 3, we consider two forthcoming suggestions: the epistemological notions of justification, and entitlement, in Crispin Wright's (2004) sense. We say more about what these notions involve.

We argue that justification is not the warrant we have for implicit commitments, since this idea is incompatible with our mathematical observations from chapter 2. Then we introduce Wright's notion of entitlement. In chapter 4, we argue that entitlement is not the warrant we have for implicit commitments, but for a different reason. However, our discussion of entitlements suggests a way forward for us. Drawing on the upshots of our discussion, in chapter 5, we formulate a different kind of warrant, which we suggest is the warrant we have for implicit commitments. We spend chapter 6 making our case. By the end of chapter 6, we arrive at an understanding of the warrant we have for implicit commitments. We propose an account of the epistemological force behind implicit commitments, by saying what the force underlying the warrant we have for them consists in. In chapter 7 we conclude, and suggest some further directions for this work.

Chapter 2

The mathematics of implicit commitments

We begin by focusing on the following central question: *mathematically*, what are implicit commitments of theories of arithmetic? Our strategy is to cash out sets of implicit commitments of a theory of arithmetic S, as axiomatized by various theories which extend S. The idea to keep in mind is that in accepting S, one is implicitly committed to accept these extensions of S. However, for the time being, we set aside the epistemological notion of acceptance.[1] With the mathematics of implicit commitments better understood, we will then be in a position to tackle the epistemology of implicit commitments.

Our approach to the central question is motivated by three existing threads of discussion, which we introduce briefly here. First: certain foundational po-

[1] If it is helpful, think of acceptance for the time being as a very broad notion, encompassing all sorts of epistemological attitudes. For example, we might take ourselves to hold a justified belief in the axioms of S. We say much more about this in chapter 3.

sitions in the philosophy of mathematics are said to be *epistemically stable*, in the sense of (Dean, 2015). We think of these foundational positions as being associated with a certain base theory S. Roughly, a foundational position with base theory S is epistemically stable if the position holds that in accepting certain theories S, one is not thereby implicitly committed to accept sentences not provable in S. Second: the epistemic stability of certain foundational positions stands in tension with the *implicit commitment thesis*, also in the sense of (Dean, 2015). Roughly, the implicit commitment thesis states that accepting a theory S implicitly commits one to accept additional statements not provable in S. Third: Nicolai and Piazza (2019) propose an account of implicit commitments which claims to offer certain foundationalists a way of reconciling the epistemic stability of their respective positions, with a version of the implicit commitment thesis.

One of the attractive features of the account in (Nicolai & Piazza, 2019) is that it accommodates minimal *soundness* assertions involving a notion of truth for an arithmetic base theory.[2] In particular, the implicit commitments of a theory S includes an *axiom soundness* principle for S, which expresses the claim that all the axioms of S are true. However, Nicolai and Piazza (2019) argue that one's implicit commitments in accepting a theory S are *fixed*. We think the idea of a fixed core of implicit commitments is a problem. We aim to offer a clear analysis of the underlying structure of the account in (Nicolai & Piazza, 2019). We propose a framework for understanding the

[2] This is a non-trivial achievement: soundness assertions involving the notion of truth are not typically expressible in the language of S, and most truth-free surrogates of soundness assertions for S are not provable in S.

mathematics of implicit commitments which is comprised of two components, *semantic* and *schematic*, axiomatized by various degrees which correspond to (sets of) principles extending a given arithmetic theory S. In contrast to the account in (Nicolai & Piazza, 2019), we argue that neither of these two components are fixed in general. Overall, our goal is to offer a more plausible understanding of the mathematics of implicit commitments, which improves on the account in (Nicolai & Piazza, 2019). Along the way, we also hope to offer a clearer understanding of the idea of epistemic stability and the implicit commitment thesis. In particular, we aim to better understand what is required, mathematically, for the notion of epistemic stability to be compatible with the implicit commitment thesis. We unfold both ideas, and use our account of theory acceptance to reveal exactly the sets of principles such that, if acceptance of those sets of principles is warranted purely on the basis of accepting S, then the resulting mathematical picture is compatible with both of these ideas.

Here is our approach in more detail. We begin in section 2.1 by identifying two threads of discussion in the literature described above. In section 2.1.1, to introduce the idea of epistemic stability, we outline Isaacson's thesis (Isaacson, 1996), a position we call *first-orderism*.[3] Second, in section 2.1.2 we outline the argument in (Dean, 2015), which claims that the epistemic stability of first-orderism conflicts with the implicit commitment thesis. We make some observations on the notion of epistemic stability and the implicit commitment thesis, and tease apart two weaker versions of both of these notions. In section

[3] Following the terminology in (Dean, 2015).

2.2, we introduce our third thread of discussion by outlining the account in (Nicolai & Piazza, 2019), which proposes the idea that there is a fixed core of implicit commitment in accepting a theory S, together with a variable component. We interleave our survey of the account in (Nicolai & Piazza, 2019) with our proposed framework for analyzing the mathematics of arithmetic theory acceptance. Our framework offers a clear method for analysis of those sets of principles, such that, if acceptance of those sets of principles is warranted purely on the basis of accepting S, then the resulting picture reconciles a version of epistemic stability with a version of the implicit commitment thesis. Our framework also helps clarify what we think the essence of the problem is with the account in (Nicolai & Piazza, 2019), which we then set out in section 2.3. In section 2.4, we provide the proof of a result, which yields a theory of implicit commitments different to the one proposed in (Nicolai & Piazza, 2019). In section 2.5 we reflect on what we have learned from our analysis, and segue into our investigation of the epistemology of implicit commitments.

2.1 Two threads of discussion

We begin by identifying two threads of discussion which motivate our account: first-orderism (Isaacson, 1996) and an introduction to the idea of epistemic stability, and Dean's (2015) claim that the epistemic stability of first-orderism conflicts with the implicit commitment thesis.

2.1.1 First-orderism and epistemic stability

First-orderism as described by Isaacson (1996) is the foundational standpoint that takes Peano Arithmetic (PA) to fully capture the notion of finitary mathematics, phrased as the claim that first-order PA "may be seen as complete for finite mathematics" (Isaacson, 1996, p. 204).[4] PA here is formulated in a first-order way, consisting of finitely many axioms together with the infinitely many axioms that comprise the first-order induction schema.[5] In particular, first-orderism takes PA itself to be justified on the basis of a Dedekindian conception of the natural numbers as "the smallest collection closed under a successor operation taking distinct elements to distinct elements and which contains not the successor of any element" (Isaacson, 1996, p. 205). Consequently, the theorems of PA "consist of those truths that can be perceived as true directly from the purely arithmetical content of a categorical conceptual analysis of the notion of natural number" (Isaacson, 1996, p. 203). In this way, PA captures a conceptually well-defined region of arithmetical truth, justified by our grasp of the structure of the natural numbers. This is phrased as the claim that PA is "complete with respect to purely arithmetical truth" (Isaacson, 1996, p. 222).

So, the theorems of PA form a subset of the class of mathematical truths. According to first-orderism, it is a proper subset. Since first-orderism holds that first-order PA is complete with respect to finitary, purely arithmetical

[4] In case it is not clear, do not confuse "the first-orderist" with Isaacson himself. We are concerned with the idea of the epistemic stability of the position, and do not intend to claim that the author of this idea occupies this position.

[5] See e.g., (Kaye, 1991) for an axiomatization of PA.

truth, sentences that we perceive to be true that lie beyond the provable reach of PA are called either not first-order, not finitary, or not purely arithmetical. Accordingly, to perceive the truth of such sentences, we require *higher-order*, *infinitary*, or *non-arithmetical* concepts. Examples of such sentences include the following: (1) the canonical Gödel sentence for PA (G(PA)), for which the justification is non-arithmetical (Isaacson, 1996, p. 214). (2) Goodstein's theorem, and Friedman's finitization of Kruskal's theorem, for which the justifications are infinitary (Isaacson, 1996, pp. 216, 219). (3) The Paris-Harrington sentence, for which the justification is higher-order (Isaacson, 1996, pp. 218–219).

To introduce the idea of epistemic stability: the completeness of PA with respect to purely arithmetical truth marks the boundary of precisely which mathematical sentences φ receive first-order, finitary, or purely arithmetical justification. A consequence of the first-orderist's acceptance of PA is that sentences φ such that $\mathsf{PA} \vdash \varphi$ are those truths that one can perceive as directly true from the purely arithmetical content of the notion of natural number. Since PA is said to be complete with respect to purely arithmetical truth, such φ are the only truths that one can perceive as directly true from the purely arithmetical content of the notion of natural number. Thus, the truth of statements beyond the theorems of PA is not (even implicitly) justified by the first-orderist's acceptance of PA, and instead any such (higher-order, infinitary, or non-arithmetical) justification must come from somewhere else. In this sense, according to the tenets of first-orderism, there exists a coherent rationale for accepting PA that does not entail or otherwise rationally oblige the first-orderist to accept statements beyond the logical consequences of PA itself. The last clause of the preceding sentence is the essence of epistemic stability.

The idea is that while one may come to accept a certain system of axioms, certain positions in the philosophy of mathematics hold that one's acceptance does not self-propagate beyond the system of axioms itself. In particular, a foundational position with associated base theory S is *epistemically stable* if there exists a coherent rationale for accepting S, that does not entail or otherwise rationally oblige a theorist to accept statements which cannot be derived from the axioms of S Dean (2015, p. 53). For example, characterizing theses of foundational positions said to be epistemically stable include: Dedekind's thesis (Dedekind, 1888/1965); Isaacson's thesis (Isaacson, 1996); Tait's thesis (Tait, 1981); the Feferman-Schütte thesis (Feferman, 1964; Kreisel, 1960; Schütte, 1965a, 1965b); and Nelson's thesis (Nelson, 1986).[6] We focus primarily on Isaacson's thesis. With epistemic stability introduced, let us turn to the implicit commitment thesis, and why Dean (2015) thinks these two notions are incompatible.

2.1.2 Conflict

One of the main goals in (Dean, 2015) paper is to argue that two foundational positions – finitism, as introduced by (Tait, 1981), and first-orderism, as above – are incompatible with what Dean calls the *implicit commitment thesis* (ICT):

[6] For discussion of these theses in the context of epistemic stability, see: (Madison & Waxman, 2021) (for Dedekind's thesis, Isaacson's thesis, Tait's thesis, and the Feferman-Schütte thesis); (Dean, 2015) (for Isaacson's thesis and Tait's thesis); (Nicolai & Piazza, 2019) (for Isaacson's thesis, Tait's thesis, and Nelson's thesis).

anyone who accepts the axioms of a mathematical theory S is thereby also implicitly committed to accepting various additional statements Γ which are expressible in the language of S but which are formally independent of its axioms (Dean, 2015, p. 32). We focus on Dean's argument in the context of first-orderism.

Essentially, Dean's argument is that the epistemic stability of PA makes first-orderism incompatible with the ICT, when the ICT is understood to include (for example) either of the following sentences among the resources Γ: the canonical Gödel sentence for PA (G(PA)), or the canonical consistency statement for PA (Con(PA)). That $\mathsf{PA} \nvdash \mathrm{G}(\mathsf{PA})$ is the content of Gödel's first incompleteness theorem. That $\mathsf{PA} \nvdash \mathrm{Con}(\mathsf{PA})$ is the content of Gödel's second incompleteness theorem (in fact, G(PA) and Con(PA) are equivalent over PA). Thus, G(PA) and Con(PA) lie beyond the provable resources of PA. On the other hand, both G(PA) and Con(PA) are examples of statements that are commonly said to be true.[7].

So, on one hand, according to the epistemic stability of first-orderism, accepting PA does *not* entail or otherwise thereby rationally oblige the first-orderist to accept G(PA) or Con(PA). On the other hand, suppose G(PA) and Con(PA) are counted among the additional resources Γ, to which one is implicitly committed as specified by the ICT. Since the first-orderist accepts PA, the ICT decrees that the first-orderist *is* rationally obliged to count those sentences among their implicit commitments on the basis of their acceptance of PA. Thus, where $\mathsf{S} = \mathsf{PA}$ and Γ includes G(PA) and Con(PA), Dean's claim is

[7] See e.g., (Dummett, 1963; Gödel, 1931/1986; Mostowski, 1952; Shapiro, 1998; Tennant, 2002; Wright, 1994)

that the first-orderist has reason to reject the ICT, and the ICT is incompatible with first-orderism. To be clear, it does not follow from this incompatibility that the first-orderist does not accept the sentences G(PA) or Con(PA). Rather, the first-orderist does not accept these sentences purely on the basis of their acceptance of PA. If the first-orderist does accept G(PA) or Con(PA), then the justification for that acceptance is grounded in higher-order/infinitary/non-arithmetical concepts.

Before we look at Nicolai and Piazza's proposed resolution in their (2019), let us pause and reflect. First, observe that the idea of epistemic stability as formulated in (Dean, 2015), and the implicit commitment thesis, are very close to being *logically* incompatible. Recall: a theory S is epistemically stable if there exists a coherent rationale for accepting S that does not entail or otherwise rationally oblige a theorist to accept statements which cannot be derived from the axioms of S. If we suppose that:

(1) if there exists a coherent rationale for accepting S, then it is possible for one to accept S;

(2) if the implicit commitment to accept, featuring in the ICT, implies an entailment or otherwise rational obligation to accept; and

(3) if accepting S implies accepting the axioms of S;

then epistemic stability is logically incompatible with the ICT. For if S is epistemically stable per Dean's definition and (1)–(3) are true, then it follows that a theory S is epistemically stable if it is possible for one to accept the axioms of S, yet this does not entail or otherwise rationally oblige one to accept statements which cannot be derived from the axioms of S. But the ICT implies that anyone's acceptance of the axioms of S entails or otherwise rationally obliges one to accept statements which cannot be derived from the axioms of S. If this is the case, one cannot rationally maintain that S is both epistemically stable, and that the ICT is true.

These observations reveal at least three ways in which the notion of epistemic stability associated with first-orderism might be reconciled with the corresponding version of the ICT. One might try and argue that at least one of (1)–(3) are false, or at least that the first-orderist would think that at least one of (1)–(3) are false. However, we do not attempt to make any such arguments here. Rather, the point of making these observations is to argue that the notion of epistemic stability defined (Dean, 2015) is particularly *strong*; so strong that in fact it is very close to being logically incompatible with the ICT per (Dean, 2015). All things considered, the strong version of epistemic stability cannot be reconciled with this version of the ICT. However, rather than leaving things here, our goal is to modify both this strong version of epistemic stability, and the articulation of the ICT, so that the modified versions are reconcilable in interesting ways. In particular, we aim to tease apart the strong version of epistemic stability from a weaker version, offer a weaker version of the ICT, and subsequently argue that it is the weaker version of epistemic stability which can be reconciled with the weaker version of the ICT in interesting ways.

We make these modifications in a couple of stages. First consider epistemic stability. Recall this notion again: a theory S is epistemically stable if there exists a coherent rationale for accepting S that does not entail or otherwise rationally oblige a theorist to accept statements which cannot be derived from the axioms of S. "Statements" here is understood as *any* statements, and our first step in isolating the weaker notion of epistemic stability that we are interested in, is to relax that requirement. Instead of ruling out the availability of *any* statements which cannot be derived from the axioms of S, we require only that *statements in the language of* S which cannot be derived from the axioms of S are ruled out. Denoting the language of S by $\mathcal{L}_{\mathsf{S}}$, we propose the following, weaker notion of epistemic stability:

> A theory S is *epistemically stable for $\mathcal{L}_{\mathsf{S}}$-sentences*, abbreviated as *$\mathcal{L}_{\mathsf{S}}$-epistemically stable*, if there exists a coherent rationale for accepting S that does not entail or otherwise rationally oblige a theorist to accept statements in the language of S which cannot be derived from the axioms of S.[8]

To isolate the weaker version of the ICT we are interested in, we make an analogous move. Recall the ICT: anyone who accepts the axioms of a mathematical theory S is thereby also implicitly committed to accepting various additional statements Γ which are expressible in the language of S but which are formally independent of its axioms. We broaden the class of additional statements Γ the acceptor is implicitly committed to accepting to *any* statements, rather than merely statements expressible in the language of S. We

[8] Articulated this way, $\mathcal{L}_{\mathsf{S}}$-epistemic stability is a property of a theory. In our discussion we consider foundational positions with an associated base theory. In these contexts we equivocate between using $\mathcal{L}_{\mathsf{S}}$-epistemic stability as a property of both the foundational position and its associated base theory.

propose the following, weaker version of the ICT:

> (Weak ICT): anyone who accepts the axioms of a mathematical theory S is thereby also implicitly committed to accepting various additional statements Γ which are formally independent of its axioms.

Weakening both the original notion of epistemic stability and the original version of the ICT in this way, we immediately have at our disposal new possible strategies for reconciling the notion of $\mathcal{L}_{\mathsf{S}}$-epistemic stability for first-orderism with the corresponding version of the weak ICT. For example, we might now try to argue that the first-orderist's acceptance of PA entails or otherwise rationally obliges the first-orderist to accept sentences *not in the language of* PA. On one hand, this would serve to make a case that the first-orderist can accept the weak ICT. On the other hand, if one could show that the extension of PA by those sentences cannot derive any consequences in the language of PA that PA cannot already derive, then the first-orderist's position is compatible with the idea of $\mathcal{L}_{\mathsf{S}}$-epistemic stability. This is the essence of the approach in (Nicolai & Piazza, 2019), so let us turn now to our third thread of discussion.

2.2 A framework for resolution

In their (2019) paper, Nicolai and Piazza propose an account of theory acceptance which aims to reconcile the notions of $\mathcal{L}_{\mathsf{S}}$-epistemic stability and the weak ICT. In section 2.2.1 we survey this account. In section 2.2.2 we propose our framework for analyzing candidate theories of implicit commitments with respect to $\mathcal{L}_{\mathsf{S}}$-epistemic stability and the weak ICT.

2.2.1 The semantic core

The central thesis in (Nicolai & Piazza, 2019) is this:

> when accepting a [mathematical] system S, we are bound to accept a fixed set of principles extending S and expressing minimal soundness requirements for S... there is also a variable component of implicit commitment that crucially depends on the justification given for our acceptance of S. (Nicolai & Piazza, 2019, p. 913)

The fixed set of principles extending S and expressing minimal soundness requirements for S is called the *semantic core of* S. These principles consist of fully compositional axioms for truth, and the axiom asserting that all of the axioms of S are true (this is the minimal soundness requirement for S). The principles of the semantic core are formulated in the extension of the language of S by a new unary predicate $T(x)$, intended as a truth predicate. The goal is to ensure that the semantic core of S is a (syntactically) conservative extension of S. A theory T_1 is *syntactically conservative* over another theory T_2 iff for every formula φ in the language of T_2, if $\mathsf{T}_1 \vdash \varphi$, then $\mathsf{T}_2 \vdash \varphi$.[9] However, whether or not the semantic core *exhausts* one's implicit commitments depends on the particular foundational standpoint that leads one to accept a given theory S in the first place in (Nicolai & Piazza, 2019, p. 929). In some

[9] We henceforth use "conservative", rather than "syntactically conservative."

cases, there are non-semantic considerations that feature in the justification for certain foundational standpoints. These considerations relate to attitudes towards schematic reasoning – in particular, the extent to which one is implicitly committed to instances of induction schema in which predicates occur that are not part of the language of S. As a result, in addition to the semantic core, there is a *variable* component of theory acceptance, one that can be articulated in terms of implicit schematic commitments.

Thus, following our remarks at the end of the section 2.1.2, the general idea is: on one hand, a foundationalist's acceptance of a given base theory S implicitly commits that foundationalist to accept sentences not in the language of S, which are not derivable in S (these sentences form the semantic core of S). As a result, that foundationalist may also hold that the weak ICT is true. But also, for suitable theories S, the semantic core of S is conservative over S. Thus, for suitable S, the foundationalist's acceptance of S does not entail or otherwise rationally oblige that foundationalist to accept sentences in the language of S, which are not derivable in S. The conservativity of the semantic core is the content of the following (Leigh, 2015, Theorem 1):

Theorem 1. (Leigh) Let S interpret $\mathsf{I}\Delta_0 + \exp$. Then the semantic core of S is a conservative extension of S.

Let us sketch the idea behind the proof. We denote the semantic core of S by $\mathsf{S}^T_{\mathrm{Ax}_\mathsf{S}}$.[10] To prove Theorem 1 we formulate $\mathsf{S}^T_{\mathrm{Ax}_\mathsf{S}}$ as a sequent calculus which includes the following cut rule for truth:

$$\frac{\Gamma \Rightarrow \Delta, T(\ulcorner\varphi\urcorner) \qquad \Gamma, T(\ulcorner\varphi\urcorner) \Rightarrow \Delta}{\Gamma \Rightarrow \Delta}\,(\mathrm{Cut}_T)$$

[10] We explain this notation shortly.

We also consider a bounded version $(\mathsf{S}^T_{\mathrm{Ax}_\mathsf{S}})^*$ of $\mathsf{S}^T_{\mathrm{Ax}_\mathsf{S}}$, where the rule (Cut_T) is replaced by the following schema of bounded cut rules, one for each $k < \omega$:

$$\frac{\Gamma \Rightarrow \Delta, T(\ulcorner \varphi \urcorner) \qquad \Gamma, T(\ulcorner \varphi \urcorner) \Rightarrow \Delta \qquad \Gamma, \mathrm{Sent}_{\mathcal{L}_\mathsf{S}}(\varphi) \Rightarrow \underset{\cdot}{d}(\ulcorner \varphi \urcorner) < \underline{k}}{\Gamma \Rightarrow \Delta} \; (\mathrm{Cut}^k_T)$$

Here $\underset{\cdot}{d}(\ulcorner \varphi \urcorner) < \underline{k}$ reads: the logical depth of the $\mathcal{L}_\mathsf{S}$-sentence φ is $< k$. Derivations in $(\mathsf{S}^T_{\mathrm{Ax}_\mathsf{S}})^*$ and $\mathsf{S}^T_{\mathrm{Ax}_\mathsf{S}}$ are defined in the usual way. The *truth rank* of a derivation is the least r such that for any rule (Cut^k_T) occurring in the derivation, $k < r$. A standard reduction argument is used to show that if the sequent $\Gamma \Rightarrow \Delta$ is derivable in $(\mathsf{S}^T_{\mathrm{Ax}_\mathsf{S}})^*$ with truth rank $r + 1$, then there is a derivation of the same sequent with truth rank r, whence $(\mathsf{S}^T_{\mathrm{Ax}_\mathsf{S}})^*$ is conservative over S. $\mathsf{S}^T_{\mathrm{Ax}_\mathsf{S}}$ is then embedded into $(\mathsf{S}^T_{\mathrm{Ax}_\mathsf{S}})^*$ using the notion of *approximations* from (Kotlarski, Krajewski, & Lachlan, 1981). Derivations in $\mathsf{S}^T_{\mathrm{Ax}_\mathsf{S}}$ are replaced by approximations with bounded depth, and so can be carried out in $(\mathsf{S}^T_{\mathrm{Ax}_\mathsf{S}})^*$. Since $(\mathsf{S}^T_{\mathrm{Ax}_\mathsf{S}})^*$ is conservative over S, so is $\mathsf{S}^T_{\mathrm{Ax}_\mathsf{S}}$.

The resulting picture is such that the foundationalist may come to accept a set of implicit commitments expressing minimal soundness requirements for S, in such a way that these implicit commitments are also compatible with the idea of $\mathcal{L}_\mathsf{S}$-epistemic stability. In this way, for a range of foundational positions, the notion of $\mathcal{L}_\mathsf{S}$-epistemic stability and the weak ICT are reconciled.

We think the account in (Nicolai & Piazza, 2019) has merits. In particular, we broadly agree that the components of implicit commitment are plausible components for an account of theory acceptance. However, we think this account falls short in supposing that one component of theory acceptance is *fixed*, and that one is *variable*. In particular, we think the idea that the semantic core is a fixed component of theory acceptance is too strong. To explain why, let us examine the idea in (Nicolai & Piazza, 2019) of a fixed semantic core of implicit commitment, and a variable component of implicit commitment, in more detail. To do this, we introduce a general framework for analyzing the two components of theory under consideration in (Nicolai & Piazza, 2019) with respect to the following three goals: (1) isolating sets of implicit commitments for suitable theories S which express minimal soundness requirements for S, (2) isolating sets of implicit commitments for suitable theories S which meet the criteria for $\mathcal{L}_{\mathsf{S}}$-epistemic stability, and (3) isolating sets of implicit commitments for suitable theories S which satisfy the weak ICT. We believe this framework offers a clear way of analyzing how these goals are to be met, and a clear way of drawing out what we think the problem is with the idea of a fixed semantic component and a variable schematic component.

2.2.2 Components of arithmetic theory acceptance

We offer a framework for analyzing two components of theory acceptance, for a fixed arithmetical base theory S with suitable coding capabilities, say S which interprets Robinson Arithmetic Q.[11] While we focus on the case where $\mathsf{S} = \mathsf{PA}$ later on, one may in general take S to be a range of other theories: *Buss arithmetic* S^1_2,[12] *Primitive Recursive Arithmetic* PRA,[13] or the fragments $\mathsf{I}\Sigma_n$ of PA (for $n \in \omega$). We denote the *language of* S by $\mathcal{L}_{\mathsf{S}}$, and we consider

[11] See e.g., (Kaye, 1991) for a definition of Q.

[12] (Buss, 1986; Simpson, 2009).

[13] In the sense of Skolem (1923/1967), essentially a reconstruction of the notion of finitary reasoning put forward by Hilbert and Bernays (1968). Strictly speaking, PRA is a theory formulated in a quantifier-free language, and uses a schema of rules in place of the schema of induction axioms. To smoothly apply the framework of this paper, it would be natural to move to the theory conservative extension $\mathsf{QF{-}IA}$ of PRA by first-order quantifiers, but we ignore this technical distinction.

the language $\mathcal{L}_T$ obtained by expanding $\mathcal{L}_{\mathsf{S}}$ with a new unary predicate $T(x)$ (intended as a truth predicate).[14]

We denote the theory of a foundationalist's implicit commitments on the basis of their acceptance of S as an $\mathcal{L}_T$-theory $\mathsf{I}(\mathsf{S})$ extending S. This aligns with the idea that the foundationalist's implicit commitments in accepting S are sentences in the extended language. Next, we axiomatize two components of theory acceptance. One of these components we call the *semantic* component of accepting S, intended to capture implicit commitments about (the behavior of) truth, along with minimal soundness principles for S. The second of these components we call the *schematic* component of accepting S, intended to capture implicit commitments about extending S's induction schema to permit the occurrence of the truth predicate. These two components of accepting S align respectively with what (Nicolai & Piazza, 2019) call the *fixed* and *variable* components of accepting S. Our choice of titles for these two components of accepting S stems from our disagreement with the use of "fixed" and "variable" as they are used by (Nicolai & Piazza, 2019) to describe the two components.

The semantic component of accepting S is axiomatized by the following four $\mathcal{L}_T$-theories extending S.

Definition 1. S^U is the $\mathcal{L}_T$ theory extending S with the schema of *uniform Tarski biconditionals* for $\mathcal{L}_{\mathsf{S}}$; i.e. all sentences of the form:

$$\forall x_1, \ldots, x_n (T(\varphi(\underline{x_1}, \ldots, \underline{x_n})) \leftrightarrow \varphi(x_1, \ldots, x_n))$$

[14] We note that $\mathcal{L}_T$ is not, strictly speaking, uniform, since $\mathcal{L}_{\mathsf{S}}$ may differ for different choices of S. But this too is a technical distinction we may ignore for our purposes.

for every $\mathcal{L}_{\mathsf{S}}$-formula $\varphi(x_1, \ldots, x_n)$.

Definition 2. $\mathsf{S}^U_{\mathrm{Ax}_\mathsf{S}}$ is the $\mathcal{L}_T$ theory $\mathsf{S}^U + \forall x(\mathrm{Ax}_\mathsf{S}(x) \rightarrow T(x))$. We call the sentence $\forall x(\mathrm{Ax}_\mathsf{S}(x) \rightarrow T(x))$ the *axiom soundness axiom for* S.

$\mathrm{Ax}_\mathsf{S}(x)$ is a Δ_0-formula expressing that x is the code of a non-logical axiom of S. In conjunction with minimal principles governing the behavior of the truth predicate (i.e., uniform Tarski biconditionals), which are really what license the name "truth" for the unary predicate $T(x)$, the axiom soundness axiom for S says that all the axioms of S are true. The axiom soundness axiom for S is the minimal soundness requirement for S aimed at in the account in (Nicolai & Piazza, 2019).

Definition 3. S^T is the $\mathcal{L}_T$ theory extending S with the following *fully compositional* truth axioms:

1. $\forall x(T(x) \rightarrow \mathrm{Sent}_{\mathcal{L}_\mathsf{S}}(x))$.

2. $\forall s, t(T(\ulcorner s = t \urcorner) \leftrightarrow (s^\circ = t^\circ))$.

3. $\forall \varphi(T(\ulcorner \neg\varphi \urcorner) \leftrightarrow \neg T(\ulcorner \varphi \urcorner))$.

4. $\forall \varphi, \psi(T(\ulcorner \varphi \vee \psi \urcorner) \leftrightarrow (T(\ulcorner \varphi \urcorner) \vee T(\ulcorner \psi \urcorner)))$.

5. $\forall v \forall \varphi(v)(T(\ulcorner \exists v \varphi(\underline{v}) \urcorner) \leftrightarrow \exists x T(\ulcorner \varphi(\underline{x}) \urcorner))$.

Here s° denotes the value of the term s, and similarly for t°. $\mathrm{Sent}_{\mathcal{L}_\mathsf{S}}(x)$ is a Δ_1-formula expressing that x is the code of a sentence of $\mathcal{L}_\mathsf{S}$.

Definition 4. $\mathsf{S}^T_{\mathrm{Ax}_\mathsf{S}}$ is the $\mathcal{L}_T$ theory $\mathsf{S}^T + \forall x(\mathrm{Ax}_\mathsf{S}(x) \rightarrow T(x))$.

We note that in a general setting, the axiom soundness axiom for S is a non-trivial addition to principles 1–5: theories that are not finitely axiomatizable cannot prove the corresponding axiom soundness axiom in the presence of principles 1–5 (Nicolai & Piazza, 2019, Lemma 1). However, finitely axiomatizable theories can.

The theory $\mathsf{S}^T_{\mathrm{Ax}_\mathsf{S}}$ is precisely the semantic core of S in (Nicolai & Piazza, 2019, p. 928). As we noted above, the axiom soundness axiom for S captures the idea of minimal soundness requirements for S. Fully compositional truth is motivated by the desiderata that the semantic core ought to be able to establish that instances of modus ponens preserve truth, and that the semantic core ought to capture a compositional notion of truth (Nicolai & Piazza, 2019, pp. 926–927). Crucially, in each of the four theories defined above, the predicate $T(x)$ is not allowed to appear in instances of S's induction schema. For a variety of arithmetical theories S, it is well-known that the result of expanding the language of S with a new unary predicate $T(x)$ which is fully compositional and allowed to occur in formulas appearing in S's induction schema is not conservative over S.[15]

For fixed S, the four theories above axiomatize four degrees of the semantic component of accepting S. They represent four possible ways of capturing the foundationalist's implicit semantic commitments in accepting S. Together with the trivial position, according to which the foundationalist has no im-

[15] For the non-conservativity result where S is S^1_2, see (Nicolai & Piazza, 2019, Proposition 3). Indeed, full compositionality of the truth predicate is not necessary; we may obtain non-conservativity even in the presence of a uniform disquotational truth. The non-conservativity results where S is any of the theories $\mathsf{I}\Sigma_n$ for $n \in \omega$, are obtained similarly. We will see a non-conservativity proof in the case where S is PA later on.

plicit semantic commitments in accepting S, we may depict five degrees of the foundationalist's implicit semantic commitment in accepting S in the following way.

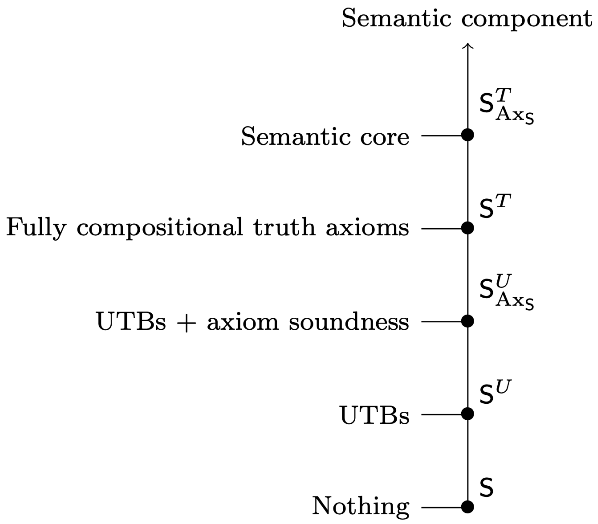
Semantic component
$\mathsf{S}^T_{\mathrm{Ax}_\mathsf{S}}$
Semantic core
S^T
Fully compositional truth axioms
$\mathsf{S}^U_{\mathrm{Ax}_\mathsf{S}}$
UTBs + axiom soundness
S^U
UTBs
S
Nothing

Figure 2.1: The semantic component of implicit commitment

This picture is more fine-grained than the picture in (Nicolai & Piazza, 2019). There, the semantic core $\mathsf{S}^T_{\mathrm{Ax}_\mathsf{S}}$ of S is a *fixed* component of implicit semantic commitment in accepting S. However, in what follows, we are interested in what happens when we consider implicit commitments which do not contain the full resources of $\mathsf{S}^T_{\mathrm{Ax}_\mathsf{S}}$.

We note that in general, since the uniform Tarski biconditionals cannot derive the fully compositional truth axioms, and the fully compositional truth axioms cannot derive the axiom soundness axiom, $\mathsf{S}^T_{\mathrm{Ax}_\mathsf{S}}$ is the strongest of these theories, and S^U is the weakest. We also note that the ordering of S^T and $\mathsf{S}^U_{\mathrm{Ax}_\mathsf{S}}$ is somewhat arbitrary, since in general, S^T cannot derive the axiom soundness axiom, but can derive the uniform Tarski biconditionals for S, and $\mathsf{S}^U_{\mathrm{Ax}_\mathsf{S}}$ cannot derive the fully compositional truth axioms. However, the ordering of S^T and $\mathsf{S}^U_{\text{Axs}}$ in Figure 2.1 does not really matter for our purposes, so without loss of generality we opt for the picture above.

To axiomatize degrees of implicit schematic commitment, we consider the case where the predicate $T(x)$ is allowed into instances of the induction schema of each of the theories S^U, $\mathsf{S}^U_{\text{Axs}}$, S^T, and $\mathsf{S}^T_{\text{Axs}}$. Instances of induction schema are stratified according to the complexity of formulas appearing in them in the usual way. We say that a formula is Δ_0 if all quantifiers it contains are bounded. We say that a formula is Σ_1 (resp. Π_1) if it is of the form $\exists x \varphi$ (resp. $\forall x \varphi$) where φ is Δ_0. We say that a formula is Σ_n (resp. Π_n) if it is of the form $\exists x \varphi$ (resp. $\forall x \varphi$) where φ is Π_{n-1} (resp. Σ_{n-1}). We say that a formula is Δ_n if it is both Σ_n and Π_n. The theory $\mathsf{I}\Gamma$ is Robinson Arithmetic Q plus induction for formulae in the class Γ. If $\mathcal{L}_P$ is a language extending $\mathcal{L}_A$ with a new predicate P, we write $\mathsf{I}\Sigma_n(\mathcal{L}_P)$ for the $\mathcal{L}_P$-theory extending PA with instantiations of the induction scheme for $\mathcal{L}_P$-formulas in the class Σ_n.

Definition 5. Let W be any of S^U, $\mathsf{S}^U_{\text{Axs}}$, S^T, or $\mathsf{S}^T_{\text{Axs}}$. Then $(\mathsf{W})_n$ is the $\mathcal{L}_T$-theory axiomatized by $\mathsf{I}\Sigma_n(\mathcal{L}_T)$. $(\mathsf{W})_\omega$ is the $\mathcal{L}_T$-theory axiomatized by $\bigcup_{n\in\omega} \mathsf{I}\Sigma_n(\mathcal{L}_T)$.

Thus, $(\mathsf{W})_n$ is the $\mathcal{L}_T$-theory extending W with instantiations of the induction scheme for $\mathcal{L}_T$-formulas in the class Σ_n ("$\Sigma_n(T)$-induction". When $n = 0$, we write "$\Delta_0(T)$-induction" in place of "$\Sigma_0(T)$-induction.") Putting everything together, Figure 2.2 below depicts the semantic and schematic components of implicit commitment in accepting S. It will turn out that some of the theories of Figure 2.2 coincide, but we address that further on.

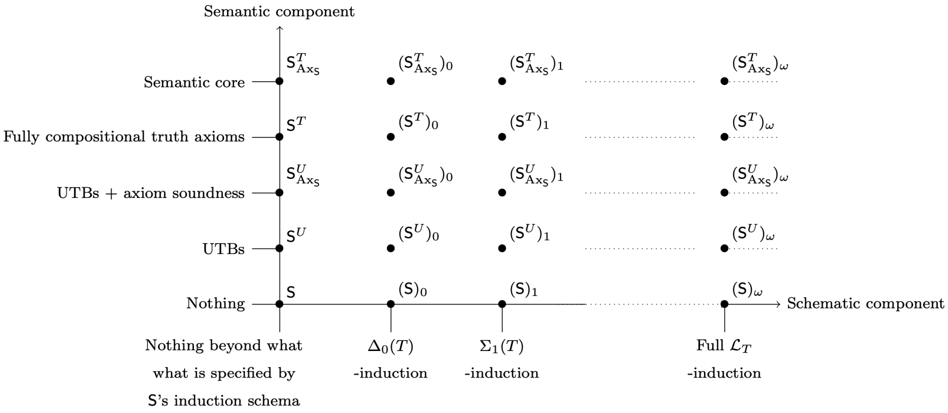
Semantic component
Semantic core
Fully compositional truth axioms
UTBs + axiom soundness
UTBs
Nothing
$\mathsf{S}^T_{\mathrm{Ax_S}}$
$(\mathsf{S}^T_{\mathrm{Ax_S}})_0$
$(\mathsf{S}^T_{\mathrm{Ax_S}})_1$
$(\mathsf{S}^T_{\mathrm{Ax_S}})_\omega$
S^T
$(\mathsf{S}^T)_0$
$(\mathsf{S}^T)_1$
$(\mathsf{S}^T)_\omega$
$\mathsf{S}^U_{\mathrm{Ax_S}}$
$(\mathsf{S}^U_{\mathrm{Ax_S}})_0$
$(\mathsf{S}^U_{\mathrm{Ax_S}})_1$
$(\mathsf{S}^U_{\mathrm{Ax_S}})_\omega$
S^U
$(\mathsf{S}^U)_0$
$(\mathsf{S}^U)_1$
$(\mathsf{S}^U)_\omega$
S
$(\mathsf{S})_0$
$(\mathsf{S})_1$
$(\mathsf{S})_\omega$
Schematic component
Nothing beyond what what is specified by S's induction schema
$\Delta_0(T)$ -induction
$\Sigma_1(T)$ -induction
Full $\mathcal{L}_T$ -induction

Figure 2.2: The semantic and schematic components of implicit commitment

Together, the semantic and schematic components of accepting S align respectively with what Nicolai and Piazza (2019) call the fixed and variable components of accepting S. According to their account, the semantic core of S is a fixed component of implicit semantic commitment in accepting S. On the other hand, whether or not the semantic core exhausts one's implicit commitments depends on the particular foundational standpoint that leads one to accept a given theory S in the first place in (Nicolai & Piazza, 2019, p. 929). In particular, if the foundationalist is implicitly committed to instances of induction schema in which the truth predicate occurs, then the foundationalist's implicit commitments in accepting S may also include non-trivial schematic implicit commitments. Thus, this type of commitment may vary from foundationalist to foundationalist.

We are now in a position to draw out what we think the problem is with the idea that the semantic component of implicit commitment in accepting S is fixed, but the schematic component of implicit commitment in accepting S varies.

2.3 The problem with a fixed semantic core

To motivate the problem, let us first consider some remarks about what Nicolai and Piazza (2019) call the variable component of implicit commitment. The variable component of acceptance is introduced to us by way of several examples of different foundational standpoints which adopt different views on extending the induction schema of the arithmetical systems associated with them (Nicolai & Piazza, 2019, Section 4). Nicolai and Piazza argue that these views depend on the justification given for a particular foundational theory itself. Here are three examples.

First, consider the case in which one does not allow extensions of the induction schema to permit extra-linguistic vocabulary at all. A paradigmatic example of this sort is finitism as articulated by (Tait, 1981). The associated foundational theory is Primitive Recursive Arithmetic (PRA). PRA is formulated in a quantifier free language, and so the schema of induction is replaced by the schema of rules:

$$\frac{\varphi(0) \qquad \varphi(x) \to \varphi(Sx)}{\varphi(x)}\,(\mathrm{IR})$$

where formulas $\varphi(v)$ appearing in (IR) are Δ_0. The finitist is committed to instances of (IR) on the basis of their acceptance of PRA insofar as those instances involve predicates that are expressible by formulas in the language of arithmetic, and are at most Δ_0. By Tarski's theorem on the undefinability of truth, a full truth predicate is not expressible by any formula in the language of arithmetic. Thus, for the finitist, the justification for claims about the totality of the natural numbers made via (IR) that involve a (fully compositional) truth predicate is not implicit in the finitist's acceptance of PRA. What's more, though, is that the finitist is reluctant to admit that (IR) even *applies* to predicates that are not expressible by formulas in the language of arithmetic. This is grounded in the justification the finitist gives for PRA itself. Instances of (IR) that involve predicates not expressible by formulas in the language of arithmetic are suspicious at best, and false at worst (Nicolai & Piazza, 2019, p. 930).

Second, consider the case in which one accepts instances of extended induction schema unrestrictedly, on the basis that one accepts the associated theory. The paradigmatic example of this sort is Feferman's reflective closure of a theory S (1991). There are two versions of this. The first is the *reflective closure* $\mathrm{Ref}(\mathsf{S})$ of a theory S. $\mathrm{Ref}(\mathsf{S})$ captures statements in the base language $\mathcal{L}$ of S that ought to be accepted on the basis of accepting the basic axioms and rules of S. The second is the *schematic reflective closure* $\mathrm{Ref}^*(\mathsf{S}(P))$ of a schematic version $\mathsf{S}(P)$ of a theory S. Given an arithmetic theory S, $\mathsf{S}(P)$ takes as axioms the usual ones governing the language of arithmetic together with the principle of induction in the form:

$$P(0) \wedge \forall x(P(x) \to P(x+1)) \to \forall x P(x),$$

where P is a new free predicate symbol/variable and is accompanied by an appropriate substitution rule (Feferman, 1991, pp. 1–2). Ref$^*(\mathsf{S}(P))$ captures the schemata $A(P)$ in the language of $\mathsf{S}(P)$ that ought to be accepted on the basis of accepting the basic schematic axioms and rules of $\mathsf{S}(P)$. Where S is PA, in the case of the reflective closure of PA, one obtains the self-applicable theory of truth KF, and Ref(PA) reaches the strength of ramified analysis up to ϵ_0. In the case of the schematic reflective closure of PA, one obtains a type-free theory of truth, and Ref(PA) reaches the strength of ramified analysis up to the Feferman-Schütte ordinal Γ_0 (Feferman, 1964; Schütte, 1965a).[16]

Third, consider the first-orderist. The two preceding positions hold different views about extending induction – views that roughly, are at either end of the extreme. According to Nicolai and Piazza, the first-orderist holds a third kind of view, which occupies what they call an intermediate position between the two preceding positions (2019, p. 931). On one hand, in the spirit of Feferman (and unlike the finitist), the first-orderist holds no particular reservations about the application of PA's induction schema to predicates that are not expressible by formulas in the language of arithmetic. On the other hand, the

[16] McGee also offers a reading of the position in which one extends induction unrestrictedly, arguing that induction schema are like the laws of logic, which we expect to persist through changes in language (1997, p. 58). Consequently acceptance of (for example) PA should commit one not only to instances of induction applied to extensions of one's language, but also to instances of induction corresponding to *any* subset of the natural numbers. An analysis of this sort yields categorical theories. For further discussion of McGee's position, see (Pedersen & Rossberg, 2010).

first-orderist's acceptance of instances of PA's induction schema that involve a truth predicate, if the first-orderist accepts such instances at all, is essentially higher-order/infinitary/non-arithmetical.[17] The thought seems to be that this is more in keeping with the spirit of the finitist idea above that the justification for claims about the totality of the natural numbers made via (IR) that involve a (fully compositional) truth predicate is not implicit in the finitist's acceptance of PRA.

Let us reflect on these examples. Our main observation is that we find the idea that the first-orderist holds an intermediate position between the finitist and foundationalists à la Feferman with respect to their views about extending induction to be a peculiar one. Recall Figure 2.2, and let S be PA (the first-orderist's base theory). On one hand, if the semantic core of PA is a fixed component of the first-orderist's implicit commitments in accepting PA, then $\mathsf{I}(\mathsf{PA})$ contains at least the theory $\mathsf{PA}^{T}_{\mathrm{Ax}_{\mathsf{PA}}}$. On the other hand, the first-orderist's views about extending induction are supposed to be such that the first-orderist may come to accept instances of PA's induction schema in which the truth predicate occurs, but if they do, then this acceptance is not purely grounded in their acceptance of PA. The theories PA, $(\mathsf{PA})_n$, for each $n \in \omega$, and $(\mathsf{PA})_\omega$, are our formal representation of accepting various (sets of) instances of PA's induction schema which permit the occurrence of the truth predicate on the basis of the first-orderist's acceptance of PA.[18] So which of the

[17] Cf. the outline of first-orderism in section 2.1.1.

[18] We acknowledge that the sense in which (for example) $(\mathsf{PA})_\omega$ is a formal representation of accepting all instances of PA's induction schema in which the *truth* predicate occurs may be a little artificial. Without at least the presence of the uniform disquotational principles, it doesn't really make sense to call the predicate T occurring in instances of PA's induction schema a *truth* predicate. Ultimately this won't be a problem, since every interesting set of

theories PA, $(\mathsf{PA})_n$, for each $n \in \omega$, and $(\mathsf{PA})_\omega$, correspond to what the first-orderist's implicit commitments about extensions of PA's induction schema are, on the basis of their acceptance of PA?

As far as we can tell, the answer according to (Nicolai & Piazza, 2019) should be PA, because if the first-orderist comes to accept instances of PA's induction schema in which the truth predicate occurs, this acceptance does not follow merely from their acceptance of PA. Thus, the first-orderist accepts *no* instances of PA's induction schema in which the truth predicate occurs *on the basis of their acceptance of* PA. But if this is the case, we do not see how the first-orderist is supposed to occupy an intermediate position between the finitist and foundationalists à la Feferman. In particular, we do not see what sets apart the first-orderist from the finitist with respect to their views on extending induction. The finitist, recall, accepts no instances of PRA's induction schema in which the truth predicate occurs simply because the finitist refuses to accept that the induction rule (IR) even *applies* to predicates that are not expressible by formulas in the language of arithmetic, and the latter must apply to $\mathcal{L}_T$ formulas. Thus, of the theories PRA, $(\mathsf{PRA})_n$, for each $n \in \omega$, and $(\mathsf{PRA})_\omega$, it is just PRA itself which captures the finitist's implicit commitments about extensions of PRA's induction schema on the basis of their acceptance of PRA. But then the first-orderist holds precisely an analogous set of implicit commitments about extensions of PA's induction schema on the

implicit commitments concerning instances of extended induction we consider in this paper also contain at least the uniform disquotational principles for the T predicate. In any case, artificial or not, we think the stratification of the schematic component of implicit commitment via the theories $(\mathsf{PA})_n$, for each $n \in \omega$, and $(\mathsf{PA})_\omega$, adds at least some pedagogical value to our framework.

basis of their acceptance of PA as the finitist does. Thus, we find it difficult to see how we are supposed to separate the first-orderist from the finitist in this respect. In general, we find it difficult to see how the position occupied by first-orderist with respect to their views about extending induction, on the basis of their acceptance of PA, is somehow in between those of the finitist and foundationalists à la Feferman.

What happens if the answer to the question above is not PA, but is one of the theories $(\mathsf{PA})_n$, for some $n \in \omega$, or $(\mathsf{PA})_\omega$? First consider $(\mathsf{PA})_\omega$. As before, the semantic core of PA is a fixed component of the first-orderist's implicit commitments in accepting PA, so $\mathsf{I}(\mathsf{PA})$ contains at least the theory $\mathsf{PA}^T_{\mathrm{Ax}_{\mathsf{PA}}}$. But if in addition $\mathsf{I}(\mathsf{PA})$ contains $(\mathsf{PA})_\omega$ we seem to be in some trouble. It is well-known that the result of extending PA by fully compositional truth and fully extended induction is a much stronger theory than PA. For instance, in the resulting theory, one easily derives the following *global reflection principle*:[19]

$$\forall\varphi(\mathrm{Pr}_{\mathsf{PA}}(\varphi) \to T(\varphi)). \tag{$\mathrm{GRP}_{\mathsf{PA}}$}$$

But from $(\mathrm{GRP}_{\mathsf{PA}})$ one can derive, for instance, $\mathrm{Con}(\mathsf{PA})$ (by instantiating the falsity $0 \neq 1$ in $(\mathrm{GRP}_{\mathsf{PA}})$). Thus, $\mathrm{Con}(\mathsf{PA})$ is also part of $\mathsf{I}(\mathsf{PA})$, and this sits in tension with one of the goals we set out to achieve: a set of implicit commitments on the basis of the first-orderist's acceptance of PA compatible with the idea of $\mathcal{L}_{\mathsf{PA}}$-epistemic stability – the idea that in accepting PA, the first-orderist is not forced by entailment or rational obligation to accept statements in the language of PA not derivable from the axioms of PA. $\mathrm{Con}(\mathsf{PA})$ is

[19] See e.g., (Wcisło & Łełyk, 2017).

exactly such a statement.

This leaves us to consider any of the theories $(\mathsf{PA})_n$, for some $n \in \omega$. But there, the situation is similar to above. Consider $(\mathsf{PA})_0$, which corresponds to the very *least* non-trivial set of implicit commitments about extensions of PA's induction schema the first-orderist may accept on the basis of their acceptance of PA. If the semantic core of PA is a fixed component of the first-orderist's implicit commitments in accepting PA, and $(\mathsf{PA})_0$ is also part of the first-orderist's implicit commitments in accepting PA, then $\mathsf{I}(\mathsf{PA})$ contains at least the theory $(\mathsf{PA}^T)_0$. And from $(\mathsf{PA}^T)_0$, we again obtain $\mathrm{Con}(\mathsf{PA})$. The reason is that $(\mathsf{PA}^T)_0$ plus the global reflection principle $(\mathrm{GRP}_{\mathsf{PA}})$ above is relatively interpretable in $(\mathsf{PA}^T)_0$.[20] This is the content of the following:[21]

Theorem 2. (Wcisło, Łełyk) $(\mathsf{PA}^T)_0 + \forall\varphi(\mathrm{Pr}_{\mathsf{PA}}(\varphi) \rightarrow T(\varphi))$ is interpretable in $(\mathsf{PA}^T)_0$ relative to PA.

The proof strategy is to recursively define a family of partial arithmetic truth predicates $T_n(x)$, for $n \in \omega$. This ensures that there is an arithmetical expression $x = T_n(v)$ representing in PA the recursive function assigning to n the code of the formula $T_n(v)$. For each $n \in \omega$, we may then apply the truth predicate to the code of $T_n(x)$ to obtain a family of predicates $T(\ulcorner T_c(x) \urcorner)$, where the parameter c is possibly nonstandard. In the presence of $\Delta_0(T)$-

[20] See for example (Lindström, 1997, Ch. 12) for a definition of relative interpretation.

[21] It is well-known that $(\mathsf{PA}^T)_1$ proves $(\mathrm{GRP}_{\mathsf{PA}})$ (Wcisło & Łełyk, 2017, Theorem 12). A natural question is whether one can relax the assumption of Π_1 T-induction, and ask whether $(\mathsf{PA}^T)_0$ proves $(\mathrm{GRP}_{\mathsf{PA}})$. Kotlarski (1968) originally published an alleged proof of a similar result using a theory of satisfaction, rather than truth, before Albert Visser and Richard Heck independently identified a gap in the proof. Theorem 2 shows that $(\mathrm{GRP}_{\mathsf{PA}})$ is arithmetically conservative over $(\mathsf{PA}^T)_0$. Wcisło and Łełyk (2017) also show that slightly modifying $(\mathsf{PA}^T)_0$ actually *proves* $(\mathrm{GRP}_{\mathsf{PA}})$.

induction, the predicates $T(\ulcorner T_c(x) \urcorner)$ are like truth predicates in the sense that they are compositional for formulas with codes less than c. The defining formula $T'(x)$ satisfying the axioms of $(\mathsf{PA}^T)_0 + (\mathrm{GRP}_{\mathsf{PA}})$ is then constructed by taking the supremum of the predicates $T(\ulcorner T_c(x) \urcorner)$ (see (Wcisło & Łełyk, 2017) or (Cieśliński, 2017, Theorem 12.3.4) for a full proof).

It follows that if $\mathsf{I}(\mathsf{PA})$ contains $(\mathsf{PA}^T)_0$, then $\mathrm{Con}(\mathsf{PA}) \in \mathsf{I}(\mathsf{PA})$.[22] Thus, again this sits in tension with one of the goals we set out to achieve: a set of implicit commitments on the basis of the first-orderist's acceptance of PA compatible with the idea of $\mathcal{L}_{\mathsf{PA}}$-epistemic stability.

Let us take stock. What emerges from this line of reasoning is that the idea of a fixed semantic core $\mathsf{PA}^T_{\mathrm{Ax}_{\mathsf{PA}}}$ of implicit commitments on the basis of their acceptance of PA is a problem for the first-orderist. There are (at least) two things we might conclude at this point. First, perhaps all this serves to show is that first-orderism is simply an incoherent view after all. We think this is a hasty move. The goal all along has been to reconcile the idea of $\mathcal{L}_{\mathsf{S}}$-epistemic stability with the weak ICT for various foundational positions said to be epistemically stable in some sense. If all we are prepared to conclude at this stage is that one of these foundational positions was incoherent all along, this does not seem very in keeping with our original goal.

Second, we might instead call into question the general idea of a fixed semantic core of implicit commitments in accepting a given theory S. Perhaps our framework can reveal a different, just as interesting, set of implicit commitments for the first-orderist, compatible with both $\mathcal{L}_{\mathsf{PA}}$-epistemic stability

[22] Strictly speaking, it follows that if $\mathsf{I}(\mathsf{PA})$ contains $(\mathsf{PA}^T)_0$, then $\mathrm{Con}(\mathsf{PA})$ is *relatively interpretable* in $\mathsf{I}(\mathsf{PA})$, but we do not think this detracts from the main point.

and the corresponding version of the weak ICT, *and* the idea that the first-orderist does not occupy a trivial position with respect to extensions of PA's induction schema. If this is the case, we may still hope to reconcile the idea of $\mathcal{L}_{\mathsf{S}}$-epistemic stability with the weak ICT for the various foundational positions which motivated this discussion after all. Next, we show that this is indeed possible.

2.4 Another resolution

Consider Figure 2.3, which depicts the semantic and schematic components of implicit commitment in accepting PA:

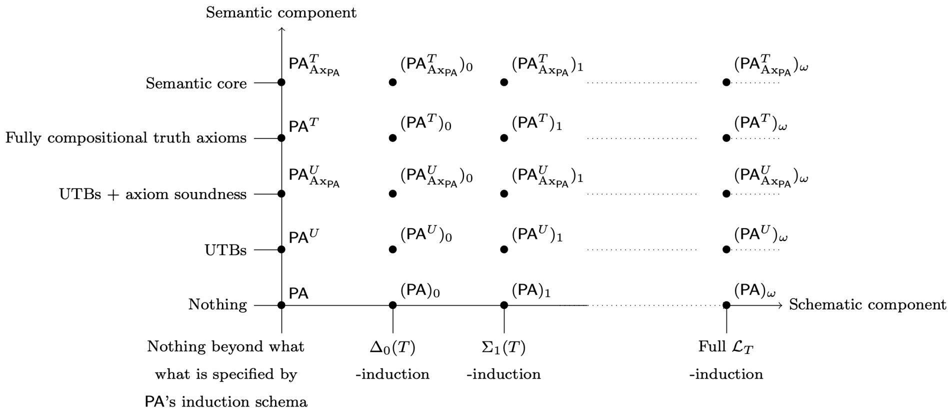
Semantic component
Semantic core
Fully compositional truth axioms
UTBs + axiom soundness
UTBs
Nothing
$\mathsf{PA}^T_{\mathrm{Ax}_{\mathsf{PA}}}$
$(\mathsf{PA}^T_{\mathrm{Ax}_{\mathsf{PA}}})_0$
$(\mathsf{PA}^T_{\mathrm{Ax}_{\mathsf{PA}}})_1$
$(\mathsf{PA}^T_{\mathrm{Ax}_{\mathsf{PA}}})_\omega$
PA^T
$(\mathsf{PA}^T)_0$
$(\mathsf{PA}^T)_1$
$(\mathsf{PA}^T)_\omega$
$\mathsf{PA}^U_{\mathrm{Ax}_{\mathsf{PA}}}$
$(\mathsf{PA}^U_{\mathrm{Ax}_{\mathsf{PA}}})_0$
$(\mathsf{PA}^U_{\mathrm{Ax}_{\mathsf{PA}}})_1$
$(\mathsf{PA}^U_{\mathrm{Ax}_{\mathsf{PA}}})_\omega$
PA^U
$(\mathsf{PA}^U)_0$
$(\mathsf{PA}^U)_1$
$(\mathsf{PA}^U)_\omega$
PA
(PA)$_0$
(PA)$_1$
(PA)$_\omega$
Schematic component
Nothing beyond what what is specified by PA's induction schema
$\Delta_0(T)$ -induction
$\Sigma_1(T)$ -induction
Full $\mathcal{L}_T$ -induction

Figure 2.3: The semantic and schematic components of implicit commitment in accepting PA

Our observations above told us that the theories $(\mathsf{PA}^T_{\mathrm{Ax}_{\mathsf{PA}}})_n$ and $(\mathsf{PA}^T)_n$, for each $n \in \omega$, are not conservative over PA.[23] Since our goal (in this context) is to meet the criteria for $\mathcal{L}_{\mathsf{PA}}$-epistemic stability, the first-orderist's implicit commitments $\mathsf{I}(\mathsf{PA})$ on the basis of their acceptance of PA can contain none of those theories. Of the remaining theories which correspond to a non-trivial implicit schematic commitment, that leaves the following theories for investigation:

- $(\mathsf{PA}^U_{\mathrm{Ax}_{\mathsf{PA}}})_n$, for each $n \in \omega$, and $(\mathsf{PA}^U_{\mathrm{Ax}_{\mathsf{PA}}})_\omega$.

[23] In fact, Theorem 2 told us that all of these theories coincide.

- $(\mathsf{PA}^U)_n$, for each $n \in \omega$, and $(\mathsf{PA}^U)_\omega$.
- $(\mathsf{PA})_n$, for each $n \in \omega$, and $(\mathsf{PA})_\omega$.

To examine which of these theories are potential candidates for the first-orderist's implicit commitments in accepting PA, we want to figure out which of these theories meets the criteria for $\mathcal{L}_{\mathsf{PA}}$-epistemic stability. The natural strategy at our disposal for showing any of these theories meets the criteria for $\mathcal{L}_{\mathsf{PA}}$-epistemic stability is to show that that theory is conservative over PA. It is well-known that each of the theories $(\mathsf{PA}^U)_n$, for each $n \in \omega$, and $(\mathsf{PA}^U)_\omega$, are conservative over PA.[24]

Our Theorem 3 below shows that each of the theories $(\mathsf{PA}^U_{\mathrm{Ax}_{\mathsf{PA}}})_n$, for each $n \in \omega$, and $(\mathsf{PA}^U_{\mathrm{Ax}_{\mathsf{PA}}})_\omega$, are also conservative over PA. Thus, we provide a complete classification of the theories of Figure 2.3 with respect to conservativity over PA.

Theorem 3 states that the theory obtained by adding to PA the uniform Tarski biconditionals, the full induction schema for $\mathcal{L}_T$-formulas, and the following axiom:

$$\forall x(D(x) \rightarrow T(x)),$$

is conservative over PA. Here $D(x)$ is a PA-*schema*, defined below. The case of interest is where $D(x)$ is $\mathrm{Ax}_{\mathsf{PA}}(x)$, the formula expressing that x is the code of an axiom of PA. We only sketch the proof of Theorem 3 here. A full proof is included in section 2.4.1.

[24] See e.g., (Halbach, 2011). Hence, so are the theories $(\mathsf{PA})_n$, for each $n \in \omega$, and $(\mathsf{PA})_\omega$.

Definition 6. Let p be a fresh unary predicate symbol not present in $\mathcal{L}_A$. An $\mathcal{L}_A$-formula D is a PA-*schema* if

1. $\mathsf{PA} \vdash D(\ulcorner \sigma \urcorner) \to \sigma$ for every formula $\sigma \in \mathcal{L}_A$, and
2. there exists a finite set U of $\mathcal{L}_A \cup \{p\}$-formulas with at most x free such that

$$\mathsf{PA} \vdash D(x) \to \exists \psi \bigvee_{\varphi \in U} (x = \ulcorner \varphi[\psi/p] \urcorner).$$

Theorem 3. Let D be a PA-schema. The theory $(\mathsf{PA}^U)_\omega + \forall x(D(x) \to T(x))$ is a conservative extension of PA.

Proof Sketch. We extend the strategy employed in (Leigh, 2015) to the theory $(\mathsf{PA}^U)_\omega$. We formulate the theory $(\mathsf{PA}^U)_\omega$ as a sequent calculus with a cut rule and an induction rule for the truth predicate. Alongside $(\mathsf{PA}^U)_\omega$ we consider $(\mathsf{PA}^U)^*_\omega$, a version of $(\mathsf{PA}^U)_\omega$ involving only bounded cuts. The presence of truth in induction means that $(\mathsf{PA}^U)^*_\omega$ does not, in general, admit cut elimination. However, it is shown in (Łełyk & Wcisło, 2017) that PA^T interprets $(\mathsf{PA}^U)_\omega$, whence the conservativity of $(\mathsf{PA}^U)^*_\omega$ over PA follows from Theorem 1.

By extending the proof of the key lemma in (Leigh, 2015) we show that $(\mathsf{PA}^U)_\omega$ embeds into $(\mathsf{PA}^U)^*_\omega$. Finally, derivations in $(\mathsf{PA}^U)_\omega$ expanded by the rule

$$\frac{\Gamma \Rightarrow \Delta, D(s)}{\Gamma \Rightarrow \Delta, T(s)} \text{(D)}$$

can be reduced to derivations in $(\mathsf{PA}^U)^*_\omega$ expanded by a corresponding rule, denoted $(\mathrm{D}_{\boldsymbol{w}})$. However, in fact $(\mathsf{PA}^U)_\omega$ interprets $(\mathrm{D}_{\boldsymbol{w}})$, whence derivations in $(\mathsf{PA}^U)_\omega + \text{(D)}$ can be carried out in $(\mathsf{PA}^U)^*_\omega$. Since $(\mathsf{PA}^U)^*_\omega$ conservatively extends PA, so does $(\mathsf{PA}^U)_\omega + \text{(D)}$. $\square$

From Theorem 3 we immediately obtain:

Corollary 1. For each $n \in \omega$, $(\mathsf{PA}^U_{\mathrm{Ax}_{\mathsf{PA}}})_n$ is conservative over PA. $\square$

We understand the import of Theorem 3 to consist in revealing a different (and interesting) set of implicit commitments for the first-orderist than the semantic core of PA, compatible with both the corresponding version of the weak ICT and $\mathcal{L}_{\mathsf{PA}}$-epistemic stability. Taking $\mathsf{I}(\mathsf{PA}) = (\mathsf{PA}^U_{\mathrm{Ax}_{\mathsf{PA}}})_\omega$, the first orderist has a set of implicit commitments which accommodates minimal soundness principles for PA, a (minimal) notion of truth, *and* fully extended arithmetic induction to the language $\mathcal{L}_T$. This set of implicit commitments satisfies the weak ICT as it applies to PA, and is compatible with the idea that the first-orderist's acceptance of PA (and hence of $(\mathsf{PA}^U_{\mathrm{Ax}_{\mathsf{PA}}})_\omega$) neither entails nor rationally obliges the first-orderist to accept statements in the language of PA not derivable from the axioms of PA. As far as the implicit commitments of the first-orderist are concerned, we think $(\mathsf{PA}^U_{\mathrm{Ax}_{\mathsf{PA}}})_\omega$ is just as plausible a set of implicit commitments as the semantic core $\mathsf{PA}^T_{\mathrm{Ax}_{\mathsf{PA}}}$ of PA.

We present the full proof of Theorem 3 below in section 2.4.1. The reader may instead skip ahead to section 2.5, where we continue our narrative.

2.4.1 Proof of Theorem 3

We extend the strategy employed in (Leigh, 2015) to the theory $(\mathsf{PA}^U)_\omega$. Let us fix some preliminaries and notational conventions.

1. We work with the language $\mathcal{L}_A^+ \supseteq \mathcal{L}_A$ which contains countably many new predicate symbols

   $$\mathcal{P} = \{p_j^i : i, j < \omega \text{ and } p_j^i \text{ is a predicate symbol with arity } i\},$$

   together with a new constant ϵ.

2. We assume a fixed Gödel coding of $\mathcal{L}_A^+$ into $\mathcal{L}_A$, which extends to finite sequences of $\mathcal{L}_A$-terms. In particular we have the following:

   (a) Unary predicates $\mathrm{Term}_{\mathcal{L}_A}(x)$ and $\mathrm{Sent}_{\mathcal{L}_A}(x)$ representing the sets of Gödel codes of arithmetical terms and sentences respectively. We extend this notation in the natural way to languages $\mathcal{L}$ extending $\mathcal{L}_A$.

   (b) The ternary substitution function $sub(x, y, z)$ defining the operation that replaces each occurrence of the variable with code y in the term or formula coded by x by the term with code z. We abbreviate $sub(x, y, z)$ by $x[z/y]$.

   (c) A unary predicate $\underset{\cdot}{d}$ defining the following operation on codes of $\mathcal{L}_A^+$ formulas:

   $$\underset{\cdot}{d}(\ulcorner \alpha \urcorner) = x \text{ iff the logical complexity of } \alpha \in \mathcal{L}_A^+ \text{ is } x.$$

   For readability, unless there is value in writing down Quine corners, we generally omit them when referring to Gödel codes of syntactic objects.

3. Greek lower-case letters α, β, γ, etc. from the start of the alphabet range over $\mathcal{L}_T$-formulas.

4. Greek lower-case letters φ, χ, etc. from the end of the alphabet range over $\mathcal{L}_A$-terms encoding $\mathcal{L}_A^+$-formulas. Greek lower-case letters in bold font $\boldsymbol{\varphi}, \boldsymbol{\psi}$, etc. denote finite sequences of $\mathcal{L}_A$-terms. If $\boldsymbol{\varphi} = \langle \varphi_0, \ldots, \varphi_k \rangle$ is a sequence of $\mathcal{L}_A$-terms, then $T(\boldsymbol{\varphi})$ denotes the set $\{T(\varphi_i) : i \leq k\}$.

5. Roman lower-case letters s, t, etc. range over $\mathcal{L}_A$-terms.

6. Greek upper-case letters $\Gamma, \Delta, \Sigma, \Pi$, etc. denote finite sets of $\mathcal{L}_T$-formulas.

Sequent calculi

We list the axioms and rules of two sequent calculi: $(\mathsf{PA}^U)_\omega$ and $(\mathsf{PA}^U)_\omega^*$. They differ only in their cut rules. To obtain $(\mathsf{PA}^U)_\omega^*$ from $(\mathsf{PA}^U)_\omega$, we replace the cut rule for the truth predicate by a version that applies only when the formula to which the truth predicate is being applied is provably of some bounded logical complexity.

Axioms.

1. $\Gamma \Rightarrow \Delta, \varphi$ if φ is an axiom of Q.

2. $\Gamma, \varphi(\underline{x}) \Rightarrow \Delta, T(\varphi(\underline{x}))$ where x is arbitrary and $\varphi(v)$ is any $\mathcal{L}_A$-formula.

3. $\Gamma, T(\varphi(\underline{x})) \Rightarrow \Delta, \varphi(\underline{x})$ where x is arbitrary and $\varphi(v)$ is any $\mathcal{L}_A$-formula.

Basic rules.

$$\frac{\Gamma \Rightarrow \Delta, \alpha}{\Gamma \Rightarrow \Delta, \forall v_i \alpha}\,(\forall\text{R}) \qquad \frac{\Gamma, \alpha(s/v_i) \Rightarrow \Delta}{\Gamma, \forall v_i \alpha \Rightarrow \Delta}\,(\forall\text{L})$$

$$\frac{\Gamma \Rightarrow \Delta, \alpha, \beta}{\Gamma \Rightarrow \Delta, \alpha \vee \beta}\,(\vee\mathrm{R}) \qquad \frac{\Gamma, \alpha \Rightarrow \Delta \qquad \Gamma, \beta \Rightarrow \Delta}{\Gamma, \alpha \vee \beta \Rightarrow \Delta}\,(\vee\mathrm{L})$$

$$\frac{\Gamma, \alpha \Rightarrow \Delta}{\Gamma \Rightarrow \Delta, \neg\alpha}\,(\neg\mathrm{R}) \qquad \frac{\Gamma \Rightarrow \Delta, \alpha}{\Gamma, \neg\alpha \Rightarrow \Delta}\,(\neg\mathrm{L})$$

Induction rule.

$$\frac{\Gamma, \varphi(x) \Rightarrow \Delta, \varphi(x+1)}{\Gamma, \varphi(\underline{0}) \Rightarrow \Delta, \varphi(t)}\,(\mathrm{Ind}_T)$$

where x is not free in the lower sequent, t is an arbitrary term, and $\varphi(v)$ is any formula in the language $\mathcal{L}_T$. $(\mathsf{PA}^U)_\omega$ and $(\mathsf{PA}^U)^*_\omega$ each include the axioms, basic rules, and induction rule. The cut rules for each are the following.

Cut rules for $(\mathsf{PA}^U)_\omega$**.**

$$\frac{\Gamma \Rightarrow \Delta, \varphi \qquad \Gamma, \varphi \Rightarrow \Delta}{\Gamma \Rightarrow \Delta}\,(\mathrm{Cut}_{\mathcal{L}_A})$$

In $(\mathrm{Cut}_{\mathcal{L}_A})$ the cut formula $\varphi \in \mathcal{L}_A$.

$$\frac{\Gamma \Rightarrow \Delta, T(\varphi) \qquad \Gamma, T(\varphi) \Rightarrow \Delta}{\Gamma \Rightarrow \Delta}\,(\mathrm{Cut}_T)$$

In (Cut_T) the formula under the truth predicate $\varphi \in \mathcal{L}_A$.

Cut rules for $(\mathsf{PA}^U)^*_\omega$**.**

$$\frac{\Gamma \Rightarrow \Delta, \varphi \qquad \Gamma, \varphi \Rightarrow \Delta}{\Gamma \Rightarrow \Delta}\,(\mathrm{Cut}_{\mathcal{L}_A})$$

For each $k < \omega$:

$$\frac{\Gamma \Rightarrow \Delta, T(\varphi) \qquad \Gamma, T(\varphi) \Rightarrow \Delta \qquad \Gamma, \mathrm{Sent}_{\mathcal{L}_A}(\varphi) \Rightarrow^* \underset{\cdot}{d}(\varphi) \leq \underline{k}}{\Gamma \Rightarrow \Delta} \,(\mathrm{Cut}_T^k)$$

where $\Rightarrow^*$ indicates that the sequent is derivable using only the axioms and arithmetical rules.

Lemma 1. $(\mathsf{PA}^U)^*_\omega$ is a conservative extension of PA.

Proof. Suppose the truth-free sequent $\Gamma \Rightarrow \Delta$ is derivable in $(\mathsf{PA}^U)^*_\omega$. Then $\Gamma \Rightarrow \Delta$ is derivable in $(\mathsf{PA}^U)_\omega$. Let d denote this derivation. It is shown in Łełyk and Wcisło, 2017, Proposition 4.15 that there exists an $\mathcal{L}_A$-conservative relative interpretation of $(\mathsf{PA}^U)_\omega$ in PA^T. That is, there is a translation $t : \mathcal{L}_T \to \mathcal{L}_T$ constant on arithmetical formulas such that for all $\varphi \in \mathcal{L}_T$:

$$\text{if } (\mathsf{PA}^U)_\omega \vdash \varphi \text{ then } \mathsf{PA}^T \vdash t(\varphi).$$

It follows that PA^T interprets d, whence $\Gamma \Rightarrow \Delta$ is derivable in PA^T, and the result follows from Leigh, 2015, Theorem 1 (i.e. Theorem 1). □

Approximations

The goal is to show that $(\mathsf{PA}^U)_\omega$ embeds into $(\mathsf{PA}^U)^*_\omega$. This is achieved via approximations, originally from (Kotlarski et al., 1981). Recall that we are working with the language $\mathcal{L}_A^+$ that extends $\mathcal{L}_A$ by a new constant ϵ and the set $\mathcal{P}$ of countably many new predicate symbols p^i_j.

Definition 7. Let $X \subseteq \mathcal{P}$ be a finite subset consisting of the predicates p^i_j. An *assignment* is any function $g : X \to \mathcal{L}_A^+$ such that for every i, j, if $p^i_j \in X$

then $g(p^i_j)$ is a formula with arity i.

If g is an assignment and $\varphi \in \mathcal{L}^+_A$, then $\varphi[g]$ denotes the result of replacing each predicate $p^i_j(s_1, \ldots, s_i)$ occurring in φ by $g(p^i_j)(s_1, \ldots, s_i)$ if $g(p^i_j)$ is defined, and by ϵ otherwise.

Definition 8. Let $\boldsymbol{\varphi} = \langle \varphi_0, \ldots, \varphi_m \rangle$ and $\boldsymbol{\psi} = \langle \psi_0, \ldots, \psi_m \rangle$ be two sequences of closed $\mathcal{L}^+_A$-formulas. We say that $\boldsymbol{\varphi}$ *approximates* $\boldsymbol{\psi}$ if there exists an assignment g such that $\psi_i = \varphi_i[g]$ for each $0 \leq i \leq m$.

We are interested in defining a particular class of approximations; n-th approximations. They are constructed in the following way. Let $w, z, z_1, z_2, \ldots$ be new variable symbols.

Definition 9. Let $\varphi \in \mathcal{L}_A$. An *occurrence in* φ is any pair $\langle \varphi', t \rangle$ such that:

1. $\varphi' \in \mathcal{L}_A \cup \{z\}$ such that z occurs in φ' exactly once;

2. $\mathrm{Term}_{\mathcal{L}_A \cup \{w\}}(t)$;

3. t is free for z in φ';

4. $\varphi = \varphi'[t/z]$.

We denote the set of occurrences in φ by $\mathcal{O}(\varphi)$.

Definition 10. Let $\varphi \in \mathcal{L}_A$. The *w-free form of* φ is the $\mathcal{L}_A \cup \{w\}$-formula $\overline{\varphi}$ obtained from φ by:

1. replacing all free variables in φ by the variable w;

2. replacing all terms in the result of 1. above in which the only variable that occurs in w, by w.

If $\langle \varphi', t \rangle$ is an occurrence in φ where φ is in w-free form, then $t = w$. We say that two $\mathcal{L}_A$-formulas φ and ψ are *weakly equivalent* if their w-free forms are equal; i.e. if $\overline{\varphi} = \overline{\psi}$.

Each $\mathcal{L}_A$-formula φ is associated with a unique function $t_\varphi : \mathcal{O}(\varphi) \to \mathrm{Term}_{\mathcal{L}_A}$ such that replacing each occurrence of the variable w in the w-free form of φ by $t_\varphi(w)$ results in φ. We say that two $\mathcal{L}_A$-formulas φ and ψ are *strongly equivalent*, which we write as $\varphi \approx \psi$, if they are weakly equivalent and in addition there exists an equivalence relation E on $\mathcal{O}(\overline{\varphi}) = \mathcal{O}(\overline{\psi})$ such that t_φ, t_ψ are well-defined on ${}^{\mathcal{O}(\Phi)}/_{E}$ and disagree on at most finitely many E-equivalence classes.

Let Φ be a set of pairwise weakly equivalent $\mathcal{L}_A$-formulas, such that each has only a finite number of free variables. There is a canonical way of defining an equivalence relation E on $\mathcal{O}(\Phi)$ as above, since $\mathcal{O}(\Phi)$ is the common value of $\mathcal{O}(\overline{\varphi})$ for each $\varphi \in \Phi$. The functions $\{t_\varphi : \varphi \in \Phi\}$ induce an equivalence relation E_Φ on $\mathcal{O}(\Phi)$ by setting:

$$\langle \varphi_0, t_0 \rangle E_\Phi \langle \varphi_1, t_1 \rangle \Leftrightarrow \bigwedge \{ t_\varphi(\langle \varphi_0, t_0 \rangle) = t_\varphi(\langle \varphi_1, t_1 \rangle) : \varphi \in \Phi \}.$$

For each $\varphi \in \Phi$, let $t'_\varphi : {}^{\mathcal{O}(\Phi)}/_{E_\Phi} \to \mathrm{Term}_{\mathcal{L}_A}$ be the map induced by t_φ. If $\varphi_0, \varphi_1 \in \Phi$ then there are at most finitely many E_Φ-equivalence classes in ${}^{\mathcal{O}(\Phi)}/_{E_\Phi}$ on which t'_{φ_0} and t'_{φ_1} disagree.

We use the notion of strong equivalence to define the *template* of a set Φ of

$\mathcal{L}_A$-formulas. Let Φ be a set of pairwise strongly equivalent $\mathcal{L}_A$-formulas. Let $\mathcal{C}_1, \ldots, \mathcal{C}_l$ enumerate the finitely many E_Φ-classes $\mathcal{C}$ in ${}^{\mathcal{O}(\Phi)}/_{E_\Phi}$ such that there are $\psi_0, \psi_1 \in \Phi$ with $t'_{\psi_0}(\mathcal{C}) \neq t'_{\psi_1}(\mathcal{C})$, or $t'_{\psi_0}(\mathcal{C})$ is a variable. Let $\varphi \in \Phi$ and consider its w-free form $\overline{\varphi}$. If $\mathfrak{o} \in \mathcal{O}(\overline{\varphi})$ is such that $\mathfrak{o} \in \mathcal{C}_i$ for some $1 \leq i \leq l$, we replace $\mathfrak{o}$ by the new variable z_i. Otherwise $\mathfrak{o} \in \mathcal{O}(\overline{\varphi})$ is not in any $\mathcal{C}_i$, so we replace $\mathfrak{o}$ by $t_\varphi(\mathfrak{o})$. The resulting formula:

$$\Theta_\Phi(z_1, \ldots, z_l)$$

is called the *template* of Φ. The template of Φ is unique up to permutation of the variables $z_1, \ldots, z_l$, and does not depend on the choice of φ. Also, for each $\varphi \in \Phi$ there exist unique terms $t_1, \ldots, t_l$ such that $\varphi = \Theta_\Phi(t_1, \ldots, t_l)$.

Next we give a sequence of definitions that culminate in the definition of n-th approximations.

Definition 11. Let $\boldsymbol{\varphi} = \langle \varphi_0, \ldots, \varphi_m \rangle$ be a non-empty sequence of $\mathcal{L}_A$-formulas. The *set of parts of* $\boldsymbol{\varphi}$, denoted $\Pi(\boldsymbol{\varphi})$, is the set of pairs $\langle \varphi', \psi \rangle$ such that:

1. $\varphi' \in \mathcal{L}_A \cup \{\epsilon\}$ is such that ϵ occurs in φ' exactly once;

2. $\psi \in \mathcal{L}_A$; and

3. $\varphi_i = \varphi'[\psi/\epsilon]$ for some $0 \leq i \leq m$.

We define an ordering $\preceq$ on $\Pi(\boldsymbol{\varphi})$ such that

$$\langle \varphi_0, \psi_0 \rangle \preceq \langle \varphi_1, \psi_1 \rangle$$

iff there exists $\chi \in \mathcal{L}_A \cup \{\epsilon\}$ with $\varphi_0 = \varphi_1[\chi/\epsilon]$ and $\psi_1 = \chi[\psi_0/\epsilon]$.

Definition 12. Let $\langle\varphi,\psi\rangle \in \Pi(\boldsymbol{\varphi})$. The *depth* of $\langle\varphi,\psi\rangle$, denoted $d(\varphi,\psi)$, is the number of logical operators of φ within whose scope ϵ falls under.

Using the notion of strong equivalence and the ordering $\preceq$, we define the following sets recursively on k.

$$\begin{aligned} \Pi^{(0)}(\boldsymbol{\varphi},n) &= \{\langle\varphi,\psi\rangle \in \Pi(\boldsymbol{\varphi}) : d(\varphi,\psi) \leq n\} \\ \Pi^{(k+1)}(\boldsymbol{\varphi},n) &= \{\langle\varphi,\psi\rangle \in \Pi(\boldsymbol{\varphi}) : \exists\langle\varphi_1,\psi_1\rangle \in \Pi^{(k)}(\boldsymbol{\varphi},n)\exists\langle\varphi_0,\psi_0\rangle \in \Pi^{(0)}(\boldsymbol{\varphi},n) \\ &\qquad\qquad \big(\psi_0 \approx \psi_1 \wedge \langle\varphi,\psi\rangle \preceq \langle\varphi_1,\psi_1\rangle \\ &\qquad\qquad \wedge\, d(\varphi,\psi) + d(\varphi_0,\psi_0) \leq d(\varphi_1,\psi_1) + n\big)\}. \end{aligned}$$

Intuitively, $\Pi^{(k+1)}(\boldsymbol{\varphi},n)$ consists of the parts of $\boldsymbol{\varphi}$ that are approximated by some $\langle\varphi_1,\psi_1 \in \Pi^{(k)}(\boldsymbol{\varphi},n)$, such that the template of φ_1 occurs in $\boldsymbol{\varphi}$ with depth at most n.

For large enough $k < \omega$, $\Pi^{(k)}(\boldsymbol{\varphi},n)$ is fixed; i.e. there exists j such that $\Pi^{(j)}(\boldsymbol{\varphi},n) = \Pi^{(j+1)}(\boldsymbol{\varphi},n)$. Fix such a j and define

$$\begin{aligned} \Gamma(\boldsymbol{\varphi},n) &= \{\psi \in \mathcal{L}_A : \exists\varphi\langle\varphi,\psi\rangle \in \Pi^{(j)}(\boldsymbol{\varphi},n)\} \\ \Gamma_I(\boldsymbol{\varphi},n) &= \{\psi \in \mathcal{L}_A : \exists\varphi\langle\varphi,\psi\rangle \text{ is } \preceq\text{-minimal in } \Pi^{(j)}(\boldsymbol{\varphi},n)\}. \end{aligned}$$

Let $\approx$ partition $\Gamma_I(\boldsymbol{\varphi},n)$ into the set of equivalence classes ${}^{\Gamma_I(\boldsymbol{\varphi},n)}/_{\approx}$. Let $\Phi_0,\ldots,\Phi_l$ enumerate the elements of ${}^{\Gamma_I(\boldsymbol{\varphi},n)}/_{\approx}$. For $0 \leq i \leq l$, let $\Theta_{\Phi_i}(z_1,\ldots,z_{l_{\Phi_i}})$ be the template of Φ_i, with arity l_{Φ_i}. For each $\varphi \in \Phi_i$, let $t_1^{\varphi},\ldots,t_{l_{\Phi_i}}^{\varphi}$ be the terms such that $\varphi = \Theta_{\Phi_i}(t_1^{\varphi},\ldots,t_{l_{\Phi_i}}^{\varphi})$.

Definition 13. Define a function:

$$F_{\boldsymbol{\varphi},n} : \Gamma(\boldsymbol{\varphi}, n) \to \mathcal{L}_A^+$$

recursively by:

1. $F_{\boldsymbol{\varphi},n}(\psi) = \psi$ if $\psi \in {}^{\Gamma_I(\boldsymbol{\varphi},n)}/\approx$ is atomic.

2. $F_{\boldsymbol{\varphi},n}(\psi) = p_i^{l_{\Phi_i}}(t_1^{\psi}, \dots, t_{l_{\Phi_i}}^{\psi})$ if $\psi \in \Phi_i \subseteq {}^{\Gamma_I(\boldsymbol{\varphi},n)}/\approx$ (for some $0 \leq i \leq l$) is not atomic.

3. If $\psi \in \Gamma(\boldsymbol{\varphi}, n) \setminus \Gamma_I(\boldsymbol{\varphi}, n)$, define:

   (a) $F_{\boldsymbol{\varphi},n}(\psi_0 \vee \psi_1) = F_{\boldsymbol{\varphi},n}(\psi_0) \vee F_{\boldsymbol{\varphi},n}(\psi_1)$.

   (b) $F_{\boldsymbol{\varphi},n}(\neg\psi) = \neg F_{\boldsymbol{\varphi},n}(\psi)$.

   (c) $F_{\boldsymbol{\varphi},n}(\exists x \psi) = \exists x F_{\boldsymbol{\varphi},n}(\psi)$.

Definition 14. Let $\boldsymbol{\varphi} = \langle \varphi_0, \dots, \varphi_m \rangle$ be a sequence of closed $\mathcal{L}_A^+$-formulas. The *n-th approximation of* $\boldsymbol{\varphi}$ is the sequence:

$$F_{\boldsymbol{\varphi},n}(\boldsymbol{\varphi}) = \langle F_{\boldsymbol{\varphi},n}(\varphi_0), \dots, F_{\boldsymbol{\varphi},n}(\varphi_m) \rangle.$$

Clearly the construction of $F_{\boldsymbol{\varphi},n}$ can be formalized within PA; in fact, in a much weaker theory. We therefore adopt the following notation.

1. $(r)_i = s$ indicates that r codes a sequence of length $m \geq i$ and s is the i-th element of the sequence.

2. If $\boldsymbol{s} = \langle s_0, \dots, s_m\rangle$ and $\boldsymbol{t} = \langle t_0, \dots, t_n\rangle$ are two sequences, $\boldsymbol{s}^\frown\boldsymbol{t}$ denotes the concatenated sequence $\langle s_0, \dots, s_m, t_0, \dots, t_n\rangle$. If in addition $m = n$ then:

   (a) $\boldsymbol{s} = \boldsymbol{t}$ denotes $\bigwedge_{i\leq m}(s_i = t_i)$;

   (b) $\boldsymbol{s}[g]$ denotes the sequence $\langle s_0[g], \dots, s_m[g]\rangle$;

   (c) $\underset{\cdot}{F}_{x,u}(\boldsymbol{s})$ denotes the sequence $\langle \underset{\cdot}{F}_{x,u}(s_0), \dots, \underset{\cdot}{F}_{x,u}(s_0)\rangle$;

   (d) $\underset{\cdot}{d}(\boldsymbol{s}) \leq u$ denotes $\bigwedge_{i\leq m} \underset{\cdot}{d}(s_i) \leq u$.

3. "$s[g] = t$" is a formula expressing that either g is not an assignment and $s = t$, or g is an assignment and t is the result of replacing each occurrence of p^i_j in the $\mathcal{L}^+_A$-formula s by $g(p^i_j)$ if $g(p^i_j)$ is defined, and by ϵ if $g(p^i_j)$ is not defined.

4. "$\underset{\cdot}{F}_{x,n}(y) = z$" is a formula expressing that there exists: a sequence $\boldsymbol{\varphi}$ with code x; $\psi \in \Gamma(\boldsymbol{\varphi}, n)$ with code y, and a term $\underset{\cdot}{F}_{\boldsymbol{\varphi},n}(\psi)$ with code z. If there is no sequence of $\mathcal{L}_T$-formulas $\boldsymbol{\varphi}$ with code x then $y = z$.

Using n-th approximations, derivations in $(\mathsf{PA}^U)_\omega$ are replaced by approximations with bounded depth. Given a sequent:

$$\Gamma, T(\boldsymbol{s}) \Rightarrow \Delta, T(\boldsymbol{t}),$$

its *n-th approximation* is the sequent:

$$\Gamma, T(\underset{\cdot}{F}_{\boldsymbol{s}^\frown\boldsymbol{t},\underline{n}}\boldsymbol{s}) \Rightarrow \Delta, T(\underset{\cdot}{F}_{\boldsymbol{s}^\frown\boldsymbol{t},\underline{n}}\boldsymbol{t}).$$

We list the properties of $F_{\boldsymbol{\varphi},n}$ here for convenience. See (Leigh, 2015, Lemmata 12, 13).

Lemma 2. The following sequents are derivable in $\mathsf{I}\Delta_0 + \exp$.

$$\begin{aligned}
\emptyset &\Rightarrow (x \underset{\cdot}{\vee} y)[z] = (x[z] \underset{\cdot}{\vee} y[x]),\\
\emptyset &\Rightarrow (\underset{\cdot}{\neg} x)[z] = \underset{\cdot}{\neg}(x[z]),\\
\emptyset &\Rightarrow (\underset{\cdot}{\forall} xy)[z] = \underset{\cdot}{\forall} x(y[z]),\\
\emptyset &\Rightarrow (y(x/w))[z] = (y[z])(x/w),\\
(x)_i = y \underset{\cdot}{\vee} z &\Rightarrow \underset{\cdot}{F}_{x,w+1}(y \underset{\cdot}{\vee} z) = \underset{\cdot}{F}_{x,w+1}(y) \underset{\cdot}{\vee} \underset{\cdot}{F}_{x,w+1}(z),\\
(x)_i = \underset{\cdot}{\neg} y &\Rightarrow \underset{\cdot}{F}_{x,w+1}(\underset{\cdot}{\neg} y) = \underset{\cdot}{\neg}\underset{\cdot}{F}_{x,w+1}(y),\\
(x)_i = \underset{\cdot}{\forall} yz &\Rightarrow \underset{\cdot}{F}_{x,w+1}(\underset{\cdot}{\forall} yz) = \underset{\cdot}{\forall} y(\underset{\cdot}{F}_{x,w+1}(z)),\\
\emptyset &\Rightarrow \underset{\cdot}{F}_{x,w}(y_0{}^\frown y_1{}^\frown y_2) = \underset{\cdot}{F}_{x,w}(y_0{}^\frown y_2{}^\frown y_1),\\
\emptyset &\Rightarrow \underset{\cdot}{d}(\underset{\cdot}{F}_{x,z}(\boldsymbol{s})) \leq l\underset{\cdot}{h}(x) \cdot 2^z.
\end{aligned}$$

Lemma 3. There is a term g with variables w, x, y and z such that the following sequents are truth-free derivable in $\mathsf{I}\Delta_0 + \exp$.

$$\begin{aligned}
\emptyset &\Rightarrow \underset{\cdot}{d}(g) \leq \underset{\cdot}{lh}(x) \cdot 2^z, \\
y < z, w = x &\Rightarrow \underset{\cdot}{F}_{w,y}(u)[g] = \underset{\cdot}{F}_{x,z}(u), \\
y < z, x = x'^\frown(x_0 \underset{\cdot}{\vee} x_1), w = x'^\frown x_i &\Rightarrow \underset{\cdot}{F}_{w,y}(w)[g] = \underset{\cdot}{F}_{x,z}(w), \\
y < z, x = x'^\frown(\underset{\cdot}{\neg} x_0), w = x'^\frown x_0 &\Rightarrow \underset{\cdot}{F}_{w,y}(w)[g] = \underset{\cdot}{F}_{x,z}(w), \\
y < z, x = x'^\frown(\underset{\cdot}{\forall} x_0 x_1), w = x'^\frown x_1[u/x_2] &\Rightarrow \underset{\cdot}{F}_{w,y}(w)[g] \\
&\qquad = \underset{\cdot}{F}_{x,z}(x')^\frown \underset{\cdot}{F}_{x,z}(x_2)[u/x_1] \\
w = x^\frown w', \forall u(\underset{\cdot}{d}(\underset{\cdot}{F}_{w,y}(u)) \leq z) &\Rightarrow \underset{\cdot}{F}_{w,y}(x)[g] = \underset{\cdot}{F}_{x,z}(x).
\end{aligned}$$

Approximating derivations in $(\mathsf{PA}^U)_\omega$

By adapting the proof of the key lemma in (Leigh, 2015) we show that $(\mathsf{PA}^U)_\omega$ embeds into $(\mathsf{PA}^U)^*_\omega$.

Definition 15. Let d be a derivation in $(\mathsf{PA}^U)_\omega$ or $(\mathsf{PA}^U)^*_\omega$.

1. The *truth depth* of d is the maximum number of truth rules occurring in d.

2. The *truth rank* of d is $\sup\{k : (\mathrm{Cut}^k_T) \text{ occurs in } d\} + 1$.

3. The *rank* of d is any pair (n, r) such that n bounds the truth depth of d and r bounds the truth rank of d.

We state the following lemma that we shall use later on. The proof is identical to that of (Leigh, 2015, Lemma 15).

Lemma 4. Let Γ, Δ be sets consisting of arithmetical formulas, and $\boldsymbol{\varphi}, \boldsymbol{\psi}$ be sequences of terms. If the sequents $\Gamma, T(\boldsymbol{\varphi}) \Rightarrow \Delta, T(\boldsymbol{\psi})$ and $\Gamma \Rightarrow \underset{\cdot}{d}(g) < \underline{k}$ are derivable with truth ranks (a, r) and $(0, 0)$ respectively, then the sequent:

$$\Gamma, T(\boldsymbol{\varphi})[g] \Rightarrow \Delta, T(\boldsymbol{\psi})[g]$$

is derivable with truth rank $(a, r + k)$.

The following two lemmata provide the key ingredients for the proof of the version of the Bounding Lemma (Lemma 7) that we need for transforming derivations in $(\mathsf{PA}^U)_\omega$ to derivations in $(\mathsf{PA}^U)^*_\omega$. The proof of Lemma 5 below is identical to that of (Leigh, 2015, Lemma 18). In what follows, we write $H(k, n) = n \cdot 2^k$.

Lemma 5. Let $lh(\boldsymbol{\varphi}) + lh(\boldsymbol{\psi}) = n$ and suppose $r \leq H(k, n + 1)$. If the k-th approximations to:

$$\Gamma, T(\boldsymbol{\varphi}) \Rightarrow \Delta, T(\boldsymbol{\psi}), T(\chi)$$

and:

$$\Gamma, T(\boldsymbol{\varphi}), T(\chi) \Rightarrow \Delta, T(\boldsymbol{\psi})$$

are derivable with rank (a, r), then the $H(k, n + 1)$-th approximation of:

$$\Gamma, T(\boldsymbol{\varphi}) \Rightarrow \Delta, T(\boldsymbol{\psi})$$

is derivable with rank $(a+1, H(k, n+1) + H(H(k, n+1), n))$.

The presence of rule (Ind_T) in our sequent calculus means that we need an analog of lemma 5 for (Ind_T). This is the content of the following.

Lemma 6. Let $lh(\boldsymbol{\varphi}) + lh(\boldsymbol{\psi}) = n$. If the k-th approximation to:

$$\Gamma, T(\boldsymbol{\varphi}), T(\chi(\underline{x})) \Rightarrow \Delta, T(\boldsymbol{\psi}), T(\chi(\underline{x+1}))$$

is derivable with rank (a, r), then the $k+1$-th approximation of:

$$\Gamma, T(\boldsymbol{\varphi}), T(\chi(\underline{0})) \Rightarrow \Delta, T(\boldsymbol{\psi}), T(\chi(t))$$

is derivable with rank $(a+1, r + H(k+1, n+2))$.

Proof. Suppose that:

$$\Gamma, T(\underset{\cdot}{F}_{\boldsymbol{w},\underline{k}}\boldsymbol{\varphi}), T(\underset{\cdot}{F}_{\boldsymbol{w},\underline{k}}\chi(\underline{x})) \Rightarrow \Delta, T(\underset{\cdot}{F}_{\boldsymbol{w},\underline{k}}\boldsymbol{\psi}), T(\underset{\cdot}{F}_{\boldsymbol{w},\underline{k}}\chi(\underline{x+1}))$$

is derivable with rank (a, r), where $\boldsymbol{w} = \boldsymbol{\varphi}^\frown\boldsymbol{\psi}^\frown(\chi(\underline{x}))^\frown(\chi(\underline{x+1}))$. Let $g(x, y, z)$ be the term given by Lemma 3 and let:

$$g' = g(\boldsymbol{w}, \underline{k}, \underline{k+1}).$$

By Lemma 4, the sequent:

$$\Gamma, T(\underset{\cdot}{F}_{\boldsymbol{w}',\underline{k+1}}\boldsymbol{\varphi}), T(\underset{\cdot}{F}_{\boldsymbol{w},\underline{k}}\chi(\underline{x}))[g'] \Rightarrow \Delta, T(\underset{\cdot}{F}_{\boldsymbol{w}',\underline{k+1}}\boldsymbol{\psi}), T(\underset{\cdot}{F}_{\boldsymbol{w},\underline{k}}\chi(\underline{x+1}))[g']$$

is derivable with rank $(a, r+H(k+1, n+2))$, where $\boldsymbol{w}' = \boldsymbol{\varphi}^\frown\boldsymbol{\psi}^\frown(\chi(\underline{0}))^\frown(\chi(t))$. By Lemma 2 and Lemma 3, and using only arithmetical cuts, we obtain a derivation of the sequent:

$$\Gamma, T(\underset{.}{F}_{\boldsymbol{w}',\underline{k+1}}\boldsymbol{\varphi}), T(\underset{.}{F}_{\boldsymbol{w}',\underline{k+1}}(\chi)(\underline{x})) \Rightarrow \Delta, T(\underset{.}{F}_{\boldsymbol{w}',\underline{k+1}}\boldsymbol{\psi}), T(\underset{.}{F}_{\boldsymbol{w}',\underline{k+1}}(\chi)(\underline{x+1}))$$

with rank $(a, r+H(k+1, n+2))$. Lemma 2 and (Ind_T) then yield a derivation of the sequent:

$$\Gamma, T(\underset{.}{F}_{\boldsymbol{w}',\underline{k+1}}\boldsymbol{\varphi}), T(\underset{.}{F}_{\boldsymbol{w}',\underline{k+1}}\chi(\underline{0})) \Rightarrow \Delta, T(\underset{.}{F}_{\boldsymbol{w}',\underline{k+1}}\boldsymbol{\psi}), T(\underset{.}{F}_{\boldsymbol{w}',\underline{k+1}}\chi(t))$$

with rank $(a+1, r+H(k+1, n+2))$. □

The following lemma provides a reduction of $(\mathsf{PA}^U)_\omega$ to $(\mathsf{PA}^U)^*_\omega$. We extend the proof of the corresponding lemma in (Leigh, 2015) and show that the functions G_1, G_2 defined in (Leigh, 2015) also satisfy applications of the rule (Ind_T).

Lemma 7. (Bounding Lemma) There are recursive functions G_1 and G_2 such that for every $a, n < \omega$, if $lh(\boldsymbol{\varphi}) + lh(\boldsymbol{\psi}) \leq n$ and the sequent:

$$\Gamma, T(\boldsymbol{\varphi}) \Rightarrow \Delta, T(\boldsymbol{\psi})$$

is derivable in $(\mathsf{PA}^U)_\omega$ with truth depth a, then its $G_1(a, n)$-th approximation is derivable in $(\mathsf{PA}^U)^*_\omega$ with rank $(a, G_2(a, n))$.

Proof. Define:

$$G_1(0, n) = 0,$$
$$G_1(m+1, n) = H(G_1(m, n+1), n+1),$$
$$G_2(m, n) = G_1(m+1, m+n).$$

Notice that for all $a, b, n, m < \omega$: if $m < n$ then $G_1(a, m) \leq G_1(a, n)$; if $a < b$ then $G_1(a, n) \leq G_1(b, n)$; and $G_1(a, n+1) \leq G_1(a+1, n)$. The proof proceeds by induction on a.

Case 1: Suppose the sequent:

$$\Gamma, T(\boldsymbol{\varphi}) \Rightarrow \Delta, T(\boldsymbol{\psi})$$

was obtained by (Cut_T) applied to $T(\chi)$ and that this derivation has height $a+1$. The proof is as in (Leigh, 2015). Let $\boldsymbol{w}' = \boldsymbol{\varphi}^\frown\boldsymbol{\psi}$ and $\boldsymbol{w} = \boldsymbol{w}'^\frown\chi$. The induction hypothesis is that the $G_1(a, n+1)$-th approximations to the sequents:

$$\Gamma, T(\boldsymbol{\varphi}), T(\chi) \Rightarrow \Delta, T(\boldsymbol{\psi})$$

and:

$$\Gamma, T(\boldsymbol{\varphi}) \Rightarrow \Delta, T(\boldsymbol{\psi}), T(\chi)$$

are each derivable in $(\mathsf{PA}^U)^*_\omega$ with rank $(a, G_2(a, n+1))$. By Lemma 5 there is a derivation with height $a+1$ of the $G_1(a, n+1)$-th approximation to the sequent:

$$\Gamma, T(\boldsymbol{\varphi}) \Rightarrow \Delta, T(\boldsymbol{\psi}).$$

This derivation has cut rank $G_2(a, n+1) + H(H(G_1(a, n+1), n+1), n)$, so it's enough to show that:

$$G_2(a, n+1) + H(H(G_1(a, n+1), n+1), n) \leq G_2(a+1, n).$$

We have:

$$\begin{aligned} G_2(a+1, n) &= G_1(a+2, a+n+1) \\ &= H(G_1(a+1, a+n+2), a+n+2) \\ &= H(H(G_1(a, a+n+3), a+n+3), a+n+2). \end{aligned}$$

Consider $G_2(a, n+1)$. For all $a, n < \omega$ we have:

$$\begin{aligned} G_2(a, n+1) &= G_1(a+1, a+n+1) \\ &= H(G_1(a, a+n+2), a+n+2) \\ &\leq H(G_1(a, a+n+3), a+n+3) \\ &\leq H(H(G_1(a, a+n+3), a+n+3), 1). \end{aligned}$$

Now consider $H(H(G_1(a,n+1),n+1),n)$. For all $a,n<\omega$ we have:

$$\begin{aligned}H(H(G_1(a,n+1),n+1),n) &\leq H(H(G_1(a,a+n+3),a+n+3),n)\\ &\leq H(H(G_1(a,a+n+3),a+n+3),a+n+1).\end{aligned}$$

Adding $G_2(a,n+1)$ and $H(H(G_1(a,n+1),n+1),n)$ we obtain the desired inequality.

Case 2: suppose the sequent:

$$\Gamma, T(\boldsymbol{\varphi}), T(\chi(\underline{0})) \Rightarrow \Delta, T(\boldsymbol{\psi}), T(\chi(t))$$

was obtained by (Ind_T) applied to:

$$\Gamma, T(\boldsymbol{\varphi}), T(\chi(\underline{x})) \Rightarrow \Delta, T(\boldsymbol{\psi}), T(\chi(\underline{x+1}))$$

and that this derivation has height $a+1$. Let $\boldsymbol{w} = \boldsymbol{\varphi}^\frown\boldsymbol{\psi}^\frown\chi(\underline{x})^\frown\chi(\underline{x+1})$. The induction hypothesis is that the $G_1(a,n+2)$-th approximation to the sequent:

$$\Gamma, T(\boldsymbol{\varphi}), T(\chi(\underline{x})) \Rightarrow \Delta, T(\boldsymbol{\psi}), T(\chi(\underline{x+1}))$$

is derivable in $(\mathsf{PA}^U)^*_\omega$ with rank $(a, G_2(a,n+2))$. By Lemma 6 there is a derivation with height $a+1$ of the $G_1(a,n+2)+1$-th approximation to the sequent:

$$\Gamma, T(\boldsymbol{\varphi}), T(\chi(\underline{0})) \Rightarrow \Delta, T(\boldsymbol{\psi}), T(\chi(t))$$

This derivation has cut rank $G_2(a, n+2) + H(G_1(a, n+2)+1, n+2)$, so it's enough to show that:

$$G_2(a, n+2) + H(G_1(a, n+2)+1, n+2) \leq G_2(a+1, n).$$

Consider $G_2(a, n+2)$. For all $a, n < \omega$ we have:

$$\begin{aligned} G_2(a, n+2) &= G_1(a+1, a+n+2) \\ &= H(G_1(a, a+n+3), a+n+3) \\ &\leq H(H(G_1(a, a+n+3), a+n+3), 1). \end{aligned}$$

Now consider $H(G_1(a, n+2)+1, n+2)$. Notice that for all $a, n < \omega$ we have:

$$\begin{aligned} G_1(a, n+2) + 1 &\leq G_1(a, a+n+3) + 1 \\ &\leq H(G_1(a, a+n+3) + 1, 1) \\ &= H(G_1(a, a+n+3), 2) \\ &\leq H(G_1(a, a+n+3), a+n+3). \end{aligned}$$

Thus for all $a, n < \omega$ we have:

$$H(G_1(a, n+2)+1, n+2) \leq H(H(G_1(a, a+n+3), a+n+3), a+n+1),$$

whence adding $G_2(a, n+2)$ and $H(G_1(a, n+2)+1, n+2)$ yields the desired inequality.[25] □

The rest of the proof is then entirely similar to that in (Leigh, 2015). By Lemma 7, if the sequent $\emptyset \Rightarrow \varphi$ has a derivation within $(\mathsf{PA}^U)_\omega$, then it has a derivation within $(\mathsf{PA}^U)^*_\omega$. The bounding lemma extends to reduce the theory $(\mathsf{PA}^U)_\omega + \forall x(D(x) \to T(x))$ into a corresponding extension of $(\mathsf{PA}^U)^*_\omega$. Since $(\mathsf{PA}^U)^*_\omega$ is conservative over PA, $\Gamma \Rightarrow \Delta$ is derivable in PA.

Proof of Theorem 3. Let D and U be as in the statement of the theorem. Let d be a derivation with truth depth a of the truth-free sequent $\Gamma \Rightarrow \Delta$ in the system obtained from $(\mathsf{PA}^U)_\omega$ by adding the following rule:

$$\frac{\Gamma \Rightarrow \Delta, D(s)}{\Gamma \Rightarrow \Delta, T(s)}\ (\mathrm{D})$$

Redefine the functions G_1 and G_2 so that $G_1(0, n)$ bounds the logical depth of the finitely many formulas in U for each n. Then the proof of Lemma 7 can be carried out to obtain a derivation with rank $(a, G_2(a, 0))$ of $\Gamma \Rightarrow \Delta$ in the system obtained from $(\mathsf{PA}^U)^*_\omega$ by adding the following rule:

$$\frac{\Pi, T(\boldsymbol{\varphi}) \Rightarrow \Sigma, T(\boldsymbol{\psi}), D(\sigma)}{\Pi, T(\boldsymbol{\varphi}) \Rightarrow \Sigma, T(\boldsymbol{\psi}), T(\underset{\cdot}{F}_{\boldsymbol{w},\underline{k}}\sigma)}\ (\mathrm{D}_{\boldsymbol{w}})$$

where Π and Σ are truth-free, $k = G_1(a, 0)$ and $\boldsymbol{w} = \boldsymbol{\varphi}^\frown\boldsymbol{\psi}^\frown\sigma$. Notice that $G_1(a, 0) \geq G_1(0, n)$ for all $a, n < \omega$.

Call this derivation d^*. Fix n such that for each instance of $(\mathrm{D}_{\boldsymbol{w}})$ occurring in d^*, $lh(\boldsymbol{w}) < n$. It is enough to show that $(\mathsf{PA}^U)_\omega$ interprets $(\mathrm{D}_{\boldsymbol{w}})$.

[25] Notice that $n+2 \leq a+n+1$ whenever $a \geq 1$, so we may invoke monotonicity whenever $a \geq 1$; but the claimed inequality also holds whenever $a = 0$ and $n < \omega$ is arbitrary.

Let:

$$U^* = \{\varphi^* : \exists\psi \bigvee_{\varphi \in U} (\varphi^* = \varphi[\psi/p]) \wedge d(\varphi^*) \leq G_2(a,n)\}.$$

Then the sequent:

$$D(x), \underset{\cdot}{d}(x) < \underline{G_2(a,n)} \Rightarrow \{x = \ulcorner\varphi\urcorner : \varphi \in U^*\} \qquad (*)$$

is derivable in PA. Now, $G_1(0,n)$ bounds the logical depth of the schematic formulas in D, and $k = G_1(a,0) \geq G_1(0,n)$ for all $a, n < \omega$. Since every occurrence of a predicate symbol p^i_j in the k-th approximation of x has depth at least k in x, it follows that if x is any instance of the schema D, then so is $\underset{\cdot}{F}_{\boldsymbol{w},\underline{k}}x$. Moreover, this fact is derivable in PA. Since $\underset{\cdot}{d}(\underset{\cdot}{F}_{\boldsymbol{w},\underline{k}}(x)) < \underline{G_2(a,n)}$ is also derivable in PA, by $(*)$, the sequent:

$$D(x) \Rightarrow \underset{\cdot}{F}_{\boldsymbol{w},\underline{k}}x \in U^*.$$

is derivable in PA. Since the sequent $D(\sigma) \Rightarrow \sigma$ is derivable in PA for all arithmetical sentences σ, and the sequent $\sigma \Rightarrow T(\sigma)$ is derivable in $(\mathsf{PA}^U)_\omega$ for all arithmetical sentences σ, the sequent:

$$D(x) \Rightarrow T(\underset{\cdot}{F}_{\boldsymbol{w},\underline{k}}x)$$

is derivable in $(\mathsf{PA}^U)_\omega$. Thus $(\mathsf{PA}^U)_\omega$ interprets $(\mathrm{D}_{\boldsymbol{w}})$, and we obtain a derivation of the sequent $\Gamma \Rightarrow \Delta$ in $(\mathsf{PA}^U)^*_\omega$. Since $(\mathsf{PA}^U)^*_\omega$ is conservative over PA,

$\Gamma \Rightarrow \Delta$ is derivable in PA. □

2.5 Morals

Let us reflect on a few upshots of our observations so far. We set out at the beginning of chapter 2 to answer the following question: mathematically, what can implicit commitments of arithmetical theories consist of? We offered the following framework for analyzing one's implicit commitments in accepting a suitable arithmetic theory S:

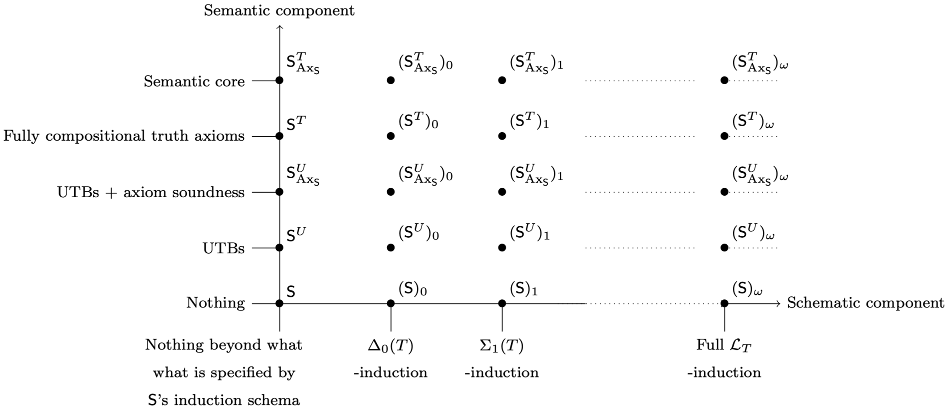
Semantic component
$\mathsf{S}^T_{\mathrm{Ax_S}}$
$(\mathsf{S}^T_{\mathrm{Ax_S}})_0$
$(\mathsf{S}^T_{\mathrm{Ax_S}})_1$
$(\mathsf{S}^T_{\mathrm{Ax_S}})_\omega$
Semantic core
S^T
$(\mathsf{S}^T)_0$
$(\mathsf{S}^T)_1$
$(\mathsf{S}^T)_\omega$
Fully compositional truth axioms
$\mathsf{S}^U_{\mathrm{Ax_S}}$
$(\mathsf{S}^U_{\mathrm{Ax_S}})_0$
$(\mathsf{S}^U_{\mathrm{Ax_S}})_1$
$(\mathsf{S}^U_{\mathrm{Ax_S}})_\omega$
UTBs + axiom soundness
S^U
$(\mathsf{S}^U)_0$
$(\mathsf{S}^U)_1$
$(\mathsf{S}^U)_\omega$
UTBs
S
$(\mathsf{S})_0$
$(\mathsf{S})_1$
$(\mathsf{S})_\omega$
Nothing
Schematic component
Nothing beyond what what is specified by S's induction schema
$\Delta_0(T)$ -induction
$\Sigma_1(T)$ -induction
Full $\mathcal{L}_T$ -induction

Figure 2.4: The semantic and schematic components of implicit commitment

There are two broad components of this framework: semantic and schematic, and each component admits fine-grained degrees. By suitably modifying the original notions of epistemic stability and the implicit commitment thesis as in (Dean, 2015), we hope to have offered a clear understanding which sets of implicit commitments are compatible with our weaker notions of epistemic stability and the implicit commitment thesis. We classified the theories of Figure 2.4 according to whether they are conservative over S. The conservative extensions reconcile the notion of epistemic stability with non-trivial versions of the implicit commitment thesis.

This brings us to our first upshot: at this point, we have an answer to our first central question of interest: *mathematically*, what are implicit commitments of theories of arithmetic? We have argued that mathematically, it makes sense to think of one's implicit commitments $\mathsf{I}(\mathsf{S})$ in justifiably believing S simply as any of the theories of Figure 2.4. That is, one's implicit commitments in justifiably believing S may include any of the following:

(1) Uniform disquotational truth principles for S.

(2) Fully compositional truth principles for S.

(3) Fully extended S-induction to the language $\mathcal{L}_T$.

(4) The axiom soundness principle for S.

The second upshot of our observations is that $\mathsf{I}(\mathsf{PA}) = (\mathsf{PA}^{U}_{\mathrm{Ax}_{\mathsf{PA}}})_\omega$ offers a clearer understanding of the sense in which the first-orderist differs from the finitist with respect to their views about extending induction. Earlier, we were worried that if the semantic core of PA is a fixed part of the first-orderist's implicit commitments in accepting PA, then actually, the first-orderist must occupy the same position with respect to the variable component of implicit schematic commitments in accepting PA as the finitist. Yet the first-orderist and the finitist really do seem to hold different views about extending induction.[26] But if instead $\mathsf{I}(\mathsf{PA}) = (\mathsf{PA}^{U}_{\mathrm{Ax}_{\mathsf{PA}}})_{\omega}$, then the first-orderist does occupy a different position with respect to their implicit schematic commitments in accepting PA as the finitist. The finitist adopts a trivial set of implicit schematic commitments when they accept PRA, and on the view we have set out, the first-orderist may count fully extended induction as part of their implicit schematic commitments when they accept PA. Thus, our understanding helps set the first-orderist apart from the finitist in this regard.

Admittedly, we are still not clear about the remark in (Nicolai & Piazza, 2019) that the first-orderist occupies an *intermediate* position about extending induction between the finitist and foundationalists à la Feferman. The implicit commitments of a foundationalist à la Feferman include *everything possible* – fully compositional truth, axiom soundness, and fully extended induction to the language $\mathcal{L}_T$. If the first-orderist's implicit commitments $\mathsf{I}(\mathsf{PA})$ are understood as the theory $(\mathsf{PA}^{U}_{\mathrm{Ax}_{\mathsf{PA}}})_{\omega}$, then purely with respect to their implicit schematic commitments, it seems that the first-orderist occupies the same position as the foundationalist à la Feferman. So perhaps we have just pushed the problem of intermediacy from one extreme to the other. However, nothing about what we've said hangs on this. The views of the first-orderist and the foundationalist à la Feferman may well overlap in this way. Our lingering worry stems only from the use of the term "intermediate."

[26] Cf. section 2.3.

The third upshot of the scenario we have advocated for is that the general idea of a fixed semantic core of implicit commitments in accepting a given base theory S is too strong. In particular, the requirement that the first-orderist be implicitly committed to fully compositional axioms for the truth predicate on the basis of their acceptance of PA is too strong. For it is precisely the presence of fully compositional truth principles which *forces* the first-orderist to give up *all* (sets of) instances of extended induction to the language $\mathcal{L}_T$ as part of their implicit commitments on the basis of their acceptance of PA.[27]

Again, this is related to our concerns about whether the first-orderist occupies the same position as the finitist with respect to their views about extending induction. We have suggested not only that these two foundationalists *do* hold different views about extending induction, but also that our understanding of the first-orderist's implicit commitments in accepting PA as the theory $(\mathsf{PA}^U_{\mathrm{Ax}_{\mathsf{PA}}})_\omega$ reflects this difference. Perhaps, though, there is a concern that our understanding of the first-orderist's implicit commitments in accepting PA as the theory $(\mathsf{PA}^U_{\mathrm{Ax}_{\mathsf{PA}}})_\omega$ does not really reflect the difference between the finitist and the first-orderist in this respect. The difference that emerged in section 2.3 was that on one hand, the finitist accepts *no* instances of PRA's induction schema in which the truth predicate occurs (whether on the basis of their acceptance of PRA or not). On the other hand, the first-orderist finds it possible to accept instances of PA's induction schema in which the truth predicate occurs, but strictly speaking, this acceptance does not follow from their acceptance of PA. On our understanding of the first-orderist's implicit

[27] Theorem 2 tells us this.

commitments in accepting PA as the theory $(\mathsf{PA}^{U}_{\mathrm{Ax}_{\mathsf{PA}}})_{\omega}$, one might complain that we have lost sight of this difference. For to say that the first-orderist's implicit commitments in accepting PA consist of the principles of the theory $(\mathsf{PA}^{U})^{*}_{\omega}$ *is just in part to say that* the first-orderist *does* accept all instances of fully extended induction for the language $\mathcal{L}_T$ on the basis of their acceptance of PA. Thus, while we have managed to separate the first-orderist from the finitist in some respect, we have lost sight of the real difference which was supposed to separate them.

We offer the following reply to this worry. By weakening the ICT in the way that we did, and by pursuing the strategy of isolating a set of implicit commitments in the extended language $\mathcal{L}_T$, *we cannot help but lose sight* of the idea that we have, in a sense, forced the first-orderist into a position whereby they accept statements beyond the logical reach of PA on the basis of their acceptance of PA. But this is true *no matter which set of implicit commitments we opt for.* The semantic core contains the fully compositional axioms for truth. These principles are $\mathcal{L}_T$-sentences, and so on a strict understanding of first-orderism, are such that while the first-orderist may come to accept those principles, the first-orderist's acceptance of those principles is not grounded in their acceptance of PA. Thus, if, on our understanding of the first-orderist's implicit commitments in accepting PA as the theory $(\mathsf{PA}^{U}_{\mathrm{Ax}_{\mathsf{PA}}})_{\omega}$, we have lost sight of the idea that forced the first-orderist into a position whereby they accept statements beyond the logical reach of PA on the basis of their acceptance of PA, then we have done exactly the same thing if instead we understand the first-orderist's implicit commitments in accepting PA as (for example) the semantic core of PA. To even *attempt* to articulate sets of implicit commitments on the basis of their acceptance of PA in the way that we have – in such a way to satisfy the weak ICT – we *must* relax the strict tenets of first-orderism.

But furthermore, we do not think there is any principled reason why we should require fully compositional truth, *rather than* extended induction, to form part of the first-orderist's implicit commitments on the basis of their acceptance of PA (or indeed vice versa). In the same vein as the previous remarks, we won't find any evidence in the tenets of first-orderism itself that favors one of these sets of principles to the other in this respect. For according to the strict tenets of first-orderism, acceptance of either set of principles does not follow from the first-orderist's acceptance of PA itself.

One objection to this line might be to claim that the theory $(\mathsf{PA}^{U}_{\mathrm{Ax}_{\mathsf{PA}}})_{\omega}$ is not a plausible theory of *truth* precisely *because* it lacks full compositionality,[28] and that this is a reason to prefer the first-orderist's implicit commitments on the basis of their acceptance of PA as the theory $\mathsf{PA}^{T}_{\mathrm{Ax}_{\mathsf{PA}}}$, rather than $(\mathsf{PA}^{U}_{\mathrm{Ax}_{\mathsf{PA}}})_{\omega}$. Essentially we think this misses the point. First, there is still a notion of uniform disquotational truth at play in the theory $(\mathsf{PA}^{U}_{\mathrm{Ax}_{\mathsf{PA}}})_{\omega}$, and since one of the underlying motivations for this project was to accommodate the assertion that all of the axioms of PA are true, we think uniform disquotational truth is enough to say we have achieved this much. But second, to say that the first-orderist's implicit commitments on the basis of accepting PA amount to the principles of $(\mathsf{PA}^{U}_{\mathrm{Ax}_{\mathsf{PA}}})_{\omega}$ is *not* to say that the first-orderist thereby *rejects* a fully compositional notion of truth. All that follows is that if the first-

[28] For example, such a view might align with defenders of a deflationary account of truth (Field, 1986, 1999; Horwich, 1990; Tennant, 2002).

orderist indeed accepts the idea that truth is fully compositional, then their acceptance of the corresponding principles is not grounded purely in their acceptance of PA. We maintain that there is no principled reason the first-orderist should prefer an implicit commitment to fully compositional truth, at the expense of an implicit commitment to extended induction, purely on the basis of acceptance of PA.

All things considered, we think the general idea of a fixed semantic core of implicit commitments in accepting a given base theory S is too strong. Rather, the more plausible mathematical understanding of implicit commitments is cashed out via the semantic and schematic components of our framework. We suggest that in general, neither component is fully fixed. If it makes sense to say that any of the principles we have considered are fixed implicit commitments in accepting a given theory S, we suggest that it is the common core of the theories $\mathsf{S}^T_{\mathrm{Ax}_\mathsf{S}}$ and $(\mathsf{S}^U_{\mathrm{Ax}_\mathsf{S}})_\omega$; that is, the theory $\mathsf{S}^U_{\mathrm{Ax}_\mathsf{S}}$. But we won't dwell on this. In any case, in general, sets of implicit commitments on the basis of accepting a given base theory S vary from foundationalist to foundationalist. Furthermore, it is not only possible to reconcile the notion of $\mathcal{L}_\mathsf{S}$-epistemic stability with the weak ICT for first-orderism, but this can be achieved with respect to a variety of implicit commitments, all of which contain the desired minimal soundness requirements for S.

Let us motivate our approach going forward. With the mathematics of implicit commitments better understood, we claim that we are now in a better position to address our second central question of interest: what is the *epistemological force* behind implicit commitments? Let us say how, and motivate our segue into chapter 3.

We have cashed out a mathematical conception of implicit commitments via the semantic and schematic components of our framework. Since neither component is fixed, this puts the first-orderist in an interesting and peculiar position. For on our mathematical conception of implicit commitments:

- In accepting the theory PA, the first-orderist is warranted in accepting the theory $\mathsf{PA}^{T}_{\mathrm{Ax}_{\mathsf{PA}}}$.
- In accepting PA, the first-orderist is warranted in accepting the theory $(\mathsf{PA}^{U}_{\mathrm{Ax}_{\mathsf{PA}}})_{\omega}$.
- But in accepting PA, the first-orderist is not warranted in accepting the union of the two theories above, the theory $(\mathsf{PA}^{T}_{\mathrm{Ax}_{\mathsf{PA}}})_{\omega}$.

This suggests a strategy for us. If we can say something about the epistemological force underlying this kind of warrant, we will be able to say something about the epistemological force underlying implicit commitments. And our observations have taught us something about this warrant: it must be such that it is not closed under conjunction in this way. So now we take up the task of trying to put our finger on what the warrant above might consist of, in a way that is responsive to our mathematical observations. Along the way, we also fix an epistemological understanding of *acceptance*.

Chapter 3

What the commitment of implicit commitments cannot be

Let $\mathsf{T} \supseteq \mathsf{S}$ be theories. Let:

$$\text{accept } \mathsf{S} \mapsto \text{ accept } \mathsf{T}$$

mean that in accepting the theory S, we are warranted by $\mapsto$ in accepting the extension T of S. Given what we learned in chapter 2, we are now interested in answering the following question: what could the warrant $\mapsto$ possibly consist in, such that we can make sense of the following scenario?

$$\begin{aligned}
\text{accept } \mathsf{PA} &\mapsto \text{ accept } \mathsf{PA}^{T}_{\mathrm{Ax}_{\mathsf{PA}}} \\
\text{accept } \mathsf{PA} &\mapsto \text{ accept } (\mathsf{PA}^{U}_{\mathrm{Ax}_{\mathsf{PA}}})_{\omega} \\
\text{accept } \mathsf{PA} &\not\mapsto \text{ accept } (\mathsf{PA}^{T}_{\mathrm{Ax}_{\mathsf{PA}}})_{\omega}
\end{aligned}$$

It will take us until the end of chapter 6 to formulate an answer to this question. To give a sense of where we are headed, over chapters 3 and 4, we survey the prospects for two kinds of warrant: *justification* and *entitlement*. We will be precise about our understandings of these terms, and in doing so, we will also fix a particular understanding of the broad notion of acceptance we have been using so far. In chapter 3 we argue that justification cannot stand in for the warrant $\mapsto$. We then introduce the notion of entitlements from (Wright, 2004). In fact, Fischer et al. (2021) have raised the suggestion that one of Wright's specific notion of entitlements, called *entitlements of cognitive projects*, are not closed under conjunction in the above sense. So perhaps entitlements of cognitive project are a promising route. In chapter 4, we argue that entitlements of cognitive project cannot stand in for the warrant $\mapsto$ either. However, we learn several valuable lessons from our discussion in chapter 4. Using these lessons, over chapters 5 and 6 we propose a new kind of warrant which we claim can stand in for $\mapsto$, and tie everything together.

3.1 No justification

First let us consider justification. In accepting PA, does it make sense to think that: (1) the first-orderist is warranted by justification in accepting $\mathsf{PA}^{T}_{\mathrm{Ax}_{\mathsf{PA}}}$, (2) the first-orderist is warranted by justification in accepting $(\mathsf{PA}^{U}_{\mathrm{Ax}_{\mathsf{PA}}})_{\omega}$, and (3) the first-orderist is not warranted by justification in accepting the $(\mathsf{PA}^{T}_{\mathrm{Ax}_{\mathsf{PA}}})_{\omega}$?

To answer these questions, we ought to be clear on what we mean by justification. However, this presents some challenges, so let us preface our discussion by highlighting those features of justification which we will have something to say about, and those features of justification which we will have nothing to say about. Our approach is as follows. We consider three kinds of justification below. One kind we refer to as "empirical justification." Two we refer to as kinds of "mathematical justification." We understand empirical justification as the sort of thing we acquire on the basis of our available worldly evidence. We understand mathematical justification as either: (1) the sort of thing inherited by certain families of foundational axioms in mathematics, in the sense of (Maddy, 1988a, 1988b), or (2) the sort of thing inherited by a mathematical sentence, which is derivable from a system of axioms justified in the sense of (1). Our goal is to draw out particular *features* of these kinds of justification, which render them incompatible with the first-orderist's scenario.

It is *not* our goal to say anything about the nature of these kinds of justifications, nor about the nature of justification more broadly. For example, we have nothing to say about whether these three kinds of justification exhaust all possible forms of justification (if such a list could even be given). Rather, we take the features drawn from our examples to be general enough to make our point, and so we rest content with what we discuss below.

All we will assume is that *some* kind of *force*, or *evidence*, underlies the kinds of justification we discuss (empirical, or mathematical). This (*whatever it is*) is what sets apart justification in either context from merely "taking for granted." But we have nothing to say about what the nature of this evidence *is*, in either the empirical context, or the mathematical context. Attempting to address these matters (if we could satisfactorily address them at all) would take us too far afield, and in any case, it is not part of the goal of this dissertation to clarify these matters. This is not a problem for us: we can say everything that we have to say while remaining silent on the nature of empirical/mathematical evidence, and leaving open questions about the nature of justification itself. So, this is what we shall do. With that, let us turn to our three kinds of justification.

First, consider justification in the following ordinary, empirical sense. Take any ordinary inference made on the basis of one's available empirical data. Perhaps, for example, I look next to my laptop and see a coffee cup there. Perhaps I take further steps to assure myself of the reliability of my empirical data. Perhaps I reach out and touch the coffee cup. Perhaps I check the lighting in my office. Perhaps I visit the optometrist and have my ocular faculties checked. The main feature of empirical justification understood this way we wish to highlight is that it *speaks to the likely truth of whatever it is that is being justified.* In the example, I understand all of my evidence *to speak to the likely truth* that there is a coffee cup next to my laptop. This is what I mean, when I say that I am *justified* in believing that there is a coffee cup next to my laptop.

So, naively, let us see what happens when we interpret the first-orderist's warrant as justification, in this ordinary, empirical sense. That is, suppose: (1) that in accepting PA, the first-orderist is warranted by ordinary empirical justification in accepting the theory $\mathsf{PA}^{T}_{\mathrm{Ax}_{\mathsf{PA}}}$, and (2) that in accepting PA, the first-orderist is warranted by ordinary empirical justification in accepting the theory $(\mathsf{PA}^{U}_{\mathrm{Ax}_{\mathsf{PA}}})_{\omega}$. On our current understanding of justification, this means that the first-orderist's acceptance of PA speaks to the likely truth of their acceptance of both the theory $(\mathsf{PA}^U_{\mathrm{Ax}_{\mathsf{PA}}})_\omega$ and the theory $\mathsf{PA}^T_{\mathrm{Ax}_{\mathsf{PA}}}$. In other words, if you were to ask the first-orderist whether they accept the theory $(\mathsf{PA}^U_{\mathrm{Ax}_{\mathsf{PA}}})_\omega$ on the basis of their acceptance of PA, you would expect them to say "yes." And similarly, if you were to ask the first-orderist whether they accept the theory $\mathsf{PA}^T_{\mathrm{Ax}_{\mathsf{PA}}}$ on the basis of their acceptance of PA, you would also expect them to say "yes."

It should be clear that this is a problem. On one hand, you would then expect the first-orderist to accept $(\mathsf{PA}^T_{\mathrm{Ax}_{\mathsf{PA}}})_\omega$ on the basis of their acceptance of PA. On the other hand, this is not at all what you would expect. For the first-orderist thinks PA is $\mathcal{L}_{\mathsf{PA}}$-epistemically stable. In particular, if you were to ask the first-orderist whether they accept $(\mathsf{PA}^T_{\mathrm{Ax}_{\mathsf{PA}}})_\omega$ on the basis of their acceptance of PA, you would expect them to say "no." The point is that justification, understood in this ordinary empirical sense, sets a high enough bar that it is closed under conjunction. So justification, understood in this ordinary empirical sense, cannot stand in for the first-orderist's warrant.

Next, to introduce our two kinds of mathematical justification: notice we said above that interpreting the first-orderist's warrant as justification in an ordinary empirical sense was naive. It was naive in the sense that it straightforwardly fails to make sense of what we were hoping to make sense of. But perhaps it was also naive to think that justification in an ordinary empirical sense is even fit for purpose here. For what is being justified is the first-orderist's acceptance of a *mathematical* theory. And it is far from clear that justification in an ordinary empirical sense behaves in the same way as *math-*

ematical justification.

For let us consider what it traditionally means for a (set of) mathematical sentence(s) to receive mathematical justification. There are a wealth of methods by which axioms are said to receive mathematical justification. For example:[1]

- Fragments of arithmetic, up to and including PA. For example:[2] Tait's (1981) thesis takes PRA to be justified on the basis of Hilbert's characterization of the natural numbers as finite sequences of symbols representable in intuition. Isaacson's (1996) thesis takes PA to be justified on the basis of a Dedekindian characterization of the natural numbers as the structure possessed by all infinite systems satisfying the axioms of second-order PA.[3]

- The axioms of ZF. Justifications derive mainly from two concepts: the idea of *limitation of size* and the *iterative conception of set* (Hallett, 1984). Limitation of size originated with (Cantor, 1883) (see (Hallett, 1984, Sec. 4)); its sentiments echoed, developed, and sometimes criticized by a long line of commentators (Bernays, 1946; Cantor, 1899/1967; Fraenkel, 1927, 1928; Fraenkel, Bar-Hillel, and Lévy, 1973; Gödel, 1944/1983a, 1947/1983b; Hessenberg, 1906; Jourdain, 1904, 1905; Lévy, 1979; Mirimanoff, 1917; Quine, 1969; Russell, 1907; Skolem, 1929; Wang, 1963; Weyl, 1949). The iterative conception descends from Fraenkel's (1927)

[1] We note that, in keeping with the general "mathematics first" approach of this dissertation, many of the traditional justificatory methods below are informed by a host of technical results. We have included these where appropriate.

[2] And as we have seen in chapter 2.

[3] See (Dean, 2015) for further discussion.

limitation of size argument (Hallett, 1984, Sec. 5.1), notably defended by Shoenfield (1967, 1977) and Wang (1974/1983). See also (Boolos, 1971/1983; Parsons, 1983; Scott, 1974).[4]

- The axiom of choice. Opinions ranged widely on the axiom of choice and its usage when it was first explicitly introduced by Zermelo (1904/1967b), and subsequently shown to imply that every set can be well-ordered (Zermelo, 1908/1967a). Hadamard (Baire et al., 1905/1982), Keyser (1905) and Hausdorff (1907) accepted a fully general version of the proof. Poincaré (1906) and Hardy (1906) accepted the axiom itself, but disputed the proof. Borel (Baire et al., 1905/1982) and Russell (1907) accepted restricted versions of the axiom. Peano (1890, 1902), Bettazzi (1892, 1896), Levi (1902), Lebesgue, Baire (Baire et al., 1905/1982), and Brouwer (1907/1975) rejected the axiom in at least one of its forms. Moore (1982) gives a detailed exposition of this history. However, over time, the indispensability of the axiom overrode any early misgivings. See (Maddy, 1988a) and (Moore, 1982).

- Determinacy axioms. For example, justifications for PD and $\mathsf{AD}^{L(\mathbb{R})}$ include plausible structural consequences (for the projective sets under PD, and for the sets of reals in $L(\mathbb{R})$ under $\mathsf{AD}^{L(\mathbb{R})}$) (Koellner, 2014; Maddy, 1988b; Moschovakis, 2009),[5] intertheoretic connections with large cardi-

[4] The iterative conception also has its critics in certain contexts. Feferman (2000) calls into question the idea that the axiom of power set does underlies the iterative conception. See also (Potter, 2004), who argues that some of the ZF axioms require alternative justifications.

[5] See the results of Mycielski and Świerczkowski (1964), Mazur and Banach (see (Mauldin, 1981, Problem 43) and Oxtoby (1957)), and Davis (1964) (in the context of PD). See also the results of Mycielski and Świerczkowski (1964), Mazur and Banach (see (Mauldin, 1981,

nal hypotheses (Koellner, 2006, 2014; Maddy, 1988b; Steel, 2000; Welch, 2015),[6] and that these axioms are implied by sufficiently strong natural theories (Koellner, 2006; Steel, 2000; Woodin, 2005).[7]

- Reflection principles. Justification is said to be *intrinsic*, in the sense of Gödel (1947/1983b). See e.g., (Koellner, 2009; Marshall, 1989; McCallum, 2021; Roberts, 2017; Tait, 2005; Welch, 2017),

- Small large cardinal axioms, e.g., those consistent with $V \neq L$. Such axioms maximize the universe of sets V by lengthening the class of ordinals, and capture in various senses the inexhaustibility of V (Gödel, 1947/1983b; Kanamori and Magidor, 1978; Koellner, 2009; Martin, 1976; McCallum, 2021; Solovay, Reinhardt, and Kanamori, 1978; Tait, 2005; Wang, 1974/1983), quoted in (Maddy, 1988a). Small large cardinal axioms are also justified to an extent by their plausible consequences (Cohen, 1971; Solovay, 1970).

- Medium large cardinal axioms. For example, measurable cardinals are justified by their plausible consequences (Maddy, 1988a, 1988b). One

Problem 43) and Oxtoby (1957)), and Martin and Steel in (Martin & Steel, 1983) (in the context of $\mathsf{AD}^{L(\mathbb{R})}$). Furthermore, the following string of results settled plausible uniformization, separation and reduction properties conjectured (Fenstad, 1971; Martin, 1976, 1977, 2012; Wang, 1974/1983) for projective sets: (Addison, 1958; Addison & Moschovakis, 1968; Blackwell, 1967; Kleene, 1950; Kondo, 1939; Kuratowski, 1936; Lusin, 1927, 1930; Lusin & Novikoff, 1935; Martin, 1968; Novikoff, 1931, 1935; Sierpinski, 1930). See (Kanamori, 1995; Maddy, 1988b; Moschovakis, 2009) for an overview of this history.

[6] For example, large cardinals are sufficient to prove versions of definable determinacy, and versions of definable determinacy implies the existence of inner models of those large cardinals. See the results in (Harrington, 1978; Martin, 1970; Martin & Steel, 1989) and the old result of Woodin recently published in (Müller, Schindler, & Woodin, 2020).

[7] See e.g., (Martin & Steel, 1989) and (Woodin, 1988). In fact this is the case *even when the theories are incompatible.* See the result of Woodin in (Schindler & Steel, 2014, Theorem 2.11.1), and (Steel, 2005).

is that $V \neq L$,[8] typically taken to be implausible (Drake, 1974; Gödel, 1944/1983a; Martin, 1976; Moschovakis, 2009; Scott, 1977; Wang, 1974/1983). Others include plausible structural properties.[9] Furthermore, the consistency of measurables is witnessed by canonical inner models that are very well-understood.[10]

- Large large cardinal axioms. The most common form of justification comes in the form of canonical inner models witnessing the existence of these cardinals. See (Neeman, 2002) for the state of the art of these results. See also (Woodin, 2017) for an overview of the progress in this area. See (Kanamori & Magidor, 1978; Reinhardt, 1974; Solovay et al., 1978) (quoted in (Maddy, 1988b)) for other forms of justification.

Justification in general has also been discussed in the context of Gödel's (1947/1983b) intrinsic/extrinsic distinction, a view analyzed, developed, and criticized in detail (Barton et al., 2020; Koellner, 2006, 2009; Maddy, 1988a, 1988b, 1990, 1997, 2011; Tait, 2001; Tiles, 2004). For general discussions, analyses, and history of axiom justifications, see e.g., (Drake, 1974; Fraenkel et al., 1973; Hallett, 1984; Maddy, 1988a, 1988b; Moore, 1982; Shoenfield, 1977).

The point of this literature survey is that factors influencing these traditional methods of mathematical justification include all sorts of pragmatic and aesthetic considerations. But what is *not* clear is that these factors *speak to*

[8] See the results in (Rowbottom, 1971; Scott, 1961; Silver, 1966; Solovay, 1967).

[9] See e.g., the results in (Solovay, 1969), also independently obtained a few months after by Mansfield (1970).

[10] See e.g., (Kunen, 1970), whose work was the culmination of Scott's (1961) result, beginning with the developments in (Gaifman, 1974). See also (Jensen, 2023).

the likely truth of the axioms they are said to justify. That is, mathematical justification, and justification in its ordinary empirical sense, seem quite different. If I say that I believe the axioms of ZF are true on the basis of the iterative conception of set, I mean something quite different to what I mean when I say that I believe there is a coffee cup next to my laptop on the basis of my empirical evidence. There seems to be little in the way of empirical evidence (whatever that is) which underlies the force (whatever that is) of mathematical justification. Thus, perhaps it was naive of us to even think that justification, understood in an ordinary empirical sense, even applies in the first-orderist's scenario.[11]

These observations are not necessarily obvious. Consider the following informal empirical proposition: the axioms of PA are consistent. This amounts to the proposition that it is impossible to derive a contradiction from the axioms of PA. Well, here is an empirical proposition, which looks like it functions as evidence for the preceding proposition in a perfectly ordinary, empirical way: so far, throughout history, no one has been able to successfully derive a contradiction from the axioms of PA. Suppose we agree that the following likelihood is small: so far, throughout history, no one has been able to successfully derive a contradiction from the axioms of PA, given that the axioms of PA are inconsistent. Then (e.g.) a Bayesian analysis reveals that "so far, throughout history, no one has been able to successfully derive a contradiction from the axioms of PA" functions as strong evidence for the proposition that the axioms of PA are consistent.

[11] To avoid a typing issue, we are equivocating here between the first-orderist's warrant for accepting certain principles, and the first-orderist's warrant for the principles themselves.

We take it that it is uncontroversial to think the former likelihood is indeed small. Mathematicians have devoted much time and energy to deriving statements from the axioms of PA, since the relevant tools became available. Some have even done so to *try* and obtain a contradiction from the axioms of PA, without success (Nelson, 1986). If the axioms of PA were inconsistent, this collective wealth of expert familiarity with PA, coupled with the fact that no one has yet discovered an inconsistency from the axioms of PA, would be extremely surprising. The overall point is that this example seems to be similar to a vast range of empirical examples of evidential justification. In particular: "so far, throughout history, no one has been able to successfully derive a contradiction from the axioms of PA" *does* speak to the likely truth of "the axioms of PA are consistent." But the natural formal counterpart to the informal, empirical assertion that the axioms of PA are consistent is the canonical consistency statement Con(PA). So perhaps there is a sense in which a formal mathematical statement is able to receive perfectly ordinary, empirical justification. Or, perhaps when we translate our informal consistency assertion to its formal counterpart, the empirical evidence justifying our informal consistency assertion is not inherited by its formal counterpart.

To avoid going down this rabbit hole, and so that we do not distract ourselves, we will not dwell on these issues here. As we move forward though, we will distinguish empirical justification from mathematical justification. We assume only that there is *some* kind of evidential force underlying both kinds of justification, even if it is not the same kind of force underlying both kinds of justification. We understand the force of empirical justification as empirical evidence, and the force of mathematical justification as mathematical evidence, whatever either kind of evidence might consist of. Whatever the force of mathematical evidence, we take it that all of our traditional methods of axiom justification carry it. Note that this does not mean we are thinking of mathematical evidence as something *out there*, and that over time we have figured out which methods of axiom justification carry this kind of force. It seems more natural to think that our traditional mathematical methods determine what mathematical evidence is, rather than the other way around (cf. (Maddy, 2011)). In any case, perhaps it was naive of us to think that justification in an ordinary empirical sense is the right type of warrant for making sense of the first-orderist's warrant, or perhaps not. Either way, justification in an ordinary empirical sense cannot stand in for the first-orderist's warrant.

Then what about mathematical justification (whatever its underlying force)? We would like to distinguish two understandings of this phrase. The first understanding aligns with our literature review above: a formal (set of) axiom(s) is said to be mathematically justified if we can provide mathematical evidence for it via any of the traditional methods above. We might think of our traditional methods of axiom justification as *independent* methods of axiom justification. The way in which we arrive at a mathematically justified belief in this or that set of axioms using our traditional methods, is independent of any of our existing beliefs about other sets of axioms.

The second understanding is this: we might think of a formal mathematical statement as being justified if we can *prove* it from a theory we take to be justified by any of our traditional methods of axiom justification. For example, if I take myself to hold a justified belief in the axioms of ZFC, which I arrived at using our traditional methods of axiom justification, then I might also take myself to also hold a justified belief in any theorem φ of ZFC simply by exhibiting a proof of φ from the axioms of ZFC. This second understanding of mathematical justification differs from the first understanding above. For the second understanding of mathematical justification is not independent in general from our existing beliefs about other sets of axioms. If I come to hold a justified belief in a theorem φ of ZFC based on my existing justified belief in the axioms of ZFC and a proof of φ from the axioms of ZFC, then my justified belief in φ *depends* on (at least) my existing justified belief in the axioms of ZFC.

We will say much more about this second understanding of mathematical justification in chapter 4. For now, we suggest that neither of these understandings of mathematical justification can stand in for the first-orderist's warrant. The first understanding of mathematical justification is simply not the right type of warrant that we are looking for. In the first-orderist's scenario, their warrant for accepting extensions of PA (hence for accepting implicit commitments of PA) *depends* on their acceptance of PA. We are not trying to make sense of a way in which the first-orderist may *independently* come to accept these extensions of PA. We are looking for a kind of warrant that makes sense of how the first-orderist may come to accept these extensions of PA *given that* they already accept PA itself. So justification, understood in the sense of traditional independent methods of axiom justification, is not the sort of warrant we are after.

Having said that, justification, understood in the sense of traditional independent methods of axiom justification, *does* help clarify one question left open in chapter 2. One thing we did not address in chapter 2 was what was meant epistemically by the first-orderist's "acceptance" of PA in the first place. In chapter 2, we used acceptance as an umbrella term, encompassing, for the time being, all kinds of epistemic attitudes. We are now in a position to fix an understanding of the first-orderist's acceptance of PA. For in chapter 2, we saw that first-orderism takes PA itself to be *justified* on the basis of a Dedekindian conception of the natural numbers. And this kind of justification is of the very sort we have just been considering: it is a traditional independent method of axiom justification. Thus, it makes sense to think of the first-orderist's acceptance (in the broad sense of chapter 2) of PA as a justified belief in the axioms of PA, where justified belief is understood exactly in the sense of traditional independent methods of axiom justification.

With this in mind, let us refine the question we have set out to investigate at this point. Let:

$$\text{justified belief in } \mathsf{S} \mapsto \ \text{justified belief in } \mathsf{T}$$

mean that on the basis of a justified belief in the theory S, we are warranted by $\mapsto$ in justifiably believing the extension T of S. Given what we learned in chapter 2, we are now interested in answering the following question: what could the warrant $\mapsto$ possibly consist in, such that we can make sense of the following scenario?

$$\begin{aligned} \text{justified belief in } \mathsf{PA} &\mapsto \text{justified belief in } \mathsf{PA}^{T}_{\mathrm{Ax}_{\mathsf{PA}}} \\ \text{justified belief in } \mathsf{PA} &\mapsto \text{justified belief in } (\mathsf{PA}^{U}_{\mathrm{Ax}_{\mathsf{PA}}})_{\omega} \\ \text{justified belief in } \mathsf{PA} &\not\mapsto \text{justified belief in } (\mathsf{PA}^{T}_{\mathrm{Ax}_{\mathsf{PA}}})_{\omega} \end{aligned}$$

What we said above is that it does not make sense to think of $\mapsto$ as mathematical justification in the sense of traditional independent methods of axiom justification. For we are not looking to tell a story about how the first-orderist arrives at an independently justified belief in the three theories on the right hand side. Rather, independent methods of axiom justification are what warrant the first-orderist's justified belief in PA on the left hand side. We want to make sense of how the same kind of epistemic attitude the first-orderist holds towards PA is inherited by the first-orderist's implicit commitments.

Furthermore, and straightforwardly, the second understanding of mathematical justification above cannot stand in for the first-orderist's warrant either. We might hope to make the following kind of argument: if the first-orderist holds a justified belief in the axioms of PA, and PA derives some consequence φ, then the first-orderist is warranted by mathematical justification in justifiably believing the extension of the collection of axioms of PA by φ. For then we might hope to tell an epistemological story about the first-orderist's implicit commitments: they are those principles which are justified in the sense of derivability. Obviously this will not work. The principles comprising the extensions of PA in the first-orderist's scenario are not even formulated in the language of PA. Nor can PA interpret the truth predicate uniformly. So the first-orderist's justified belief in these extensions of PA cannot be warranted by mathematical justification understood in the sense of derivability.

So, overall, we claim that justification is off the table. Empirical justification, even if it is the right *type* of warrant, fails to exhibit the kind of non-closure property we were looking for. And if empirical justification is not the right type of warrant, then we cannot hope to substitute it for $\mapsto$. Mathematical justification, understood as an independent method of axiom justification, is not what we are after. And mathematical justification understood in the sense of derivability cannot work. We must look beyond the idea of justification.

To motivate where else we are going to look, let us turn to the literature. For suitable S, various recent attempts have been made to identify the nature of the warrant that reflection principles for S receive, on the basis of accepting S itself (Fischer et al., 2021; Horsten, 2021; Horsten & Leigh, 2016; Łełyk & Nicolai, 2022; Nicolai & Piazza, 2019). In some of these accounts, the warrant for reflection is viewed through the lens of the distinction between the epistemic notions of entitlement and justification (Burge, 1993; Wright, 2004). For example, Horsten and Leigh (2016) argue that when we are justified in believing a theory, we are thereby entitled to adopt a corresponding reflection principle.[12]

[12] In other accounts, Horsten (2021) proposes an epistemological analysis of the process of reflection which assumes the consistency of the accepted theory, and argues that this vindicates the idea of epistemic stability. Łełyk and Nicolai (2022) propose an axiomatization of the minimal commitments implicit in the acceptance of a theory S, concluding that justified belief in the axioms of a theory are preserved to the corresponding reflection principle.

In particular, and perhaps most promising for us, Crispin Wright's specific notion of entitlements of cognitive project have been recently discussed in the context of rational theory acceptance (Fischer et al., 2021). The authors of that paper adopt the view that in accepting a theory S, one is entitled to accept a reflection principle for S. Furthermore, at the end of that paper, we are offered a suggestion about entitlements of cognitive project of just the sort that we are interested in; roughly, the suggestion that entitlements of cognitive project are not closed under conjunction.[13] So next, we will take up the question of whether entitlement of cognitive project can stand in for the first-orderist's warrant, and make sense of things for us.[14] Ultimately we argue that entitlements of cognitive project cannot stand in for the first-orderist's warrant, but it will take us a little while to say why. So at this point, we will leave open our problem about the nature of the first-orderist's warrant. We will spend the remaining pages of chapter 3 examining (in section 3.2) the context in which entitlements of cognitive project were introduced in (Wright, 2004), and extracting the salient features of this context for our purposes. Finally, in

[13] This is the reason motivating our choice of Wright's notions of entitlements of cognitive project, rather than some other notion(s) of epistemic warrant. For example, entitlements and justifications are also discussed by Tyler Burge (1993, 1998, 2003, 2013), although Burge's and Wright's views on the distinction between entitlements and justifications differ, and the authors have different goals in mind. See (Graham & Pedersen, 2020). See also (Casullo, 2007; Coliva, 2012; Graham & Pedersen, 2020; Majors, 2015; Neta, 2009; Silins, 2012; Wright, 2012) for recent commentary on both authors' views. We note also that the issue of whether the acceptability of the conjunction of two sets of principles such that, when considered by themselves, are acceptable, has been raised in the context of abstraction principles (Ebert & Rossberg, 2016). In the context of that discussion, any two abstraction principles that considered individually are judged to be admissible, are also jointly admissible. This is called *irenicity* (Cook, 2016; Shapiro & Uzquiano, 2016; Weir, 2003).

[14] Of course, we are trying to make sense of how the first-orderist may *avoid* accepting strong reflection principles. Really it is the suggested non-closure property of entitlements of cognitive project that makes them worth pursuing.

section 3.3, we examine the context in which entitlements of cognitive project are used in (Fischer et al., 2021). This will lead us in to chapter 4, where we point to the crucial problem with entitlements of cognitive project. In chapter 5 we will attempt to solve this problem for the context in (Fischer et al., 2021). Finally, we will return to the first-orderist's scenario in chapter 6, and apply a version of our solution in that context too.

3.2 Entitlements of cognitive project

Wright introduces entitlements of cognitive project in his (2004) paper. Entitlements of cognitive project are a particular kind of Wright's general notion of entitlements, so first we set out what the general notion of entitlements involves, before we narrow down our focus to entitlements of cognitive project.

3.2.1 Entitlements

Entitlements are introduced, as (what Wright calls) a unified solution to forms of skeptical paradox.[15] In particular, Wright offers a schematization of lines of reasoning which have led to skeptical conclusions of various sorts. Thus, it is the lines of reasoning which have led to these various skeptical conclusions which are unified.[16] The schema (or an instance of it) looks something like this:

[15] We are setting out to investigate how entitlements might function in a *mathematical* setting. If entitlements are supposed to solve some kind of skeptical paradox, you might be wondering what the *paradox* is, in the mathematical setting. We don't think there is one, but we'll address this later on.

[16] Wright calls this the *unified strategy* for a solution.

(I) My current experience is in all respects as if P.

(II) P.

(III) Q.

By substituting various propositions for P and Q above, we obtain lines of reasoning which form the basis for various kinds of skeptical conclusions.

We will look at two examples. First let P be "there is a hand in front of me" and Q be "there is an external world." Then we obtain a line of reasoning which forms the basis for what Wright calls a Humean skeptical paradox:[17]

(i) My current experience is in all respects as if there is a hand in front of me.

(ii) There is a hand in front of me.

(iii) There is an external world.

Let us spell out carefully how we are led to skeptical paradox from (i)–(iii). Here is the brief version of the argument: suppose the body of evidence (i) warrants proposition (ii). Proposition (ii) entails (perhaps together with other warranted premises) proposition (iii). So we might think that the body of evidence (i) also warrants proposition (iii). But (Wright thinks) it does not: this is phrased as the claim that the warrant we have for proposition (ii) is not *transmitted* to proposition (iii). In fact, for the body of evidence (i) to warrant proposition (ii) in the first place, we *already* have to have warrant for

[17] With P and Q instantiated as above, we get a Moore-type argument, which purports to give a proof an external world. Moore-type arguments were part of Wright's motivating context for entitlements, actually stretching back to his (1965). See also (Coliva, 2020).

proposition (iii). So: if the body of evidence (i) warrants proposition (ii), then the body of evidence (i) does not warrant proposition (iii). But even worse: we could not possibly obtain *any* other warrant for proposition (iii), other than the kind of evidence in (i). So, if the body of evidence (i) warrants proposition (ii), then we don't have any warrant for (iii). And without warrant for proposition (iii), we don't have any warrant for proposition (ii) (or any propositions of a similar kind).[18] This contradicts the first premise, that the body of evidence (i) warrants proposition (ii). Overall, we are led to conclude that we have no warrant for proposition (ii) (or for any proposition of a similar kind) after all.

Let us elaborate on these steps. The body of evidence (i) warrants proposition (ii) *inductively*. Our warrant for (ii) is grounded in a body of evidence, construed to include observations made on the basis of our usual range of empirical methods of investigation.[19]

Now, proposition (ii) (perhaps together with other warranted premises) implies proposition (iii). But Wright does not think that the evidence (i) provides for (ii) transmits to (iii). This is because he thinks the evidence (i) provides for (ii) (via inductive inference) is grounded in a broader informational context, in which we find proposition (iii) itself. In particular, the warrant (i) provides for (ii) is *information-dependent*, in Wright's (2002) sense, and the information on which this warrant depends is just the sort of information included in proposition (iii).

Let us explain what this means. A body of evidence e is an information-dependent warrant for a proposition p if regarding e as warranting p rationally

[18] This is a closure property.

[19] We could also extend "evidence" to include a priori reflection. See (Coliva, 2020).

requires certain kinds of collateral information (Wright, 2002, pp. 335–336). For instance: I am hiking in a national park and I hear the sound of a large creature growling and rooting around in the forest. Is this evidence of a bear? Yes, if I know I am in Sequoia National Park in California, where bears live. No, if I know I am in the Peak District National Park in England, where there are no bears. My warrant for believing there is a bear nearby depends on collateral information concerning my geographical location. In particular, if my evidence *does* warrant my believing that there is a bear nearby, I require collateral warrant for the proposition which states I am currently in a place where bears live.

Of course, this is not necessarily a problem, in and of itself. But Wright thinks that a problem does occur when the proposition for which one requires collateral warrant, is itself entailed by the proposition warranted by our body of evidence in the first place. Failure of warrant transmission occurs in cases like these. In particular, Wright thinks failure of warrant transmission occurs in (i)–(iii) above. My warrant for proposition (ii) depends on certain kinds of collateral information. But the kind of collateral information on which this depends includes information about whether there is an external world. If there is *no* external world, my current experience *cannot* warrant my belief that there is a hand in front of me. If there *is* an external world, my current experience *does* warrant my belief that there is a hand in front of me. So, my warrant for proposition (ii) already requires collateral warrant for proposition (iii). But proposition (ii) (perhaps together with other warranted premises) implies proposition (iii), and so the warrant for (ii) fails to transmit to (iii).

At this point, if the body of evidence (i) warrants proposition (ii), then the body of evidence (i) does not warrant proposition (iii). But after some reflection, we concede that there is *no* other way of acquiring a warrant for (iii). If (iii) is not warranted by our current experience, the idea is supposed to be that we have *no idea* how to provide a warrant for (iii) at all. And without warrant for proposition (iii), we don't have any warrant for proposition (ii) (or any propositions of a similar kind). Contradiction. So the body of evidence (i) does not warrant proposition (ii) after all, and overall, we are led to conclude that we have no warrant for proposition (ii) (or for any proposition of a similar kind).

Let us look at a second example. Let P be "there is really a hand in front of me" and Q be "I am not in the midst of a lucid and persistent dream." Then we obtain a line of reasoning which forms the basis for what Wright calls a Cartesian skeptical paradox:

(i′) My current experience is in all respects as if there is a hand in front of me.

(ii′) There really is a hand in front of me.

(iii′) I am not now in the midst of a lucid and persistent dream.

We are led to skeptical paradox from (i′)–(iii′) in a similar way as before. The body of evidence (i′) warrants proposition (ii′) inductively. But the warrant (i′) provides for (ii′) is information-dependent, and the information on which this warrant depends is just the sort of information included in proposition (iii′). If I *am* now in the midst of a lucid and persistent dream, my current experience *cannot* warrant my believing there really is a hand in front of me. If I am *not* now in the midst of a lucid and persistent dream, my current experience *does* warrant my believing there really is a hand in front of me.

So, my warrant for proposition (ii′) already requires collateral warrant for proposition (iii′). But proposition (ii′) (perhaps together with other warranted premises) implies proposition (iii′), and so the warrant for (ii′) fails to transmit to (iii′). Thus, the collateral warrant for proposition (iii′) cannot be underwritten by the body of evidence (i′). Again, after some reflection, we concede that there is *no* other way of acquiring a warrant for (iii′). If (iii′) is not warranted by our current experience, the idea is supposed to be that we have *no idea* how to provide a warrant for (iii′) at all. And without warrant for proposition (iii′), we don't have any warrant for proposition (ii′) (or any propositions of a similar kind). Contradiction. So the body of evidence (i) does not warrant proposition (ii) after all, and overall, we are led to conclude that we have no warrant for proposition (ii) (or for any proposition of a similar kind).

In the two examples above, propositions like (iii) and (iii′) share something in common: they are such that a lack of warrant for those propositions entails a lack of warrant for a large class of our beliefs in some region of thought (the sort typified by (ii) and (ii′) above). As a result, Wright calls propositions like (iii) and (iii′) *cornerstones* for large classes of our beliefs. The general line of skeptical reasoning makes the case that certain propositions are in fact cornerstones for broad classes of our beliefs, yet (on pain of circularity) there is no warrant for those cornerstones. Thus, there is no warrant for broad classes of our beliefs.

Not only do the lines of reasoning (i)–(iii) and (i′)–(iii′) share general structural similarities, but the examples also serve to show that we are led from (i)–(iii) and (i′)–(iii′) to a skeptical conclusion in the same kind of way. Call general propositions of the form (I) above (of which (i) and (i′) above are instances) *type*-I propositions, and similarly for general propositions of the form (II) and (III) above.[20] Wright draws out the similarities in the arguments resulting in skeptical conclusions by way of the following schematization:

> (i) Type-II propositions can only be justified on the evidence of (by inductive inference) type-I propositions.
>
> (ii) The evidence provided by type-I propositions for type-II propositions is information-dependent, requiring (among other things) collateral warrant for a type-III proposition.
>
> (iii) So: type-III propositions cannot be warranted by transmission of evidence provided by type-I propositions for type-II propositions across a type-II to type-III entailment – rather it is only if one already has warrant for the type-III proposition that any type-II propositions can be justified in the first place.
>
> (iv) Type-III propositions cannot be warranted any other way.
>
> If all four propositions are accepted, then type-III propositions are cornerstones for type-II propositions (thesis ii) which cannot themselves be warranted (theses iii and iv). So

[20] We understand "experience" in type-I propositions broadly. Cf. the examples in (Wright, 2004, p. 171) concerning observations about others' behavior and physical condition, and an agent's memory.

(v) There is no warrant for any type-II propositions. (Wright, 2004, p. 172)

Unifying both the (I)–(III) lines of reasoning that form the basis for skeptical arguments, and also the structure of the skeptical arguments themselves, sets the stage for Wright's unified solution. (i)–(v) above reveals what Wright calls the common lacuna between various instances of skeptical reasoning. On one hand, the line of reasoning (i)–(v) establishes at most this: that evidence for a cornerstone cannot be acquired by any justificatory process (thesis (iii)). On the other hand, what is needed for the skeptical conclusion is this: that there is no warrant for a cornerstone. Thus, there is a gap between thesis (iii) and the skeptical conclusion (v) above. Entitlements are then introduced, as an alternative kind of warrant for type-III propositions, and thus as a way to block the skeptical move from (iii) to (v) by making thesis (iv) false:

> Suppose there is a type of rational warrant which one does not have to *do any specific evidential work* to earn: better, a type of rational warrant whose possession does not require the existence of evidence – in the broadest sense, encompassing both *a priori* and empirical considerations – for the truth of the warranted proposition. Call it *entitlement*. (Wright, 2004, pp. 174–175).

Unfortunately, a general notion of entitlements will not solve the problem uniformly, because the content of the type-III proposition in an instance of the line of reasoning (I)-(III) might differ from instance to instance. For example, in the Humean example (i)–(iii) above, the content of proposition (iii) concerns ontology. We are interested in giving some sort of warrant for material objects. On the other hand, in the Cartesian example (i$'$)–(iii$'$) above, the content of proposition (iii$'$) concerns cognitive dislocation; it is a presupposition about pure enquiry. There, we are interested in giving some sort of warrant for the reliability of our methods of cognitive enquiry.

As a result of all these differences, we are introduced to four species of entitlement: strategic entitlements, entitlements of cognitive project, entitlements of rational deliberation, and entitlements of substance. Strategic entitlements and entitlements of cognitive project aim to cure skeptical paradoxes about the reliability of our methods of cognitive enquiry. Entitlements of rational deliberation aim to cure skeptical paradoxes which claim that we cannot rationally choose between alternatives. Entitlements of substance aim to cure skeptical paradoxes about existence. Investigations of each kind of entitlement, and of whether various kinds of entitlement do in fact alleviate the kind of skeptical worries they were introduced for, have been given both by (Wright, 2004) and subsequently elsewhere in a number of places.[21] We will focus on entitlements of cognitive project, and we leave the other three kind of entitlement introduced by Wright to one side here.[22] So next let us say what entitlements of cognitive project are.

[21] See e.g., (Coliva, 2015, 2020; Moretti, 2021; Pederson, 2009).

[22] We note that in particular, entitlements of cognitive project have raised a number of concerns (see e.g., (Coliva, 2020)). But we omit discussion of these related concerns here, since our goal is to argue that there is *no* reason to think that entitlements of cognitive project are well-suited in general for contexts of rational theory acceptance.

3.2.2 Entitlements of cognitive project

Entitlements of cognitive project exist as a relation between cornerstone propositions and *cognitive projects*. A *cognitive project* is understood to be some kind of cognitive undertaking, resulting in the achievement of a cognitive goal (Pederson, 2009, p. 445). For example, I might be interested in arriving at a belief about the width of my desk (my cognitive goal), and I might come to believe that it is one and a half meters wide when I measure it and observe the result (my undertaking of the project, and resulting cognitive achievement). To define an entitlement of a particular cognitive project, we are invited to consider any cognitive project, the failure of which would not be worse than the costs implied by not undertaking it, and the success of which would be better (Wright, 2004, p. 192).

Before we give Wright's clauses for a proposition's being an entitlement of some cognitive project, let us comment on the latter condition on the cognitive project itself. This condition seems to function as a sort of criterion for the kinds of cognitive projects we will be discussing: we are interested in cognitive projects which are in some sense *worth* undertaking. But what makes a cognitive project worth undertaking? it seems difficult to draw a precise line between kinds of cognitive projects which satisfy this criterion, and kinds of cognitive projects which do not. For example, suppose I set out to see what follows from an inconsistent set of sentences. Broadly speaking, there seems to be little value in this undertaking. But in certain contexts there does seem to be *something* valuable about it. Suppose I am taking a course on propositional logic, in which the notion of inconsistency is defined as "derives $\bot$."

Suppose I set out to see what sentences can be derived from the set $\{\varphi, \neg\varphi\}$. The cognitive achievement at the end of my undertaking is my realization that *every* sentence in the language of propositional logic can be derived from $\{\varphi, \neg\varphi\}$. Implicit in my cognitive achievement is my realization that the defined notion of inconsistency is equivalent to the notion that anything follows from an inconsistent set of sentences. Thus, I have learned something. So in this restricted context, there is some value in my cognitive undertaking.

In any case, we will leave this issue here. We are happy to adopt a very general understanding of cognitive projects, whose worthiness of investigation might be determined by various factors. For example, the goals of cognitive projects which are sufficiently worthy of investigation may be influenced by pragmatic considerations, or (aligning with the context of chapters 5 and 6) influenced by the aims of mathematical practice. In what follows (particularly in chapters 5 and 6), we take it that the goals of the cognitive projects we define are suitably broad, so as to qualify as worthy of investigation in Wright's sense. Thus, in defining these cognitive projects, we take it that we have not stretched the idea too thin. With that, let us say what an entitlement of cognitive project is.

First, a preliminary definition. We say that a proposition p is a *presupposition* of a particular cognitive project c if to doubt p would rationally commit one to doubting the significance or competence of the project c (Wright, 2004, p. 191). Then:

Definition 16. An *entitlement of cognitive project c* is any p satisfying the following conditions (Wright, 2004, pp. 191–192):

1. p is a presupposition of c.

2. We have no sufficient reason to believe that p is untrue.

3. The attempt to justify p would involve further presuppositions in turn of no more secure a prior understanding... and so on without limit; so that someone pursuing the relevant enquiry who accepted that there is nevertheless an onus to justify p would implicitly undertake a commitment to an infinite regress of justificatory projects, each concerned to vindicate the presuppositions of its predecessors.

The general idea is that entitlements of cognitive project allow us to rationally place trust, without evidence, in presuppositions of rational enquiry. The kind of presupposition we are allowed to rationally place trust in, without evidence, are those which: (1) believing the results of a cognitive enquiry rationally requires us not to doubt, and (2) are beyond vindication by evidence, except at the cost of further presuppositions of the same kind. The "trust," which entitlements of cognitive project allow us to rationally place in certain presuppositions of rational enquiry, is the sort of thing capable of underwriting some sort of rational belief in the achievements of (a successful execution of) the cognitive project (Wright, 2004, p. 193).

The idea behind clause 1 is that in the context of the cognitive project c, p is an unavoidable commitment, in which we must rationally hold a *positive* epistemic attitude. The notion of doubt in clause 1 is weak, understood not only to be compatible with a positive attitude towards the negation of p but also to be compatible with agnosticism about p, where one holds no positive attitude towards p or its negation (Pederson, 2009, p. 445). So clause 1 states that the success of c rationally requires us to hold a stronger-than-agnostic epistemic attitude towards p. In particular, the notion of doubt *excludes* the sort of belief underwritten by entitlements of cognitive project.[23]

Clause 2 reflects the non-evidential nature of entitlements of cognitive project. Positive evidence for the truth of p is not required. Only a *lack* of sufficient countervailing evidence matters. This is what makes entitlements of cognitive project fundamentally non-evidential. In particular, entitlements of cognitive project do not speak to the likely truth of the propositions warranted by them. As a result, notice that the sort of rational belief underwritten by an entitlement of cognitive project cannot be *justified* belief. For *justified* belief is the sort of belief which requires evidence as an input. This aligns with our remarks in section 3.2. On the other hand, beliefs formed on the basis of entitlement of cognitive project have no evidence as input. This idea will be of central importance to us in chapter 4. Notice also that since the doubt excludes the sort of belief underwritten by entitlements of cognitive project, doubt also excludes the stronger notion of justified belief.

Finally, we point out that the idea of infinite justificatory regress in clause 3 requires three independent conditions are met. We reach the first stage of infinite justificatory regress in the sense of clause 3 just in case any attempt to justify p is such that: (1) in making that attempt, we end up relying on

[23] That is, if I doubt p, then I do not believe p, where belief is the sort underwritten by entitlements of cognitive project. Notice we are *not* claiming that if I doubt p, then I believe $\neg p$, nor are we claiming that if I do not believe p, then I doubt p. This general relationship between doubt and belief is consistent with existing independent discussions of these notions (Meadows, 2021).

another *presupposition* q of c, (2) q is of no more secure a prior understanding than p itself, and (3) if we accept that there is nevertheless an onus to justify p, we would implicitly undertake a commitment to justify q. Later on, we will largely be concerned with cases where condition (1) comes apart from conditions (2) and (3).

Let us get some examples on the table. We will look at three, discussed across (Wright, 2004) and (Wright, 2005).

The first kind of example is an entitlement to the proper functioning of certain of our cognitive faculties, and returns us to the Cartesian scenario above. In particular, the cornerstone (type-III) proposition “I am not now in the midst of a lucid and persistent dream” is an entitlement of *any* cognitive project involving perceptual interaction with the world, on the basis of which I am trying to form some sort of belief (like my belief that there really is a hand in front of me). There are three underlying ideas. (1) Doubting the proposition “I am not now in the midst of a lucid and persistent dream,” would rationally commit me to doubt the significance or competence of any such cognitive project. If I were agnostic or worse about whether I am not now in the midst of a lucid and persistent dream, I could not rationally maintain that I am also able to *believe* there really is a hand in front of me. (2) I have no sufficient reason to suppose the proposition “I am not now in the midst of a lucid and persistent dream” is false. Perhaps, for example, I look around and see nothing I would consider to be out of the ordinary, which would otherwise result in my thinking that I might be dreaming. (3) Any attempt to justify the proposition “I am not now in the midst of a lucid and persistent dream” would involve a further presupposition of the cognitive project I am currently engaged in, of no more secure a prior understanding than the proposition "I am not now in the midst of a lucid and persistent dream" itself, such that if I were to accept an onus to justify "I am not now in the midst of a lucid and persistent dream," I would implicitly undertake a commitment to a series of justificatory projects, each concerned to vindicate the presuppositions of its predecessor. Let us try and sketch the first stage of such a regress.

Suppose I attempt to justify that I am not now in the midst of a lucid and persistent dream by pinching myself and noticing that I do not wake up to find myself having been lying in bed asleep. Well, if my resultant experience upon pinching myself functions as a proper justification of "I am not now in the midst of a lucid and persistent dream," it had better be the case that I did not merely *dream* that I pinched myself and noticed that I did not wake up to find myself having been lying in bed asleep.

But now I have hit upon a further presupposition of my cognitive project. If I am not now in the midst of a lucid and persistent dream, then when I pinched myself and noticed that I did not wake up to find myself having been lying in bed asleep, I must be sure that I did not merely dream that I pinched myself and noticed that. So if I were to doubt that I did *not* merely *dream* that I pinched myself and noticed that I did not wake up to find myself having been lying in bed asleep, I would also doubt that I am not now in the midst of a lucid and persistent dream. And doubting the latter rationally commits me to doubting the significance or competence of my cognitive project. So, "I did not merely dream that I pinched myself and noticed that I did not wake up to find myself having been lying in bed asleep" is a presupposition of my cognitive project.

Furthermore, it does not seem like my understanding of "I did not merely dream that I pinched myself and noticed that I did not wake up to find myself having been lying in bed asleep" is any more secure than my understanding of "I am not now in the midst of a lucid and persistent dream." And if I were to accept an onus to justify "I am not now in the midst of a lucid and persistent dream" in this way, it seems right to think that I would thereby undertake a commitment to justify "I did not merely dream that I pinched myself and noticed that I did not wake up to find myself having been lying in bed asleep." If I were to *deny* that I had to justify the latter, it hardly seems like I have taken my onus to justify 'I am not now in the midst of a lucid and persistent dream" seriously. So, I seem to have reached the first stage of justificatory regress. If I were to carry on in this way, my attempt to justify 'I am not now in the midst of a lucid and persistent dream" looks infinitely regressive in Wright's sense.

The second kind of example is an entitlement to the co-operativeness of the prevailing circumstances in the successful operation of certain of our cognitive faculties. Consider again any cognitive project involving perceptual interaction with the world, on the basis of which I am trying to form some sort of belief. Perhaps I look across the street (my perceptual interaction with the world) and come to believe that there is a dog resting on my neighbor's porch (my cognitive achievement).

What I am entitled to trust in the context of this particular cognitive project is that the prevailing circumstances are not such that I am led astray in my belief formation – for instance, that the creature I can see resting on my neighbor's porch is not a cat cleverly disguised to look like a dog. Then the proposition "the creature I can see resting on my neighbor's porch is not a cat cleverly disguised to look like a dog" is a presupposition of my cognitive project. If I were agnostic or worse about whether the creature I can see resting on my neighbor's porch is not a cat cleverly disguised to look like a dog, I could not rationally maintain that I am also able to believe that there is a dog resting on my neighbor's porch. Successful execution of my cognitive project depends on the co-operation of these prevailing circumstances. Furthermore, suppose I strain my ears and listen for the sound of a cat, but to no avail. Then I may suppose I have no sufficient reason to believe the creature I can see resting on my neighbor's porch *is* a cat cleverly disguised to look like a dog. Finally, suppose I were to attempt to justify that the creature I can see resting on my neighbor's porch is not a cat cleverly disguised to look like a dog by investigating my environment for dogs disguised as cats. But if this investigation holds up, I have to take it that the prevailing circumstances under which I carried out my investigation were also co-operative. So in my justificatory attempt, I have to rely on a further presupposition of my cognitive project in turn of no more secure a prior understanding. And as before, if I were to sincerely accept an onus to justify "the creature I can see resting on my neighbor's porch is not a cat cleverly disguised to look like a dog," I would thereby implicitly undertake a commitment to justify this further presupposition. If I were to continue in this way, my attempt to justify "the creature I can see resting on my neighbor's porch is not a cat cleverly disguised to look like a dog" looks infinitely regressive in Wright's sense.

The third kind of example is an entitlement to rely on the validity of basic inferential rules. This example is much closer in kind to the examples we will be considering shortly, so we will pay close attention to it. Consider any cognitive project which involves the use of my intellectual capacities towards some cognitive achievement. We focus on the case of modus ponens. For example, I am reasoning using some object language, and I come to believe that a conclusion ψ is true whenever two premises φ and $\varphi \to \psi$ are true (the cognitive achievement), by deriving ψ from φ and $\varphi \to \psi$ by modus ponens (the use of my intellectual capacities). Then I am entitled in the context of such a project to rely on the validity of the corresponding instance of modus ponens. That is, I am entitled in the context of such a project to rely on the following: $\varphi, \varphi \to \psi \models \psi$.

First, $\varphi, \varphi \to \psi \models \psi$ is a presupposition of my cognitive project. For suppose I were to doubt that $\varphi, \varphi \to \psi \models \psi$. In the context of my project, this means that I doubt that every model $\mathcal{M}$ such that $\mathcal{M} \models \varphi$ and $\mathcal{M} \models \varphi \to \psi$ is also such that $\mathcal{M} \models \psi$. On the other hand, my cognitive goal is to believe that ψ is true whenever φ and $\varphi \to \psi$ are true. That is, my cognitive goal is to believe that every model $\mathcal{M}$ such that $\mathcal{M} \models \varphi$ and $\mathcal{M} \models \varphi \to \psi$ is also such that $\mathcal{M} \models \psi$. But doubt excludes belief, so I cannot rationally doubt and believe this simultaneously.

Let us suppose further that I also have no sufficient reason to believe that the relevant instance of modus ponens is invalid (perhaps I assure myself that

I have not affirmed the consequent during my reasoning). We are particularly interested in what an attempt to justify $\varphi, \varphi \to \psi \models \psi$ looks like, and how it leads to infinite justificatory regress in Wright's sense. We will reconstruct several other examples along these lines in chapters 4 and 5. So, suppose I attempt to justify $\varphi, \varphi \to \psi \models \psi$.

First, I fix a model $\mathcal{M}$ such that $\mathcal{M} \models \varphi$ and $\mathcal{M} \models \varphi \to \psi$. Next, I say what my assumptions mean. This step is carried out by invoking an informal metalanguage. In particular, $\mathcal{M} \models \varphi \to \psi$ *means* that:

$$\text{if } \mathcal{M} \models \varphi, \text{ then } \mathcal{M} \models \psi.$$

Then by my assumptions:

$$\mathcal{M} \models \varphi$$

and

$$\text{if } \mathcal{M} \models \varphi, \text{ then } \mathcal{M} \models \psi.$$

I want to conclude that $\mathcal{M} \models \psi$. But to conclude $\mathcal{M} \models \psi$ on the basis of $\mathcal{M} \models \varphi$ and (if $\mathcal{M} \models \varphi$, then $\mathcal{M} \models \psi$), I have to rely on the validity of an informal metalinguistic inference from $\mathcal{M} \models \varphi$ and (if $\mathcal{M} \models \varphi$, then $\mathcal{M} \models \psi$) to $\mathcal{M} \models \psi$. That is, I have to rely on the validity of a corresponding instance of modus ponens in my metalanguage.

But the validity of the corresponding instance of modus ponens in my metalanguage is a presupposition of my cognitive project. The point is that the notion of → in my object language, and the metalinguistic notion of → I am using in my metatheoretic reasoning, when I say "if... then... ," *mean* the same thing. When I say "if... then... " in my metalanguage, I have essentially invoked a metatheoretic version of →, denote it by $\to^{\text{Meta}}$, such that:

$$(\mathcal{M} \models \varphi) \to^{\text{Meta}} (\mathcal{M} \models \psi) \text{ iff } \mathcal{M} \models \varphi \to \psi. \tag{$*$}$$

So suppose I were to doubt the validity of the corresponding instance of modus ponens in my metatheory. Then I doubt that whenever $\mathcal{M} \models \varphi$ and $\mathcal{M} \models \varphi \to^{\text{Meta}} \mathcal{M} \models \psi$ it follows that $\mathcal{M} \models \psi$. By $(*)$, this is just to say that I doubt whenever $\mathcal{M} \models \varphi$ and $\mathcal{M} \models \varphi \to \psi$, it follows that $\mathcal{M} \models \psi$. And *this* is just to say that I doubt $\varphi, \varphi \to \psi \models \psi$. But $\varphi, \varphi \to \psi \models \psi$ is a presupposition of my cognitive project. So I am thereby rationally committed to doubting the significance or competence of c.

Furthermore, it also seems right to say that my understanding of the validity of modus ponens in my metalanguage is no more secure than my understanding of the validity of modus ponens in my object language (again, the notion of → in my object language, and the notion of $\to^{\text{Meta}}$ in my metalanguage, mean the same thing). And if I were to accept an onus to justify the validity of modus ponens in my object language, I would presumably implicitly undertake a commitment to justify the validity of modus ponens in my metalanguage. So, it seems that my attempt to justify $\varphi, \varphi \to \psi \models \psi$ has led to the first stage of infinite justificatory regress in Wright's sense. If I then set out to justify the validity of modus ponens in my metalanguage, I would again move one level up, and end up in a similar position as before. So, if I were to carry on in this way, I would be led to infinite justificatory regress in Wright's sense.

We will return to attempts to justify the validity of certain inference rules in due course. At this point, hopefully these three examples have helped put a little meat on the bones of the idea of entitlements of cognitive project. Let us pause and draw out the salient points from what we have seen so far, with respect to where we are headed.

3.2.3 Summary

First, we have seen that the skeptic's demand for warrant for certain cornerstone propositions is the motivating context for entitlements. We've seen three examples of attempts to justify cornerstones which plausibly seem to lead to infinite justificatory regress in Wright's sense. We note, though, that in and of itself, this is not what drives through the skeptical conclusion. The real problem with cornerstones is that without an independent response to the skeptical argument, *the question of how we might justify cornerstones at all is completely mysterious.* This is the very *essence* of the kind of skeptical challenge entitlements are meant to assuage. The skeptic demands a justification of cornerstone propositions, but at the same time takes all of our usual justificatory resources off the table.[24] Entitlements are introduced to satisfy this demand for free on our behalf. So, if entitlements of cognitive project are

[24] Cf. (Maddy, 2016), who calls this a demand for extraordinary evidence.

bona fide examples of entitlements in cases of rational theory acceptance, we ought to expect the sort of principles to which one is entitled in cases of rational theory acceptance to be similar in kind: without entitlements, we ought to expect difficulties with the idea of how we might justify these principles at all. But as far as *arithmetic* theory acceptance is concerned, we do not think there are any real difficulties.

Second, and relatedly to the first point, entitlements are introduced to *rescue* the everyday justified beliefs we think we hold, from skeptical threat. By epistemically "propping up" (so to speak) corresponding cornerstone propositions in other ways, the status of the everyday justified beliefs we already think we hold is restored. If entitlements of cognitive project are bona fide examples of entitlements in cases of rational theory acceptance, we should expect things to work in a similar way. But we will see shortly that in the context of (Fischer et al., 2021), *new* beliefs enter the picture as a *result* of our entitlements. What, then, is the status of these *new* beliefs?

Third, and finally, for entitlements of cognitive project to succeed, they have to be *non-evidential*. Entitlements of cognitive project cannot speak to the likely truth of the propositions they are supposed to warrant. As a result, the beliefs we form on the basis of entitlements are just that – beliefs. They are not, nor can they be, *justified* beliefs. For justified beliefs require evidence, and entitlements of cognitive project do not carry any. Combining this observation with the last point, we should be on high alert if we encounter new beliefs which enter the picture purely as a result of an entitlement, which are also supposed to be *justified* beliefs. Yet this is just the kind of story we find in (Fischer et al., 2021).

With these three points in mind, let us finish setting the stage for our main argument, by turning to the context of rational theory acceptance in (Fischer et al., 2021) in which entitlements of cognitive project appear.

3.3 Hypatia and her justification of new mathematical beliefs

The goal in (Fischer et al., 2021) is to argue that the non-classically governed concept of type-free truth can play an essential role in justifying new mathematical beliefs. The argument for this claim is supported by framing it in the context of Wright's entitlements of cognitive projects. There are two core ideas: (1) when we are justified in believing a given mathematical theory S, we are entitled to extend S with principles governing a non-classical type-free disquotational concept of truth. (2) When we are justified in believing a given mathematical theory S, we are entitled to extend S by reflection principles for S (Fischer et al., 2021, p. 64). Thus, we are warranted in extending S by these principles by entitlement of a particular cognitive project; a project in which we begin with a justified belief in the principles of a suitable theory of arithmetic S, and whose goal is to justify new mathematical beliefs. In particular, the goal of this cognitive project is to arrive at newly justified mathematical beliefs in the principles of Predicative Analysis.[25]

There is a lot going on here, both epistemically and logically. Our goal for

[25] See e.g., (Feferman, 1964).

the remainder of chapter 3 is to slow things down, and attempt to precisely enumerate the stages of this justificatory process. First, let us clarify what the type-free disquotational concept of truth consists of, what the underlying logic is, and what the reflection principles for S consist of.

The unrestricted type-free disquotational concept of truth corresponds to two *sequents*, expressions of the form $\Gamma \Rightarrow \Delta$, where Γ, Δ are finite sets of formulas of a language containing a disquotational truth predicate T. They are:

(T1) $\varphi \Rightarrow T(\ulcorner \varphi \urcorner)$, and

(T2) $T(\ulcorner \varphi \urcorner) \Rightarrow \varphi$,

where φ is any formula in the language of S expanded to include a new unary predicate $T(x)$. There are well-known problems with classically governed type-free notions of disquotational truth, which motivates the choice of non-classical ambient logic. In particular, the ambient logic governing the notion of disquotational truth[26] is the four-valued background logic FDE. FDE is a sublogic of classical logic, the precise details of which do not matter to us here. Because of the Liar Paradox, not all theories of truth are such that $T(\ulcorner \varphi \urcorner)$ is equivalent to φ in general. For example, this is not the case in Feferman's type-free theory of truth KF. In cases like these, the theory of truth is not a (fully) *disquotational* theory of truth. It is possible to add stronger, compositional principles for truth, but the resulting theory allows for both truth value gaps and gluts. To exclude gluts one may add a consistency axiom $T(\ulcorner \neg\varphi \urcorner) \Rightarrow \neg T(\ulcorner \varphi \urcorner)$ to

[26] *And* the arithmetical theory S. Cf. (Fischer et al., 2021, Footnote 33).

the compositional principles. To exclude gaps, one may add a completeness axiom $\neg T(\ulcorner\neg\varphi\urcorner) \Rightarrow T(\ulcorner\neg\varphi\urcorner)$ to the compositional principles. Essentially, adding consistency and completeness axioms correspond to paracomplete and paraconsistent approaches to a logic for a theory of transparent truth. But neither approach sits well with global reflection principles – principles of the form $\mathsf{Pr}_{\mathsf{S}}(\ulcorner\varphi\urcorner) \Rightarrow T(\ulcorner\varphi\urcorner)$, expressing the claim that all the theorems of S are true. This motivates the choice of FDE logic in (Fischer et al., 2021), since FDE logic stays neutral regarding the choice between paracompleteness and paraconsistency. The choice between a paracomplete and a paraconsistent approach is what the character Hypatia in the title of this paper is maintaining silence about.

Finally, the rules corresponding to reflection principles for suitable theories S are essentially *uniform reflection principles* of the following form:[27]

$$\frac{\Rightarrow \mathrm{Pr}_{\mathsf{S}}(\ulcorner\Gamma \Rightarrow \Delta\urcorner)}{\Gamma \Rightarrow \Delta}\,(\mathrm{r}_{\mathsf{S}})$$

where $\mathrm{Pr}_{\mathsf{S}}(\ulcorner\Gamma \Rightarrow \Delta\urcorner)$ is a unary provability predicate representing that the sequent $\Gamma \Rightarrow \Delta$ is derivable in S. The idea is that if $\mathrm{Pr}_{\mathsf{S}}(\ulcorner\Gamma \Rightarrow \Delta\urcorner)$ is derivable in a suitable background theory, then we may infer $\Gamma \Rightarrow \Delta$. We may take the background theory to be some weak fragment of arithmetic, say *Kalmar Elementary Arithmetic* EA.[28] If S is an axiomatizable theory, we define the *reflection on* S as the closure of S under the reflection rule $(\mathrm{r}_{\mathsf{S}})$ by:

$$\mathrm{r}(\mathsf{S}) = \mathsf{S} + (\mathrm{r}_{\mathsf{S}}).$$

[27] We will meet a stronger version of this reflection principle later on, for technical reasons.

[28] EA is a finitely axiomatizable subtheory of QF–IA, a conservative extension of Primitive Recursive Arithmetic with first-order quantifiers and the induction schema for Δ_0-formulas.

This helps us be a little more precise about things, but we can do better before we outline the basic stages of the justificatory project described above.

Let us introduce Hypatia, one of the characters featured in (Fischer et al., 2021). Hypatia works in the foundations of mathematics, and her goal is to undertake such a justificatory project. First, we are offered a broad overview of Hypatia's project. Hypatia is:

> happy to rely on full disquotational truth in her reasoning. She is persuaded that at least a portion of arithmetic can be fully justified. In regard to "stronger" infinitistic methods she is more careful. Although she is not strictly refusing these infinitistic parts, she intends to justify them by extending her justification of arithmetic to richer areas of mathematics. (Fischer et al., 2021, p. 74)

Shortly after this passage, we are offered a more detailed version of Hypatia's project, which we quote here in full. For concreteness, let's say Hypatia starts out with a justified belief in the principles of the weak fragment of arithmetic EA. After that:

> she is entitled to rely on FDE logic in a fully schematic form so that she knows any arithmetical sentence that can be seen to follow from the axioms of EA. Now she is warranted in introducing a notion of disquotational truth... she is entitled to embark on a cognitive project that involves adopting [the rules (T1) and (T2)]. Thus Hypatia comes to accept the theory S formulated in the language expansion with a truth predicate and closed under FDE logic and the disquotational rules for truth: call this theory TS_0.

> When she does so, her acceptance of TS_0 includes her firm belief that all the theorems of this theory are true. She comes to accept the stronger theory obtained by reflecting on the basic disquotational theory TS_0. If all is well, she is *entitled* to embrace reflection principles or rely on reflection rules for TS_0.
>
> Hypatia is justified in believing all the mathematical theorems of this extended theory. Moreover, the reliability of the disquotational truth concept and the process of reflection allows her to believe in the truth of everything that the extension of TS_0 with reflection proves. Hypatia is then again entitled to adopt reflection principles or rules for the stronger theory and justified in accepting all the (mathematical) theorems of a further iteration of reflection over TS_0. (Fischer et al., 2021, p. 75)

This is the key passage, which together with our observations above about the rules and the logic at play, suggests a way to enumerate the basic stages of this process. We achieve this in two stages, the first of which is below. The emphasized terms will focus our subsequent discussion. So, as a first attempt at enumerating the basic stages of the process involved in Hypatia's cognitive project of justifying new mathematical beliefs, consider the following process (1′)–(9′):

(1′) Hypatia starts out with a justified belief in the principles of the theory EA.

(2′) Hypatia is entitled to rely on FDE logic in a fully schematic form.

(3′) Hypatia is *now* entitled to (*embark on a cognitive project such that she comes to*) adopt the sequents (T1) and (T2).

(4′) Hypatia thereby comes to accept the theory TS_0.

(5′) Hypatia is thereby justified in believing the principles of the theory TS_0.

(6′) Hypatia is (*now*?) entitled to (*embark on a cognitive project such that she comes to*?) adopt the rule (r_{TS_0}) for TS_0.

(7′) Hypatia thereby comes to accept the theory $\mathsf{TS}_0 + (\mathrm{r}_{\mathsf{TS}_0}) = \mathrm{r}(\mathsf{TS}_0)$.

(8′) Hypatia is thereby justified in believing the principles of the theory $\mathrm{r}(\mathsf{TS}_0)$.

(9′) The process now continues iteratively as in step (6′), where instead of an entitlement to (*embark on a cognitive project such that she comes to*) adopt the rule (r_{TS_0}) for TS_0, Hypatia is now entitled to (*embark on a cognitive project such that she comes to*) adopt the rule ($r_{\mathrm{r}(\mathsf{TS}_0)}$) for $\mathrm{r}(\mathsf{TS}_0)$.

We think (1′)–(9′) is a particularly awkward reconstruction. Let us pay attention to the two emphasized instances across (3′) and (6′) of "now," and the two emphasized mentions of embarkment upon cognitive projects. The instances of these phrases in (3′) are part of the key passage in (Fischer et al., 2021) quoted above. The instances of these phrases in (6′) are our addition, and do not appear in the quoted passage. The reason for these additions will become clear shortly.

First let us consider the occurrence of these phrases in (3′). As a motivator, one reason we emphasized where entitlement occurs in this process is because we are interested in exactly what principles Hypatia's entitlements are *for*. But also, Hypatia's entitlements, are entitlements *of cognitive project*. Thus, let us try and pin down exactly what cognitive project each of these entitlements corresponds to. Hypatia's entitlement in (2′): to rely on FDE logic in a fully schematic form, *must* be an entitlement of her overall cognitive project, to justify new mathematical knowledge from her starting point of a justified belief in the principles of the theory EA. This is the *only* cognitive project on the table up to that point; what else could this entitlement relate to? On the other hand, in (3′), we are introduced to another cognitive project, in which Hypatia comes to adopt the sequents (T1) and (T2). What is this cognitive project supposed to consist of? We consider two possibilities.

First, perhaps this new cognitive project is just the same as Hypatia's original cognitive project. That is: (a) Hypatia's adoption of the sequents (T1) and (T2) is an entitlement of her cognitive project of justifying new mathematical knowledge from her starting point of a justified belief in the principles of the theory EA. If this is the case, this raises a couple of questions. What are we to make of the force of the emphasized word "now" in (3′)? The connotation, as we understand it, is that *first* Hypatia is entitled to rely on FDE logic in a fully schematic form, and *after that*, she is entitled to adopt the sequents (T1) and (T2). But if both of these entitlements are entitlements of her cognitive project of justifying new mathematical knowledge from her starting point of a justified belief in the principles of the theory EA, why should it matter in what *order* she arrives at them? Second, what are we to make of the phrase "she is entitled to embark on a cognitive project that involves adopting [the sequents (T1) and (T2)]"? We already know Hypatia has embarked on a cognitive project. To tell us she has embarked on this cognitive project suggests that it is supposed to be a new, different one.

Thus, consider instead a second possibility: (b) Hypatia's adoption of the sequents (T1) and (T2) is an entitlement of a new cognitive project, different than her original one (but presumably *somehow* related). If this is the case, then essentially, her cognitive context has shifted slightly, since it is no longer *merely* to the end of justifying new mathematical beliefs. This also raises questions. For Hypatia's entitlement to adopt the sequents (T1) and (T2) is not *merely* an entitlement of her cognitive project of justifying new mathematical beliefs. Rather, Hypatia's entitlement to adopt the sequents (T1) and (T2) is an entitlement of a (presumably stronger) cognitive project. But this is a different claim than the claim that the sequents (T1) and (T2) for EA are entitlements of Hypatia's cognitive project of justifying new mathematical beliefs. Indeed, iteratively, and even ignoring for now the question of whether Hypatia embarks on a new (different) cognitive project in (6), *every* time Hypatia iterates the process $(1')$–$(9')$, her cognitive project shifts. This would involve *transfinitely many* different claims, than (e.g.) the claim that the sequents (T1) and (T2) are entitlements of Hypatia's cognitive project of justifying new mathematical beliefs.[29]

[29] Eventually, Hypatia arrives at the theorems of Predicative Analysis, in the sense of (Feferman, 1964), via iterations of the reflection process of length up to the *Feferman-Schütte ordinal* Γ_0 (Feferman, 1964; Schütte, 1965b).

Second, consider our addition of the interrogative emphasized instances of the above phrases in (6′). Our intention behind this addition was to ask: if Hypatia's entitlements in (3′) *do* indeed implicitly shift the cognitive project upon which she has embarked, then does the same kind of thing happen when she acquires her entitlements in (6′)? If so, then this serves to reinforce our worries above. The process (1′)–(9′) then involves transfinitely many different claims, than (e.g.) the claim that the sequents (T1) and (T2) are entitlements of Hypatia's cognitive project of justifying new mathematical knowledge. If not, though, we are left with the following mysterious question: why does Hypatia's project shift in (3′) but not in (6′)?

Given these observations, we think the best way forward at this point is to offer a second reconstruction of Hypatia's cognitive project, a reconstruction which avoids having to deal with the sorts of questions we have just raised. In particular: (a) we set aside our worries about the connotations of "now," and the mention of embarkment upon a cognitive project in (3′), by removing them altogether (along with our own additions in (6′)). (b) We avoid the peculiar question about whether the order in which Hypatia arrives at her entitlements matters in (2′) and (3′), by bundling them into one set of entitlements. We will continue to refer to this set of entitlements by Hypatia's entitlement to rely on the sequents (T1) and (T2), but the reader should keep in mind that the ambient logic governing those rules is FDE. (c) Finally, in most of what follows in chapter 4, we will consider *one* iteration (the first iteration) of Hypatia's overall project, and do away with the final stage, where she iterates. Nothing will be lost by doing so: if we can cast doubt on one iteration of Hypatia's cognitive project, we have thereby cast doubt on every iteration. We will return to iterations of Hypatia's project in chapter 5. Putting everything together, consider the following enumeration of the basic stages involved in Hypatia's justificatory process:

(1) Hypatia starts out with a justified belief in the principles of the theory EA.

(2) Hypatia is entitled to adopt the sequents (T1) and (T2).

(3) Hypatia thereby comes to accept the theory TS_0.

(4) Hypatia is thereby justified in believing the principles of the theory TS_0.

(5) Hypatia is entitled to adopt the rule $(\mathrm{r}_{\mathsf{TS}_0})$ for TS_0.

(6) Hypatia thereby comes to accept the theory $\mathsf{TS}_0 + (\mathrm{r}_{\mathsf{TS}_0}) = \mathrm{r}(\mathsf{TS}_0)$.

(7) Hypatia is thereby justified in believing the principles of the theory $\mathrm{r}(\mathsf{TS}_0)$.

We write JMB for the process (1)–(7), for the "justification of new mathematical beliefs." We are now almost ready to begin teasing apart the notion of entitlements (of cognitive project), as they feature in JMB above, and the notion of entitlements of cognitive project per Wright (2004).

Before we move on, we make two remarks. First, we note that there is one obvious sense in which these two contexts differ, which is related to our earlier remarks in this chapter. The context in (Wright, 2004) is *epistemic.* But the context in (Fischer et al., 2021) is *mathematical.* We have suggested that justification in these two contexts is different. On the face of things, evidence does not work the same way in mathematics as it does in ordinary epistemic settings. Hypatia's justified belief in EA, in stage (1) of JMB above, is formed on the basis of independent methods of axiom justification. Hypatia's justified belief in stages (4) and (7) is mathematically justified belief, which is supposed to depend in some sense on her justified belief in EA. Going forward, we will implicitly distinguish Hypatia's entitlements, as they occur in JMB, from entitlements as they were introduced in (Wright, 2004). For entitlements were introduced in an epistemic setting, and carry no *epistemic* evidence (whatever that consists of). So we will suppose that Hypatia's entitlements in JMB play an analogous role in the mathematical setting: they carry no *mathematical* evidence (whatever that consists of). This way, we aim to ensure that we are speaking about the right *kind* of thing, when we move between the epistemic and mathematical contexts.

Second, there seems to be a slight typing error in the way we have formulated JMB. Entitlements of cognitive project are defined for propositions. But Hypatia's entitlements in JMB are entitlements to adopt a principle. For example, the proposition asserting the validity of the sequents (T1) and (T2) is an entitlement of Hypatia's cognitive project *c*. But as JMB is written, the principles (T1) and (T2) themselves are what Hypatia is entitled to. Over chapter 4, we will avoid this issue by saying that Hypatia is warranted by entitlement in adopting a principle (P) just in case the proposition asserting the validity of (P) is an entitlement of her cognitive project *c*. In chapter 5, when we set out our solution to the problems we are about to encounter, we will distinguish between the propositions asserting the validity of some principle (P), and the warrant we have for extending a theory by the principle (P) itself. (This distinction seems to be obscured in (Fischer et al., 2021).)

With that, let us turn to our analysis of JMB.

Chapter 4

The problem with entitlements of cognitive project

We carry out our analysis in two broad stages. First, in section 4.1, we examine the *soundness* of stages (2) and (5) of JMB. We ask: are Hypatia's warrants for adopting the sequents (T1) and (T2), and for adopting the rule ($\mathrm{r}_{\mathsf{TS}_0}$) for TS_0, *really* bona fide examples of entitlements of cognitive project? Second, in section 4.2, we examine the *validity* of JMB. We ask: if Hypatia's warrants for adopting the sequents (T1) and (T2), and for adopting the rule ($\mathrm{r}_{\mathsf{TS}_0}$) for TS_0, really are entitlements of cognitive project, *does* she thereby ultimately come to hold a *justified* belief in the principles of the theory $\mathrm{r}(\mathsf{TS}_0)$? Along the way, we highlight important junctures, which will ultimately help inform our solution to the problems we are about to encounter.

4.1 Soundness of JMB

Are Hypatia's warrants for adopting the sequents (T1) and (T2), and for adopting the rule ($\mathrm{r}_{\mathsf{TS}_0}$) for TS_0, bona fide examples of entitlements of her cognitive project? Over sections 4.1.1 and 4.1.2, our strategy is to take Wright's definition of entitlement of cognitive project, and explore the extent to which Hypatia's attitude towards the principles she adopts, meet the criteria of that definition. First, in section 4.1.1, we consider Hypatia's warrant for adopting the sequents (T1) and (T2). After that, in section 4.1.2, we consider Hypatia's warrant for adopting the rule ($\mathrm{r}_{\mathsf{TS}_0}$) for TS_0.

4.1.1 Disquotational truth

We begin by examining the proposition asserting the validity of the sequents:

(T1) $\varphi \Rightarrow T(\ulcorner \varphi \urcorner)$, and

(T2) $T(\ulcorner \varphi \urcorner) \Rightarrow \varphi$,

where φ is any formula in the language $\mathcal{L}_{\mathsf{EA}} \cup \{T\} = \mathcal{L}_T$. Let c be Hypatia's cognitive project, of justifying new mathematical beliefs from her currently held justified belief in the principles of EA. Recall that a proposition p is an entitlement of Hypatia's cognitive project c just in case p satisfies the following conditions:

1. p is a presupposition of c.

2. Hypatia has no sufficient reason to believe that p is untrue.

3. Hypatia's attempt to justify p would involve further presuppositions in turn of no more secure a prior understanding... and so on without limit; so that if Hypatia accepted that there is nevertheless an onus to justify p, she would implicitly undertake a commitment to an infinite regress of justificatory projects, each concerned to vindicate the presuppositions of its predecessors.

Let p be the proposition "the sequents (T1) and (T2) are valid." Our aim is to explore the extent to which p meets the three conditions for being an entitlement of Hypatia's cognitive project c. We aim to tell a plausible story about why we might think p meets conditions 1 and 2. However, we aim to cast some doubt on the idea that p meets condition 3. We first address clause 2, since it appears to be the most straightforward. We then address clauses 1 and 3.

Lack of sufficient countervailing evidence

Does Hypatia have any sufficient reason to believe that p is false? To answer this question, it is useful to consider what a sufficient reason to believe that p is false would consist of. We think there are two plausible suggestions. For the moment, without loss of generality, consider any instance of (T1), say $\varphi \Rightarrow T(\ulcorner \varphi \urcorner)$, where φ is a fixed sentence in the language $\mathcal{L}_T$. First, Hypatia would have sufficient reason to believe that p is false if EA, the theory in which Hypatia holds a justified belief, could derive φ on one hand, and $\neg T(\ulcorner \varphi \urcorner)$ on the other. But in general, neither φ nor $T(\ulcorner \varphi \urcorner)$ are even sentences in the language of EA, since in general, the predicate T occurs in both. Perhaps Hypatia sets out to see if EA can nonetheless *interpret* the predicate T. She isolates the complexity of φ, say Σ_n for some $n \in \omega$, and comes to realize that EA can define a unary Σ_n-predicate T_n such that EA interprets φ. But in particular, doing things this way, Hypatia finds that EA (if it is consistent) cannot derive φ on one hand and $\neg T_n(\ulcorner \varphi \urcorner)$ on the other. Furthermore, Hypatia also quickly comes to realize that while EA can define disquotational truth predicates for $\mathcal{L}_T$-sentences of a given logical complexity, due to the usual Tarskian considerations, EA cannot define a disquotational truth predicate that applies uniformly to $\mathcal{L}_T$-sentences of arbitrary logical complexity. Overall, she realizes that EA cannot interpret φ on one hand, and $\neg T(\ulcorner \varphi \urcorner)$ on the other. This rules out the first possibility.

The natural move now would be for Hypatia to consider what happens when she adds the sequents to her starting theory EA. So, Hypatia considers a second possibility: she would have sufficient reason to believe that p is false if the $\mathcal{L}_T$-theory TS_0, consisting of EA together with (T1) and (T2), was inconsistent. But (maybe even after trying) she finds herself unable to derive an inconsistency from the axioms of TS_0. Hypatia's lesson from the liar paradox is that if there is a coherent notion of type-free truth, then it is governed by non-classical logic. But she has come to embrace FDE logic (in a fully schematic form) for her formal notion of truth. Thus, there is no obvious way for Hypatia to obtain a contradiction in TS_0 using her formal notion of truth. Furthermore, she comes to realize that TS_0 is conservative over EA. Thus, if an *arithmetical* contradiction could be derived in TS_0, then the same contradiction could be derived in EA alone. Hypatia finds the latter possibility extremely unlikely. Overall, she concludes that there is no reason to think that TS_0 is inconsistent. This rules out the second possibility. With these two possibilities considered, Hypatia finds no sufficient reason to believe that p is false.[1] Thus, it seems plausible enough that p meets condition 2 above.

Presuppositional clause

Next, consider clause 1. We are interested in the following question: is p a presupposition of Hypatia's cognitive project c? Let us try and tell a story about why it might be plausible to think that p is a presupposition of c. We will give the argument for the validity of (T1). We then remark on the argument for (T2).

The basic idea is to distinguish between Hypatia's justified belief in sentences φ in the language $\mathcal{L}_{\mathsf{EA}}$ of EA, and her justified belief in the corresponding sentences $T(\ulcorner\varphi\urcorner)$ in the expanded language $\mathcal{L}_T$. We argue that the way in which Hypatia's justified belief in $\mathcal{L}_{\mathsf{EA}}$-sentences φ propagates to the corresponding $\mathcal{L}_T$-sentences $T(\ulcorner\varphi\urcorner)$, presupposes the validity of the sequent (T1). But this propagation of beliefs, from $\mathcal{L}_{\mathsf{EA}}$-sentences φ to the corresponding $\mathcal{L}_T$-sentences $T(\ulcorner\varphi\urcorner)$, must occur at some point during Hypatia's cognitive project c. Thus, if Hypatia were to doubt the proposition asserting the validity of (T1), she would be rationally committed to doubting the overall significance or competence of c.

To explain the basic idea, consider the following cognitive project c_T, dif-

[1] Maybe there are other possibilities, but we rest content with the point made here.

ferent from Hypatia's current cognitive project c. When Hypatia sets out to undertake c_T, she begins with a currently held justified belief in every *arithmetical* consequence of EA, where "consequence" is understood to exclude trivial consequences (i.e. the axioms of EA). So c_T is similar to Hypatia's current cognitive project c in this respect. But unlike Hypatia's current cognitive project c, the cognitive goal in successfully undertaking out c_T is only to arrive at a justified belief in any $\mathcal{L}_T$-sentence $T(\ulcorner\varphi\urcorner)$, where φ is an arithmetical consequence of EA. (Hypatia's cognitive goal in successfully carrying out c is quite different, and much more ambitious – she aims to arrive at a justified belief in the principles of Predicative Analysis).

Hypatia's cognitive project c_T certainly seems well-defined, insofar as there does seem to be a difference between her justified belief in the arithmetical consequences of EA, and her justified belief in any $\mathcal{L}_T$-sentence $T(\ulcorner\varphi\urcorner)$, where φ is an arithmetical consequence of EA. For example, we are assuming that Hypatia arrived, somehow, at a justified belief in the principles of EA. We haven't said anything about how Hypatia came to hold this set of justified beliefs in the first place, but for concreteness, suppose she came to hold a justified belief in the principles of EA because the natural number structure is somehow intuitable to her, and she verified that the axioms of EA hold in the standard model.[2] Then at that particular point in time, Hypatia holds a justified belief in all and only the principles of EA. She does not, for example, yet hold a justified belief in the $\mathcal{L}_T$-sentence $T(\ulcorner\varphi\urcorner)$, where φ is any non-trivial arithmetical consequence of EA.

[2] We will say much more about the how Hypatia's justified belief in the *axioms* of EA propagates to the general *theory* of EA in section 4.1.2.

How, then, *would* Hypatia come to hold a justified belief in the $\mathcal{L}_T$-sentence $T(\ulcorner\varphi\urcorner)$? One thing she might do is try to exhibit a proof of $T(\ulcorner\varphi\urcorner)$ from the axioms of EA, and assure herself that justified belief is preserved throughout this process.[3] But this is not possible: $T(\ulcorner\varphi\urcorner)$ is not formulated in the language of EA, and as we said above, EA cannot interpret $T(\ulcorner\varphi\urcorner)$ uniformly for all φ.

So how else might she come to hold a justified belief in $T(\ulcorner\varphi\urcorner)$? We suggest Hypatia's one remaining option is this: she argues that since the truth predicate T is disquotational, then φ *means the same thing* as $T(\ulcorner\varphi\urcorner)$. Thus, since she holds a justified belief in φ, her justified belief is preserved in virtue of the identical meaning of these two sentences. Insofar as the identical meanings of φ and $T(\ulcorner\varphi\urcorner)$ consist in their identical truth conditions, then Hypatia infers a justified belief in $T(\ulcorner\varphi\urcorner)$ on the basis of her justified belief in φ, and the following principle:

(†) Whenever φ is true, $T(\ulcorner\varphi\urcorner)$ is true.

The point is that for Hypatia's justified belief in the arithmetical consequences φ of EA to transfer to the corresponding $\mathcal{L}_T$-sentences $T(\ulcorner\varphi\urcorner)$, she requires collateral warrant for the principle (†).

But the principle (†) for which Hypatia requires collateral warrant *is* just the proposition asserting the validity of the corresponding instance of the sequent (T1): that whenever $\mathcal{M} \models \varphi$, we have $\mathcal{M} \models T(\ulcorner\varphi\urcorner)$. Thus, if Hypatia infers justified belief in $T(\ulcorner\varphi\urcorner)$ on the basis of her justified belief in φ and the principle (†), then in particular, she presupposes the validity of the sequent

[3] Much more on this possibility in section 4.1.2.

$\varphi \Rightarrow T(\ulcorner \varphi \urcorner)$.

So, if Hypatia were to doubt the proposition asserting the validity of instances of the form $\varphi \Rightarrow T(\ulcorner \varphi \urcorner)$ (where doubt might include even agnosticism about the validity of instances of the form $\varphi \Rightarrow T(\ulcorner \varphi \urcorner)$), she would also be rationally committed to doubting the significance or competence of the cognitive project c_T. For the goal of c_T requires her to hold a *justified belief* in any $\mathcal{L}_T$-sentence $T(\ulcorner \varphi \urcorner)$, where φ is an arithmetical consequence of EA. And justified belief excludes doubt.

Finally, we claim that Hypatia's cognitive project c_T is embedded in her more ambitious cognitive project c. By that, we mean that the cognitive goal of c_T is necessary for the cognitive goal of c. Why? The principles of EA, understood to include all the arithmetical consequences of EA, form the basis of Hypatia's cognitive project c. It is those principles which Hypatia first comes to hold a justified belief in. If she could not come to hold a justified belief in the principles of the theory TS_0, where (e.g.) she first encounters all the $\mathcal{L}_T$-sentences $T(\ulcorner \varphi \urcorner)$ (where φ is an arithmetical consequence of EA), then how can she expect to achieve anything further, and ultimately justify new mathematical beliefs in the principles of Predicative Analysis on the basis of her justified belief in the axioms of EA?

Putting all this together, we claim to have offered a plausible understanding of why the proposition asserting the validity of the sequent (T1) is a presupposition of Hypatia's larger cognitive project c. For achieving the cognitive goal of Hypatia's narrower cognitive project c_T is necessary for achieving the cognitive goal of her overall cognitive project c, and the proposition asserting the validity of (T1) is a presupposition of c_T. Thus, the proposition asserting the validity of (T1) is a presupposition of c.

Let us remark on the argument for the proposition asserting the validity of the sequent (T2). To make a similar argument, we should want to identify a stage during Hypatia's JMB process during which her justified beliefs in $\mathcal{L}_T$-sentence $T(\ulcorner \varphi \urcorner)$ propagate to the corresponding arithmetical consequences φ of EA. We admit that it seems a little strange to think of this as happening during Hypatia's JMB process. After all, she starts out with a justified belief in purely arithmetical sentences, and after that, her beliefs are supposed to propagate to sentences in the expanded language, rather than the other way around. But the validity of the sequent (T2) seems to ensure a certain *coherence* between her justified beliefs in arithmetical sentences φ, and the corresponding $\mathcal{L}_T$-sentences $T(\ulcorner \varphi \urcorner)$. It would presumably be rather strange if Hypatia were able to infer a justified belief in $\mathcal{L}_T$-sentences $T(\ulcorner \varphi \urcorner)$ on the basis that she holds a justified belief in φ and the principle (†), but *not* have this cohere with the idea that if she holds a justified belief in an arbitrary $\mathcal{L}_T$-sentence $T(\ulcorner \varphi \urcorner)$, she can also infer a justified belief in the corresponding arithmetical sentence φ on that basis coupled with the following principle:

(††) Whenever $T(\ulcorner \varphi \urcorner)$ is true, φ is true.

And (††) is just the proposition asserting the validity of (T2). So it seems like doubting the proposition asserting the validity of (T2) would result in a peculiar epistemic non-equivalence between arithmetical sentences and $\mathcal{L}_T$-sentences, rather than rationally compel Hypatia to think she was unable to carry out c in a significant or competent way. So, from now on, we will assume that this coherence is something Hypatia wants, and rest content with the story we have offered here. It seems to be the most plausible way of thinking of the proposition p asserting the validity of the sequents (T1) and (T2) as presuppositions of her cognitive project c. Let us turn now to condition 3, which is the most interesting case.

Infinite justificatory regress

Would Hypatia's attempt to justify p involve further presuppositions in turn of no more secure a prior understanding (and so on without limit), such that if she accepted that there is nevertheless an onus to justify p, she would implicitly undertake a commitment to an infinite regress of justificatory projects, each concerned to vindicate the presuppositions of its predecessors?

To address this question, it will be helpful for us to first walk through a few other examples in which justifying the validity of a rule leads to the sort of justificatory regress of condition 3. First recall the third kind of example of entitlement of cognitive project we discussed earlier, concerning the validity of modus ponens.

Let c^* be *any* cognitive project in which we hold a justified belief in the principles of EA, and which involves the use of our intellectual capacities towards some cognitive achievement. We saw in section 3.2.2 how an attempt to justify the validity of the relevant instance of modus ponens leads to regress in the context of such a cognitive project c^*. The basic idea was that any attempt to give such a justification must rely on the validity of a corresponding instance of modus ponens itself in the metalanguage. Thus, if we accept the onus to provide a justification for the validity of a particular instance of modus ponens, we are implicitly committed to justifying an infinite regress of corresponding instances of the validity of modus ponens, each in turn of no more secure a prior understanding than its predecessor.

Recall the argument: suppose, during our cognitive project c^*, we have relied on the validity of an inference from $\varphi, \varphi \to \psi$ to ψ in the object language of EA by modus ponens (perhaps this occurs in a proof of a theorem we have set out to give from the axioms of EA). Let us attempt to justify the validity of this inference.

Fix a model $\mathcal{M}$ such that $\mathcal{M} \models \varphi$ and $\mathcal{M} \models \varphi \to \psi$. Moving to our informal metalanguage, $\mathcal{M} \models \varphi \to \psi$ means that *if* $\mathcal{M} \models \varphi$, *then* $\mathcal{M} \models \psi$. At this point, we have essentially invoked a metatheoretic version of $\to$, denote it by $\to^{\text{Meta}}$, such that:

$$(\mathcal{M} \models \varphi) \to^{\text{Meta}} (\mathcal{M} \models \psi) \text{ iff } \mathcal{M} \models \varphi \to \psi.$$

Then by our assumptions, we have that $\mathcal{M} \models \varphi$ and $(\mathcal{M} \models \varphi) \to^{\text{Meta}} (\mathcal{M} \models \psi)$. Now, we want to conclude that $\mathcal{M} \models \psi$. But to conclude $\mathcal{M} \models \psi$, we seem to have to rely on the validity of the corresponding metalinguistic instance of modus ponens. Since the notion of $\to$ in our object language, and the metalinguistic notion $\to^{\text{Meta}}$ we are using in our metatheoretic reasoning, mean the same thing, the validity of modus ponens in our metalanguage is a presupposition of c^*. (And is of no more secure a prior understanding than that of modus ponens in our object language, and is such that if we accept an onus to justify the validity of modus ponens in our object language, we are thereby committed to justifying the validity of modus ponens in our metalanguage as well.) If we were to carry on justifying, we are led to infinite justificatory regress in Wright's sense.

Next, we note that the line of reasoning above is not unique to modus ponens, and in fact seems to apply to propositions asserting the validity of *any* rule among our most basic inferential apparatus. Let us briefly reconstruct the above lines of reasoning for the validity of another of these rules: conjunction elimination. Let c^* be the same cognitive project as above, and suppose at some point during our undertaking of c^*, we want to assure ourselves of the validity of an inference from $\varphi \wedge \psi$ to φ by conjunction elimination, where φ, ψ are sentences in the language of EA. So, we attempt to justify $\varphi \wedge \psi \models \varphi$.

Fix a model $\mathcal{M}$ such that $\mathcal{M} \models \varphi \wedge \psi$. Moving to an informal metalanguage, $\mathcal{M} \models \varphi \wedge \psi$ means that $\mathcal{M} \models \varphi$ *and* $\mathcal{M} \models \psi$. At this point, we have essentially invoked a metatheoretic version of $\wedge$, denote it by $\wedge^{\text{Meta}}$, such that:

$$(\mathcal{M} \models \varphi) \wedge^{\text{Meta}} (\mathcal{M} \models \psi) \text{ iff } \mathcal{M} \models \varphi \wedge \psi.$$

Now, we want to conclude that $\mathcal{M} \models \varphi$. But to conclude $\mathcal{M} \models \varphi$, we seem to have to rely on the validity of an informal metalinguistic inference from $(\mathcal{M} \models \varphi) \wedge^{\text{Meta}} (\mathcal{M} \models \psi)$ to $\mathcal{M} \models \varphi$. Furthermore, the notion of $\wedge$ in our object language, and the metalinguistic notion $\wedge^{\text{Meta}}$ we are using in our metatheoretic reasoning, mean the same thing. As a result, the validity of conjunction elimination in our metalanguage is a presupposition of c^*, is of no more secure a prior understanding than that of conjunction elimination in our object language, and is such that if we accept an onus to justify the validity of conjunction elimination in our object language, we are thereby committed to justifying the validity of conjunction elimination in our metalanguage as well. If we were to carry on justifying, we are led to infinite justificatory regress in Wright's sense.

But now let us turn to Hypatia's cognitive project c. We will reconstruct an attempt to justify the validity of an instance of the sequent (T2), say $T(\ulcorner\varphi\urcorner) \Rightarrow \varphi$, for some $\mathcal{L}_T$-formula φ (the argument for (T1) is similar). During this justificatory attempt, we will pay close attention to what happens when we start reasoning in our informal metalanguage. We claim that something quite different happens, compared to the two previous examples.

So let us attempt to justify $T(\ulcorner\varphi\urcorner) \models \varphi$. We start off reasoning like this: fix a model $\mathcal{M}$ such that $\mathcal{M} \models T(\ulcorner\varphi\urcorner)$. Now, at this point in the previous two examples, when we articulated in our informal metalanguage what our assumptions *meant*, we invoked a metalinguistic version of the very notion we set out to justify in the first place. In the context of the current example, this would amount to saying the following: $\mathcal{M} \models T(\ulcorner\varphi\urcorner)$ *means* that it is *true* that $\mathcal{M} \models \varphi$. In saying this, we have invoked a metalinguistic notion of T, denote it by T^{Meta}, such that:

$$T^{\text{Meta}}(\mathcal{M} \models \varphi) \text{ iff } \mathcal{M} \models T(\ulcorner\varphi\urcorner).$$

But our point is that we do not see why we should have to invoke *any* such metalinguistic notion of truth when we say what $\mathcal{M} \models T(\ulcorner\varphi\urcorner)$ means. This is an artifact of the *triviality* of the predicate T: T "does nothing" to φ. So all $\mathcal{M} \models T(\ulcorner\varphi\urcorner)$ *means* is simply that $\mathcal{M} \models \varphi$. But $\mathcal{M} \models \varphi$ was the very conclusion we set out to draw. We reached this conclusion in one step, simply by articulating what our assumption means. So in our justificatory attempt, we have not appealed to any further presupposition of our cognitive project c. We appealed to nothing but the meaning of $\mathcal{M} \models T(\ulcorner\varphi\urcorner)$. In particular, we do not see how this attempt to justify the validity of $T(\ulcorner\varphi\urcorner) \models \varphi$ meets the conditions for infinite justificatory regress in Wright's sense at all.

Let us take stock of what we have learned so far. We have been investigating the claim that the proposition p asserting the validity of the sequents (T1) and (T2) is a genuine entitlement of Hypatia's cognitive project c. We have argued that it is plausible to think that Hypatia has no sufficient reason to believe that p is false. We have also attempted to outline a plausible story about why p is a presupposition of c. But we think it is perfectly possible for Hypatia to justify p without committing herself to infinite justificatory regress in Wright's sense. Let us turn next to the reflection rule ($\mathrm{r}_{\mathsf{TS}_0}$), which is also said to be an entitlement of Hypatia's current cognitive project.

4.1.2 Reflection

Now we are interested in examining the proposition asserting the validity of the rule:

$$\frac{\Rightarrow \mathrm{Pr}_{\mathsf{TS}_0}(\ulcorner\Gamma \Rightarrow \Delta\urcorner)}{\Gamma \Rightarrow \Delta}\,(\mathrm{r}_{\mathsf{TS}_0})$$

Recall that ($\mathrm{r}_{\mathsf{TS}_0}$) captures the notion that if we have established our background theory can provably recognize that the sequent $\Gamma \Rightarrow \Delta$ is provable in TS_0, then $\Gamma \Rightarrow \Delta$ holds.

Let now p be the proposition "the rule ($\mathrm{r}_{\mathsf{TS}_0}$) is valid."[4] Again, we are interested in whether p meets the criteria for being an entitlement of Hypatia's cognitive project c, of justifying new mathematical beliefs in the principles of Predicative Analysis from her currently held justified belief in the principles of EA. An important juncture in our argument occurs during our discussion of p's being a presupposition of c. The key point is that there is a fundamental difference between the reflection principle ($\mathrm{r}_{\mathsf{TS}_0}$), and the kind of Wrightean cornerstone propositions we encountered in section 3.2.1. As in section 4.1.1, we first consider clause 2, then consider clauses 1 and 3.

Lack of sufficient countervailing evidence

Does Hypatia have any sufficient reason to believe that p is false? As in the case of the validity of the sequents (T1) and (T2), there are two plausible suggestions that would help answer this question. First, Hypatia would have sufficient reason to believe that p is false if TS_0, the theory in which she currently holds a justified belief, could derive an invalid instance of the rule ($\mathrm{r}_{\mathsf{TS}_0}$). However, by the usual Gödelian considerations, Hypatia quickly comes to realize that TS_0 cannot derive every instance of the rule ($\mathrm{r}_{\mathsf{TS}_0}$).[5] In fact,

[4] In what follows, we are assuming that Hypatia is now at stage (4) of JMB, where she has come to hold a justified belief in the theory TS_0. But there is nothing special about this. We could just as well reformulate what follows for a corresponding rule (r_{S}), where S is any suitable theory in which Hypatia currently holds a justified belief.

[5] This is Löb's theorem.

she realizes that the only instances of ($\mathrm{r}_{\mathsf{TS}_0}$) derivable in TS_0 are such that the consequent is already derivable in TS_0, in which case the corresponding instance of ($\mathrm{r}_{\mathsf{TS}_0}$) is valid.[6] Thus, she rules out this first possibility. Second, Hypatia would have sufficient reason to believe that p is false if the $\mathcal{L}_T$-theory $\mathsf{TS}_0 + (\mathrm{r}_{\mathsf{TS}_0})$, consisting of TS_0 together with ($\mathrm{r}_{\mathsf{TS}_0}$), was inconsistent. But again (maybe even after trying), Hypatia finds herself unable to derive an inconsistency from the axioms of $\mathsf{TS}_0 + (\mathrm{r}_{\mathsf{TS}_0})$. Moreover, she comes to realize that $\mathsf{TS}_0 + (\mathrm{r}_{\mathsf{TS}_0})$ is consistent if TS_0 is. She finds the latter possibility extremely unlikely, and so overall, she concludes that there is no reason to think that $\mathsf{TS}_0 + (\mathrm{r}_{\mathsf{TS}_0})$ is inconsistent. With these two possibilities considered, Hypatia finds no sufficient reason to believe that p is false.[7] Thus, it seems plausible enough that p meets condition 2 above.

Presuppositional clause

Next consider clause 1 of the definition of entitlement of cognitive project, which states that p is a presupposition of c. By definition: doubting p would rationally commit Hypatia to doubting the significance of competence of her project c. As in section 4.1.1, let us tell a story about why it might be plausible to think that p is a presupposition of c.

This time, the basic idea is to distinguish between Hypatia's justified belief in the *axioms* of TS_0, and her justified belief in the entire *theory* TS_0. We argue that the way in which her justified belief in the axioms of TS_0 propagates to

[6] This is the content of Löb's theorem.

[7] Again, maybe there are other possibilities, but we rest content with the point made here.

the entire theory TS_0 presupposes the validity of $(\mathrm{r}_{\mathsf{TS}_0})$. But this propagation of beliefs, from axioms to theory, must occur at some point during Hypatia's cognitive project c. Thus, if she were to doubt the validity of $(\mathrm{r}_{\mathsf{TS}_0})$, she would also be rationally committed to doubting the overall significance or competence of c.

To spell out the basic idea, let us start with a warm up example. Consider the following cognitive project c_{EA}, different from Hypatia's current cognitive project c. When Hypatia sets out to undertake c_{EA}, she begins with a currently held justified belief in the *axioms* of EA (c_{EA} is like her current cognitive project c in this respect). But unlike the goal of her current cognitive project c, Hypatia's cognitive goal in successfully carrying out c_{EA} is only to arrive at a justified belief in the entire *theory* of EA (her cognitive goal in successfully carrying out c is to arrive at a justified belief in the entire theory of Predicative Analysis).

Hypatia's cognitive project c_{EA} certainly seems well-defined, insofar as there does seem to be a difference between her justified belief in the axioms of EA, and her justified belief in the entire theory of EA. Again, this does not depend on the particular methods by which Hypatia arrived at a justified belief in the axioms of EA. (As before, for concreteness, we may suppose she came to hold a justified belief in the axioms of EA because the natural number structure is somehow intuitable to her, and she verified that the axioms of EA hold in the standard model.)

Then at that particular point in time, Hypatia holds a justified belief in all and only the axioms of EA. She does not, for example, yet hold a justified belief in an arbitrary disjunction, one million disjuncts long, one of which is one of the axioms of EA. Indeed, by taking a large enough number of disjuncts we may assume Hypatia has never even *thought* about such a disjunction, in which case there is little sense in the idea that she holds a justified belief in it.

How, then, *would* Hypatia come to hold a justified belief in this theorem of EA? A plausible suggestion, we claim, is that she would exhibit a proof of this theorem, say $\Gamma \Rightarrow \Delta$, from the axioms of EA.[8] But then Hypatia's justified belief in $\Gamma \Rightarrow \Delta$ is grounded in *more* than just her justified belief in the *axioms* of EA. It is Hypatia's reliance on the inference rules of classical logic when she writes down proofs from the axioms of EA, coupled with her justified belief in the axioms of EA, by which she arrives at a justified belief in $\Gamma \Rightarrow \Delta$.[9] Thus, the difference between Hypatia's justified belief in the axioms of EA, and her justified belief in the entire theory of EA, is this: on one hand, her justified belief in the axioms of EA arrives via some independent traditional justificatory method. But in general, her justified belief in the theory of EA arrives via her justified belief in the axioms of EA, and her reliance on the inference rules of classical logic when she writes down proofs from the axioms of EA.

So, consider a scenario in which Hypatia writes down a proof of some theorem $\Gamma \Rightarrow \Delta$ of EA from the axioms of EA (we suppose $\Gamma \Rightarrow \Delta$ is not itself an axiom of EA). The point is that for Hypatia's justified belief in the axioms of EA to transfer to $\Gamma \Rightarrow \Delta$, she requires collateral warrant for the following principle: if $\Gamma \Rightarrow \Delta$ is provable from the axioms of EA, then $\Gamma \Rightarrow \Delta$ holds.

[8] Or at least convince herself that such a proof could, in principle, be written down.

[9] A similar story is told in (Horsten, 2021) with respect to the consistency of a theory.

This last principle is nothing more than an informal expression of Hypatia's reliance on the inference rules of classical logic when she constructs proofs in EA. Let us say that it is an informal expression of Hypatia's *trust* in the theory EA. It is *why* she must think that any $\Gamma \Rightarrow \Delta$ provable from the axioms of EA inherits her justified belief. Put another way, if Hypatia is to infer a justified belief in $\Gamma \Rightarrow \Delta$ on the basis of her proof coupled with her justified belief in the principles of TS_0, then she must trust that proof in EA is capable of delivering statements in which she is to come to hold a justified belief.

But the natural way of formalizing the informal principle for which Hypatia requires collateral warrant *is* just the proposition asserting the validity of the following reflection rule (r_{EA}):

$$\frac{\Rightarrow \mathrm{Pr}_{\mathsf{EA}}(\ulcorner \Gamma \Rightarrow \Delta \urcorner)}{\Gamma \Rightarrow \Delta}\ (\mathrm{r}_{\mathsf{EA}})$$

Thus, if Hypatia infers justified belief in $\Gamma \Rightarrow \Delta$ on the basis of her justified belief in the axioms of EA and her reliance on the inference rules of classical logic when she constructs proofs in EA, then she presupposes the validity of (r_{EA}). If she were to doubt the validity of (r_{EA}) (where doubt might include even agnosticism about the validity of (r_{EA})), she would be rationally committed to doubting the significance or competence of the cognitive project c_{EA}. For the success of c_{EA} requires her to hold a *justified belief* in arbitrary theorems of EA. And justified belief excludes doubt.

Finally, we claim that Hypatia's cognitive project c_{EA} is embedded in her more ambitious cognitive project c. By that, we mean that the cognitive goal of c_{EA} is necessary for the cognitive goal of c. For the axioms of EA form the basis of Hypatia's cognitive project c – it is those axioms which she first comes to hold a justified belief in. If she could not come to hold a justified belief in the principles of the *theory* EA, how can she expect to achieve anything further, and ultimately justify new mathematical beliefs in the principles of Predicative Analysis on the basis of her justified belief in the axioms of EA?

Putting all this together, we claim to have offered a plausible understanding of why the validity of (r_{EA}) is a presupposition of Hypatia's larger cognitive project c. For the cognitive goal of Hypatia's narrower cognitive project c_{EA} is necessary for the cognitive goal of her cognitive project c, and (r_{EA}) is a presupposition of c_{EA}. Thus, (r_{EA}) is a presupposition of c. This concludes our warm up example.

Now let us turn to the validity of ($\mathrm{r}_{\mathsf{TS}_0}$), which we are currently investigating. We claim that a roughly similar line of reasoning to the above serves to show why ($\mathrm{r}_{\mathsf{TS}_0}$) is a presupposition of Hypatia's cognitive project c. First, we may suppose that there is a moment during Hypatia's JMB process at which her *warranted belief* in the *axioms* of the theory TS_0 propagates to general *justified belief* in the *theory* TS_0. We say "warranted belief" here, rather than "justified belief," because while Hypatia's warrant for the axioms of EA is justified belief, her warrant for the sequents (T1) and (T2) is mere entitlement (according to Fischer et al. (2021)). But by stage (4) of JMB, she is *justified* in believing the principles of the theory TS_0. This moment corresponds to nothing more than Hypatia's successful undertaking of the following cognitive project c_{TS_0}, different from her current cognitive project c. When Hypatia sets out to undertake c_{TS_0}, she begins with warranted belief in the *axioms* of TS_0. But unlike the goal of her current cognitive project c, Hypatia's cognitive goal in successfully carrying out c_{TS_0} is only to arrive at a justified belief in the *theory* of TS_0 (where the *theory* of TS_0 is understood to exclude the trivial theorems, i.e., the axioms of TS_0. For Hypatia only has entitlement to some of the axioms of TS_0, namely (T1) and (T2)).

Then analogously to our warm up example, it is Hypatia's reliance on the inference rules of classical logic when she constructs proofs in TS_0, in conjunction with her warrant for the axioms of TS_0, by which she arrives at a justified belief in an arbitrary theorem of TS_0. Hypatia infers justified belief in an arbitrary theorem $\Gamma \Rightarrow \Delta$ of TS_0 on the basis of her warranted belief in the axioms of TS_0 and her reliance on the inference rules of classical logic when she constructs proofs in TS_0. Thus, for Hypatia's warranted belief in the axioms of TS_0 to transfer to an arbitrary theorem $\Gamma \Rightarrow \Delta$ of TS_0, she requires collateral warrant for the following informal proposition: if $\Gamma \Rightarrow \Delta$ is provable from the axioms of TS_0, then $\Gamma \Rightarrow \Delta$ holds.

But the natural way of formalizing this informal claim *is* just the proposition p asserting the validity of the reflection rule ($\mathrm{r}_{\mathsf{TS}_0}$). Thus, if Hypatia infers justified belief in $\Gamma \Rightarrow \Delta$ on the basis of her warranted belief in the axioms of TS_0 and her reliance on the inference rules of classical logic when she constructs proofs in TS_0, then she presupposes the validity of ($\mathrm{r}_{\mathsf{TS}_0}$). So the proposition asserting the validity of ($\mathrm{r}_{\mathsf{TS}_0}$) is a presupposition of c_{TS_0}.

But achieving the goal of Hypatia's narrower cognitive project c_{TS_0} is necessary for achieving the goal of her overall cognitive project c. As before: if Hypatia could not come to hold a justified belief in the principles of the *theory* TS_0, and TS_0 is the first theory extending EA that she encounters in her JMB process, how can she expect to achieve anything further, and ultimately justify new mathematical beliefs in the principles of Predicative Analysis on the basis of her justified belief in the axioms of EA? Putting all this together, we have argued that achieving the goal of Hypatia's narrower cognitive project c_{TS_0} is necessary for achieving the goal of c, and the proposition asserting the validity of $(\mathrm{r}_{\mathsf{TS}_0})$ is a presupposition of c_{TS_0}. Thus, the proposition asserting the validity of $(\mathrm{r}_{\mathsf{TS}_0})$ is a presupposition of c.

We note that if what we have said is correct, then similar lines of reasoning serve to show why the validity of *any* of the reflection rules Hypatia comes to adopt during her JMB process is a presupposition of her cognitive project c. For each of those reflection rules $(\mathrm{r}_{\mathsf{S}})$ corresponds to a theory S which Hypatia meets during her JMB process, such that her warranted belief in the axioms of S propagates to a general justified belief in the (non-trivial) theorems of S. And for her warranted belief in the axioms of S to propagate to a general justified belief in the (non-trivial) theorems of S, she must presuppose the validity of $(\mathrm{r}_{\mathsf{S}})$.

Here we reach an important juncture. Let us offer an alternative reconstruction of the line of reasoning above. Given what we said, we might instead frame the situation as a Wrightean I–III line of reasoning (from Hypatia's perspective), as in section 3.2.1:

(a) $\Rightarrow \mathrm{Pr}_{\mathsf{TS}_0}(\ulcorner \Gamma \Rightarrow \Delta \urcorner)$.

(b) $\Gamma \Rightarrow \Delta$.

(c) $(\mathrm{r}_{\mathsf{TS}_0})$ is valid.

What we said above is essentially this: Hypatia infers a justified belief in (b) on the basis of (a), coupled with her justified belief in the axioms of TS_0. But coupled with her justified belief in the axioms of TS_0, the evidence (a) provides for (b) requires collateral warrant for (c). Thus, if Hypatia were to doubt the proposition asserting the validity of ($\mathrm{r}_{\mathsf{TS}_0}$), she would be rationally committed to doubting the significance or competence of her narrower cognitive project c_{TS_0}. Yet achieving the cognitive goal of c_{TS_0} is necessary for achieving the cognitive goal of c. Thus, the proposition asserting the validity of ($\mathrm{r}_{\mathsf{TS}_0}$) is a presupposition of c.[10]

Now we note an important point: there is a fundamental difference between the reflection principle ($\mathrm{r}_{\mathsf{TS}_0}$), and the kind of Wrightean cornerstone propositions we encountered in section 3.2.1. For comparison, consider the sort of Cartesian line of reasoning we considered in section 3.2.1:

(i′) My current experience is in all respects as if there is a hand in front of me.

(ii′) There really is a hand in front of me.

(iii′) I am not now in the midst of a lucid and persistent dream.

Our observation is that there is a crucial difference between (c) and the cornerstone proposition (iii′) above. To explain the difference, let us characterize the method by which each type-I proposition transmits evidence to its

[10] We do not mean to suggest that representing the situation this way shows that the validity of ($\mathrm{r}_{\mathsf{TS}_0}$) is an *entitlement* of Hypatia's cognitive project c at this point. For we haven't yet said anything about clause 3 of the definition of entitlement of cognitive project. But representing the situation this way does give us a good feel for the idea.

corresponding type-II counterpart by calling it the *current means for providing justification.*[11] In (a)–(c) above, the theory TS_0 is Hypatia's current means for providing justification. If she writes down a proof of a theorem from the axioms of TS_0, Hypatia infers a justified belief in that theorem. In (i′)–(iii′) above, we might say that the current means for providing justification consists of our best empirical theories, broadly construed. I check, in perfectly ordinary empirical ways, that my current experience is in all respects as if there is a hand in front of me (Perhaps I check that the current lighting conditions are optimal, or check the proper functioning of my ocular faculties by consulting an optometrist, etc.)

Then the difference between (c) and the cornerstone proposition (iii′) above, is that the rule ($\mathrm{r}_{\mathsf{TS}_0}$), when added to Hypatia's current means TS_0 for providing justification, yields *a host* of new mathematical consequences (since $\mathsf{TS}_0 + (\mathrm{r}_{\mathsf{TS}_0})$ is not conservative over TS_0). But the cornerstone proposition "I am not now in the midst of a lucid and persistent dream," considered in conjunction with one's current realm of means for providing justification in the Cartesian scenario, which consists of our best empirical theories, yields *no* new empirical consequences.[12] Our entitlement to (iii′) above serves to rescue all of the existing empirically justified beliefs we already think we hold. But something different happens when Hypatia is entitled to (c) above: *new beliefs enter the picture as a result of her entitlement.* Consider an arithmetical con-

[11] This aligns with our earlier remarks about justification in chapter 3.

[12] There seem to be other differences between (c) and the cornerstone proposition (iii′). In the Wrightean scenario, according to the Cartesian skeptic, our inability to justify (iii′) results in something we are supposed to find paradoxical. But there doesn't seem to be anything *paradoxical* about the idea that Hypatia cannot justify ($\mathrm{r}_{\mathsf{TS}_0}$) on the basis of her justified belief in the principles TS_0 (and e.g., her reliance on classical logic).

sequence of the theory $\mathsf{TS}_0 + (\mathrm{r}_{\mathsf{TS}_0})$ which is not a consequence of TS_0 alone, like Con(EA). The source of Hypatia's warrant for Con(EA) lies with $(\mathrm{r}_{\mathsf{TS}_0})$. For it is only when $(\mathrm{r}_{\mathsf{TS}_0})$ is added to TS_0 that Con(EA) enters the picture. The emergent question, aligning with our remarks at the end of chapter 3, and to which we will return in section 4.2, is: if Hypatia's warrant for $(\mathrm{r}_{\mathsf{TS}_0})$ is mere entitlement to believe, then why isn't she also merely *entitled* to believe Con(EA)? And more generally: if Hypatia's warrant for $(\mathrm{r}_{\mathsf{TS}_0})$ is mere entitlement to believe, then why isn't she also merely *entitled* to believe in the principles of $\mathsf{TS}_0 + (\mathrm{r}_{\mathsf{TS}_0})$?

We will return to answer these questions in due course. At this point, we have on the table a story about why it might be plausible to think that p is a presupposition of c. For now, let us finish off our discussion of p's being an entitlement of cognitive project c.

Infinite justificatory regress

Consider clause 3 of the definition of entitlement of cognitive project. Would Hypatia's attempt to justify p involve further presuppositions in turn of no more secure a prior understanding (and so on without limit), such that if she accepted that there is nevertheless an onus to justify p, she would implicitly undertake a commitment to an infinite regress of justificatory projects, each concerned to vindicate the presuppositions of its predecessors?

This is a tricky question to answer. To get a better handle on it, we first introduce a warm up example, before returning to Hypatia's context. Setting $(\mathrm{r}_{\mathsf{TS}_0})$ to one side for the moment, consider instead the proposition asserting the validity of the reflection rule (r_{EA}) for EA:

$$\frac{\Rightarrow \mathrm{Pr}_{\mathsf{EA}}(\ulcorner \Gamma \Rightarrow \Delta \urcorner)}{\Gamma \Rightarrow \Delta}\,(\mathrm{r}_{\mathsf{EA}})$$

Let us suppose we are currently engaged in a certain cognitive project. Suppose that our starting point (like Hypatia) is a currently held justified belief in the principles of EA. Let us suppose further that our specific cognitive goal is to arrive at a justified belief in the principles of the stronger theory PA (this is much less ambitious than Hypatia's overall goal). For simplicity, let us also suppose that we are only considering reflection for EA as a means of achieving our goal (so we set truth aside, for the time being). Let us denote this cognitive project by c_{PA}. Our undertaking c_{PA} therefore looks like this:

(1) We start out with a justified belief in the principles of the theory EA.

$(\vdots)$ $\vdots$

(n) We thereby arrive at a justified belief in the principles of the theory PA.

where we are going to leverage reflection principles to fill in the $\vdots$ and get us from start to finish.

Now, logically, our JMB-like process is a relatively simple one, due to the following result of Kreisel and Lévy (1968):

Theorem 4. $\mathsf{EA} + (\mathrm{r}_{\mathsf{EA}}) \equiv \mathsf{PA}$.

Thus, by adding (r_{EA}) to EA, logically, we reach PA in one step. It remains to tell an epistemic story about how our justified belief in EA leads to justified belief in the principles of PA. The first question we face is: what is our warrant for adopting (r_{EA})?

Suppose we are considering whether our warrant for adopting (r_{EA}) is entitlement of our current cognitive project c_{PA}. In particular, we want to know whether our attempt to justify the validity of (r_{EA}) involves further presuppositions of c_{PA} in turn of no more secure a prior understanding (and so on without limit), such that if we accepted that there is nevertheless an onus to justify the validity of (r_{EA}), we would implicitly undertake a commitment to an infinite regress of justificatory projects, each concerned to vindicate the presuppositions of its predecessors.

Before we explore the idea of regress, first let us leverage some of our earlier discussion about the presuppositional status of (the validity of) reflection rules, since this will be important for us. In the spirit of that discussion, consider again the narrower-than-c_{PA} cognitive project c_{EA}, which we begin with our currently held justified belief in the *axioms* of EA, and whose cognitive goal is to arrive at a justified belief in the entire *theory* of EA. By what we said above, this narrower cognitive project c_{EA} is such that: (1) the proposition asserting the validity of the corresponding reflection rule (r_{EA}) is a presupposition of it, and (2) achieving the goal of the narrower cognitive project c_{EA} is necessary for achieving the goal of the overall cognitive project c_{PA}. For if our justified beliefs fail to propagate from the axioms of EA to the theory of EA, then the prospects for our more ambitious cognitive project c_{PA} are surely in doubt.

Now, we want to generalize this scenario. So, let S be *any* of the theories that we meet during our JMB-like process (in this case, we have seen that there are two: EA and PA). Analogously to what we said for EA, any cognitive project in which we currently hold a justified belief in the *axioms* of S, and the cognitive goal of which is to arrive at a justified belief in the entire *theory* of S, is such that: (1) the proposition asserting the validity of the corresponding reflection rule (r_S) is a presupposition of such a narrower cognitive project, and (2) achieving the goal of the narrower cognitive project is necessary for achieving the goal of our overall cognitive project. For if our justified beliefs fail to propagate at *any* stage of our JMB-like process, it seems fair to say that the prospects for our larger cognitive project are in doubt. The point here is that the validity of (r_S) propagates our justified belief from the axioms of S to the entire theory of S. Thus, for our justified belief to be preserved during our JMB-like process, this must happen for *every* theory S we meet during that process. So, if S is any theory that we meet during our JMB-like process, then the proposition asserting the validity of the corresponding rule (r_S) is a presupposition of our current cognitive project.

Next let us approach the idea of regress. In particular, let us try and reconstruct an attempt at justifying the validity of (r_EA) in the spirit of our examples from section 4.1.1. So, suppose we try to justify $\mathrm{Pr}_\mathsf{EA}(\varphi) \models \varphi$.

Fix a model $\mathcal{M}$ such that $\mathcal{M} \models \mathrm{Pr}_\mathsf{EA}(\varphi)$. We will suppose for simplicity that $\mathcal{M}$ is a standard model. Aligning with our earlier examples, the next thing to do is to say what our assumption means. In the cases of modus ponens and conjunction elimination, when we said what our assumptions meant, we invoked a metalinguistic version of the very connective occurring in the inference we set out to justify in the first place. In the case of the sequents (T1) and (T2), we argued that we can arrive at our conclusion by doing no such thing.

In the current context of reflection, something interesting happens. For the assumption $\mathcal{M} \models \mathrm{Pr}_{\mathsf{EA}}(\varphi)$ means that in $\mathcal{M}$, there is a proof of φ from the axioms of EA. We want to conclude that $\mathcal{M} \models \varphi$. The natural way of obtaining this conclusion is the following: by induction over $\mathcal{M}$ on the length of proofs, we obtain the soundness theorem for EA: if $\mathsf{EA} \vdash \varphi$, then $\mathsf{EA} \models \varphi$. Since $\mathsf{EA} \vdash \varphi$, we have $\mathsf{EA} \models \varphi$. Then since EA is true in $\mathcal{M}$, we conclude $\mathcal{M} \models \varphi$.

Let us pause and consider what has happened here. What we have done is *proved* that $\mathcal{M} \models \varphi$ in order to conclude that $\mathcal{M} \models \varphi$. In particular, we have relied on the validity of *some* kind of inference rule, of a piece with the rule (r_{EA}) we set out to justify in the first place. Let us write:

$$\frac{\Rightarrow \mathrm{Pr}_{\mathsf{S}}(\ulcorner \Gamma \Rightarrow \Delta \urcorner)}{\Gamma \Rightarrow \Delta}\ (\mathrm{r}_{\mathsf{S}})$$

for the formal counterpart to the inference rule we relied on during our reasoning. The key observation is that the rule (r_{S}) *cannot* be (r_{EA}) itself. For EA cannot derive the statement of its own soundness, yet we relied on the soundness of EA during our metatheoretic proof. So when we gave our metatheoretic proof, we were implicitly employing a theory S strictly stronger than EA itself. In other words, we were not reasoning using an informal metatheoretic version of EA. Thus, our reasoning has involved a formal jump: we set out to justify the validity of (r_{EA}), and in doing so, we relied on the validity of a(n) (informal) version of the rule (r_{S}), for some theory S stronger than EA.

So, whether this counts as a regress, per condition 3 of the definition of entitlement of cognitive project, depends (at least) on whether the proposition asserting the validity of (r_{S}) is a presupposition of our current cognitive project.

We argue that this in turn depends crucially on the theory S and what we will refer to as the *scope* of the cognitive project one is currently engaged in.[13] For the time being, think of the scope of our cognitive project as the strongest theory in which we are hoping to achieve a justified belief by successfully carrying it out. Thus, the scope of our current cognitive project c_{PA} is PA.

To spell out the details, we argue that if S is PA itself, then the proposition asserting the validity of $(\mathrm{r}_{\mathsf{S}})$ is a presupposition of our current cognitive project c_{PA}. But if S lies well outside the scope of our cognitive project, say S is the theory ZF, then the proposition asserting the validity of $(\mathrm{r}_{\mathsf{S}})$ is not a presupposition of our current cognitive project c_{PA}. Thus, in the context of our current cognitive project c_{PA}, if S is PA, then our attempt to justify the validity of $(\mathrm{r}_{\mathsf{EA}})$ starts to look regressive in Wright's sense. But if S is ZF, we claim that we do not regress, when we attempt to justify the validity of $(\mathrm{r}_{\mathsf{EA}})$.

So let S be PA. The first thing we ought to check is whether PA *is* strong enough to derive the soundness theorem for EA, but this follows from Theorem 4 above. Thus, it makes sense to think that our metatheoretic reasoning took place using PA. But now we recall our earlier observation: that if S is a theory we meet during our JMB-like process, then the proposition asserting the validity of the corresponding rule $(\mathrm{r}_{\mathsf{S}})$ is a presupposition of c_{PA}. PA is a theory we meet during our JMB-like process (it is the final theory we meet). Thus, the proposition asserting the validity of the corresponding rule $(\mathrm{r}_{\mathsf{PA}})$ is a presupposition of c_{PA}. So in our attempt to justify the validity of $(\mathrm{r}_{\mathsf{EA}})$, when we relied on the validity of the corresponding metatheoretic version of

[13] We will make the idea of scope precise in chapter 5.

the rule (r_{PA}), we appealed to a presupposition of our current cognitive project c_{PA}. We are happy to agree that our understanding of (r_{PA}) rests on no more secure a foundation than our understanding of (r_{EA}) itself. We are also happy to agree that if we were to accept an onus to justify the validity of (r_{EA}), we appear to thereby implicitly undertake a commitment to justify the further presupposition of c_{PA} which asserts the validity of the rule (r_{PA}). So, we appear to be in danger of regress in Wright's sense.

But let us push the idea of an *infinite* regress of this sort. It is clear that if we were to reconstruct a similar attempt to justify the validity of the rule (r_{PA}), we would rely on the validity of a metatheoretic version of a further reflection rule (r_{S}), where now S is strong enough to be able to derive the soundness of PA. But if *that* theory S is such that the proposition asserting the validity of the rule (r_{S}) is *not* a presupposition of our current cognitive project, then it appears that our justificatory regress has consisted of no more than one step. This looks like a way of avoiding the sort of regress per clause 3 of the definition of entitlement of cognitive project c_{PA}. For we would no longer appeal to a further *presupposition* of c_{PA}. Thus, we may hope to avoid the sort of infinite justificatory regress per clause 3 of the definition of entitlement of cognitive project. (*Even if* our understanding of the proposition asserting the validity of (r_{S}) *does* rest on no more secure a foundation than our understanding of the proposition asserting the validity of (r_{EA}), and so on.) In the context of our current cognitive project c_{PA}, (justified belief in the theory) PA is the limit of our potential cognitive achievement. So *any* such theory S would provide a way out of the sort of regress we appear to be falling into. Indeed, ZF is such a theory. So next, let us explore what happens with respect to our attempted justification of the validity of (r_{EA}), when S is the theory ZF, rather than PA.[14]

Let S be ZF. ZF is indeed strong enough to derive the soundness of EA. For example, ZF derives each of the axioms of PA, and PA is strong enough to derive the soundness theorem for EA. So it makes sense to think that our metatheoretic reasoning took place using ZF. But the crucial idea is that ZF is *not* a theory that we meet during our current cognitive project c_{PA}. This is simply an artifact of the goal of c_{PA}: all we hoped to reach (mathematically speaking) when we set out to undertake c_{PA} was the theory PA (and epistemically, we hoped to somehow arrive a justified belief in the principles of PA). Thus, (full) ZF lies far beyond the scope of c_{PA}. What reason do we have to think that the proposition asserting the validity of the rule (r_{ZF}), whose metatheoretic version we relied on during our attempt at justifying the validity of the rule (r_{EA}), is a presupposition of c_{PA}?

We claim that there is no such reason. We have seen what kind of cognitive project the validity of the rule (r_{ZF}) *would* count as a presupposition for – any cognitive project whose success depends on justified belief in the axioms of ZF propagating to the entire theory of ZF. But we do not require this in order for our current cognitive project c_{PA} to succeed. Indeed, suppose we were to doubt the proposition asserting the validity of (r_{ZF}). In what sense would it be *irrational* if we were to nonetheless maintain that c_{PA} is still significant or competent? We do not think it *would* be irrational. The success of c_{PA} does not depend on our epistemic attitude towards the proposition asserting the

[14] This is equivalent, for all intent and purpose, to first justifying the validity of (r_{EA}) using PA in the metatheory, and *then* justifying the validity of (r_{PA}) using ZF in the metatheory.

validity of the rule (r_{ZF}). And in general, the success of c_{PA} does not depend on our epistemic attitude towards propositions asserting the validity of rules of the form (r_{S}), where S is *any* theory which lies beyond the scope of what we are hoping to achieve by undertaking c_{PA}. Irrationality only emerges in these contexts when we simultaneously doubt and believe (or justifiably believe) some principle we meet at some point during our cognitive project. For our understanding of doubt excludes both (entitled) belief and justified belief. But the success of c_{PA} does not rationally require us to believe (even justifiably) that the rule (r_{ZF}) is valid. So, while it might be a strange position to hold, we claim that it is perfectly possible to doubt the validity of the rule (r_{ZF}) and rationally maintain that c_{PA} is still significant or competent. The proposition asserting the validity of the rule (r_{ZF}) is not a presupposition of c_{PA}.

Putting everything together, our appeal to the validity of the rule (r_{ZF}) in our attempted justification of the validity of the rule (r_{EA}) does not involve a further presupposition of our cognitive project c_{PA}. So even if our understanding of the proposition asserting the validity of the rule (r_{ZF}) rests on no more secure a foundation than our understanding of the proposition asserting the validity of the rule (r_{EA}), and so on, our attempted justification of the validity of (r_{EA}) fails to meet clause 3 of the definition of entitlement of our current cognitive project c_{PA}. We are perfectly able to justify the validity of the rule (r_{EA}) in a non-regressive way, by stepping far enough outside the scope of c_{PA}. This concludes our warm up example.

Now, let us return to Hypatia's context. The goal of her current cognitive project is to arrive at a justified belief in the principles of Predicative Analysis.

Would Hypatia's attempt to justify p (the proposition asserting the validity of ($\mathrm{r}_{\mathsf{TS}_0}$)) involve further presuppositions of her cognitive project in turn of no more secure a prior understanding (and so on without limit), such that if she accepted that there is nevertheless an onus to justify p, she would implicitly undertake a commitment to an infinite regress of justificatory projects, each concerned to vindicate the presuppositions of its predecessors? Logically, the details are not as simple as in our warm up example, but the underlying idea is the same. We argue that Hypatia is perfectly able to justify the validity of the rule ($\mathrm{r}_{\mathsf{TS}_0}$) in a non-regressive way, by stepping far enough outside the scope of her cognitive project.

Analogously to our warm up example, let us leverage some of our earlier discussion. Above, we considered one of Hypatia's narrower cognitive projects c_{TS_0}, which she begins with a currently held warranted belief in the *axioms* of TS_0, and whose cognitive goal is to arrive at a justified belief in the entire *theory* of TS_0. We argued that c_{TS_0} is such that: (1) the proposition asserting the validity of the corresponding reflection rule ($\mathrm{r}_{\mathsf{TS}_0}$) is a presupposition of c_{TS_0}, and (2) achieving the goal of c_{TS_0} is necessary for achieving the goal of c. Like before, we want to generalize this scenario. So, let S be *any* of the theories that Hypatia meets during her JMB process (there are transfinitely many such theories). Then *any* cognitive project c_{S}, which Hypatia begins with a warranted belief in the *axioms* of S, and the cognitive goal of which is to arrive at a justified belief in the entire *theory* of S, is such that: (1) the proposition asserting the validity of the corresponding reflection rule (r_{S}) is a presupposition of c_{S}, and (2) achieving the goal of the narrower cognitive project c_{S} is necessary for achieving the goal of Hypatia's larger cognitive project c. For if Hypatia's justified beliefs fail to propagate at *any* stage of her JMB process, it seems fair to say that the prospects for her larger cognitive project are in doubt. Again, the point is that if S is a theory that Hypatia meets during her JMB process, then the proposition asserting the validity of the corresponding rule (r_{S}) is a presupposition of c.

Now let us try and reconstruct Hypatia's attempt at justifying the validity of ($\mathrm{r}_{\mathsf{TS}_0}$). So suppose Hypatia wants to justify $\mathrm{Pr}_{\mathsf{TS}_0}(\varphi) \models \varphi$. She reasons as follows: fix a standard model $\mathcal{M}$ such that $\mathcal{M} \models \mathrm{Pr}_{\mathsf{TS}_0}(\varphi)$. This means that in $\mathcal{M}$, there is a proof of φ from the axioms of TS_0. We want to conclude that $\mathcal{M} \models \varphi$. The natural way of obtaining this conclusion is the following: by induction over $\mathcal{M}$ on the lengths of proofs, we obtain the soundness theorem for TS_0: if $\mathsf{TS}_0 \vdash \varphi$, then $\mathsf{TS}_0 \models \varphi$. Since $\mathsf{TS}_0 \vdash \varphi$, we obtain $\mathsf{TS}_0 \models \varphi$. Then since TS_0 is true in $\mathcal{M}$, we conclude $\mathcal{M} \models \varphi$.

Again, what Hypatia has done is *proved* that $\mathcal{M} \models \varphi$ in order to conclude that $\mathcal{M} \models \varphi$. In particular, she has relied on the validity of *some* kind of inference rule, of a piece with the rule ($\mathrm{r}_{\mathsf{TS}_0}$) whose validity she set out to justify in the first place. Let us write:

$$\frac{\Rightarrow \mathrm{Pr}_{\mathsf{S}}(\ulcorner \Gamma \Rightarrow \Delta \urcorner)}{\Gamma \Rightarrow \Delta}\ (\mathrm{r}_{\mathsf{S}})$$

for the formal counterpart to the inference rule she relies on during her reasoning. Again, the key observation is that the rule (r_{S}) *cannot* be ($\mathrm{r}_{\mathsf{TS}_0}$) itself. For TS_0 cannot prove the statement of its own soundness, yet Hypatia relied on the soundness of TS_0 during her metatheoretic proof. So when she gave her metatheoretic proof, Hypatia was implicitly employing a theory S strictly stronger than TS_0 itself. Hypatia's reasoning has involved a formal jump: she set out to justify the validity of $(\mathrm{r}_{\mathsf{TS}_0})$, and in doing so, she relied on the validity of a rule $(\mathrm{r}_{\mathsf{S}})$, for some theory S strictly stronger than TS_0.

So, whether this counts as a regress, per clause 3 of the definition of entitlement of Hypatia's current cognitive project, depends on whether the validity of $(\mathrm{r}_{\mathsf{S}})$ is a presupposition of her current cognitive project. Exploring the idea of regress, suppose the theory S which Hypatia boosts the theory TS_0 with just the kind of resources that would allow her to carry out her metatheoretic reasoning, so that her metatheoretic reasoning is performed using the theory $\mathsf{TS}_0 + (\mathrm{r}_{\mathsf{TS}_0}) = \mathrm{r}(\mathsf{TS}_0)$. We argued above that if S is a theory that Hypatia meets during her JMB process, then the proposition asserting the validity of the corresponding rule $(\mathrm{r}_{\mathsf{S}})$ is a presupposition of her current cognitive project c. Hypatia meets the theory $\mathrm{r}(\mathsf{TS}_0)$ during her JMB process (during the first iteration). Thus, the proposition asserting the validity of the corresponding reflection rule $(\mathrm{r}_{\mathrm{r}(\mathsf{TS}_0)})$ is a presupposition of her cognitive project c. But the validity of this reflection rule corresponds to the validity of Hypatia's inference in her metatheoretic reasoning, when she inferred that $\mathcal{M} \models \varphi$ on the basis of her proof of $\mathcal{M} \models \varphi$ in the theory $\mathrm{r}(\mathsf{TS}_0)$. Thus, Hypatia has regressed. In her attempt to justify the validity of the reflection rule $(\mathrm{r}_{\mathsf{TS}_0})$, she relied on the validity of the reflection rule $(\mathrm{r}_{\mathrm{r}(\mathsf{TS}_0)})$. But the proposition asserting the validity of the latter rule is a presupposition of her current cognitive project.

Now in fact, if Hypatia were to continue with this strategy, it looks like she *would* be led to an infinite regress of the sort specified by clause 3 of the definition of entitlement of her cognitive project c. To say why, let us simplify notation. Recall that for an axiomatizable theory S, the reflection $\mathrm{r}(\mathsf{S})$ on S is defined as the closure of S under the reflection rule $(\mathrm{r}_{\mathsf{S}})$:

$$\mathrm{r}(\mathsf{S}) = \mathsf{S} + (\mathrm{r}_{\mathsf{S}}).$$

We may then define iterations of reflection on S. For example, $\mathrm{r}(\mathrm{r}(\mathsf{S}))$ is the result of closing $\mathrm{r}(\mathsf{S})$ under the rule $(\mathrm{r}_{\mathrm{r}(\mathsf{S})})$. We denote $\mathrm{r}(\mathrm{r}(\mathsf{S}))$ by $\mathrm{r}^2(\mathsf{S})$.

Continuing Hypatia's justificatory attempt, suppose now that she attempts to justify the validity of the rule $(\mathrm{r}_{\mathrm{r}(\mathsf{TS}_0)})$ by boosting the theory $\mathrm{r}(\mathsf{TS}_0)$ with exactly the resources that would allow her to carry out a proof of the validity of the rule $(\mathrm{r}_{\mathrm{r}(\mathsf{TS}_0)})$. That is, she moves from the theory $\mathrm{r}(\mathsf{TS}_0)$ to the theory $\mathrm{r}^2(\mathsf{TS}_0)$. Then since $\mathrm{r}^2(\mathsf{TS}_0)$ is a theory which Hypatia meets during her JMB process (during the second iteration), the proposition asserting the validity of the corresponding reflection rule $(\mathrm{r}_{\mathrm{r}^2(\mathsf{TS}_0)})$ is a presupposition of her cognitive project. Yet Hypatia relies on the validity of the rule $(\mathrm{r}_{\mathrm{r}^2(\mathsf{TS}_0)})$ in her attempt to justify the validity of the rule $(\mathrm{r}_{\mathrm{r}(\mathsf{TS}_0)})$. Thus, her attempt to justify the validity of the rule $(\mathrm{r}_{\mathrm{r}(\mathsf{TS}_0)})$ involves a further presupposition of her cognitive project. And so on, *ad infinitum*.

Our point here is that Hypatia's strategy – boosting her initial theory with reflection in order to try and justify the validity of the corresponding reflection rule for the initial theory – will *never* exceed the limits of her current cognitive project c, since transfinitely iterating the reflection operation on EA is, in part, how her JMB process is structured. This makes the current example different from the warm up example. In that example, we only needed two iterations of reflection on EA to exceed the limits of the cognitive project we defined there. But in the current example, if Hypatia proceeds with this strategy, then there is a danger of *infinite* regress.

Consequently, if Hypatia's attempt to justify the validity of the rule ($\mathrm{r}_{\mathsf{TS}_0}$) proceeds in this way, then her justificatory attempt does involve further presuppositions of her cognitive project c in turn of no more secure a prior understanding than the validity of the rule ($\mathrm{r}_{\mathsf{TS}_0}$). If she accepts an onus to justify the validity of the rule ($\mathrm{r}_{\mathsf{TS}_0}$), then proceeding in this way, she would be implicitly committed to an infinite regress of justificatory projects, each concerned to vindicate the presuppositions of its predecessor. So, if she attempted to justify the validity of the rule ($\mathrm{r}_{\mathsf{TS}_0}$) in this way, then the proposition asserting the validity of the rule ($\mathrm{r}_{\mathsf{TS}_0}$) looks like it meets clause 3 of the definition of entitlement of her cognitive project c.

But thinking back to our warm up example, it should also be clear that Hypatia can also justify the validity of the rule ($\mathrm{r}_{\mathsf{TS}_0}$) in another way, using a strong enough theory which exceeds the limits (Predicative Analysis) of her current cognitive project c. ZF is just such a theory. Full ZF lies far beyond the logical scope of Hypatia's cognitive project. We have argued that there is no reason to think that the proposition asserting the validity of the corresponding reflection rule (r_{ZF}) is a presupposition of Hypatia's current cognitive project. For the success of c does not rationally depend on her epistemic attitude toward the validity of the rule (r_{ZF}). Thus, if all she relies on in her metatheoretic reasoning is the validity of a metatheoretic version of the rule (r_{ZF}), Hypatia is perfectly able to justify the validity of the rule ($\mathrm{r}_{\mathsf{TS}_0}$) without involving a further presupposition of her cognitive project c. If this is the case, we conclude that Hypatia's attempt to justify the validity of the rule ($\mathrm{r}_{\mathsf{TS}_0}$) does not meet clause 3 of the definition of entitlement of her cognitive project.

The point of this long discussion is this: whether Hypatia's attempt to justify p involves further presuppositions in turn of no more secure a prior understanding (and so on without limit), depends crucially on the scope of her cognitive project c. Given the structure of Hypatia's current cognitive project, if she is not careful about her choice of metatheory, she does face the possibility of infinite justificatory regress. But there are choices for her metatheory that do seem to allow Hypatia to avoid infinite justificatory regress. In particular, if ZF is on the table as a candidate metatheory for Hypatia, we think p fails to meet clause 3 of the definition of entitlement of her cognitive project c.

Here we reach another important juncture in our argument. We have referred to ZF repeatedly in what we said above. But what reason do we have to think that ZF *is* on the table for Hypatia? Does she, for example, hold a justified belief in ZF, so it would make sense to think of ZF as a plausible metatheoretic candidate in the lines of reasoning above; a plausible means for providing justification?

In fact, there is a passage in (Fischer et al., 2021) which *does*, on the face of things, suggest that ZF is available to Hypatia as a means for providing justification; in particular, for justifying new mathematical beliefs in the principles of Predicative Analysis:

> there are also other ways in which Hypatia may come to accept Predicative Analysis. For instance, she may straightaway, i.e.,

> without going through the iterative reflection process described above, acquire a belief in Zermelo-Fraenkel set theory, perhaps by coming to understand and accept a version of the iterative conception of set. If ZF can indeed be justified from the iterative conception, then Hypatia can in this way come to know a mathematical theory that is much stronger than Predicative Analysis. (Fischer et al., 2021, p. 79)

The possibility of Hypatia coming to hold an independently justified belief in the principles of ZF is interesting, and it is a point to which we will return. On the face of things, it is also peculiar, and raises a few questions. Among these are: if Hypatia can justify new mathematical beliefs (in a theory that far exceeds the strength of EA) using methods other than a JMB-like process, isn't her current cognitive project undermined, in some sense? We are offered an answer in (Fischer et al., 2021). The thought is that no, her current cognitive project is not undermined. The reason, we are told, is that extending the scope of our mathematical knowledge by independently coming to hold a justified belief in the principles of ZF

> differs structurally from extension by reflection. In order to accept a new axiom (strong principle of infinity, combinatorial principle), we need to do justificatory work, whereas no new justification is needed to adopt the global reflection rule for a theory that you are already justified to believe in.

Thus, the idea is that Hypatia, armed only with her justified belief in the principles of EA, has two alternative routes by which she might come to hold a justified belief in the principles of Predicative Analysis. Either she undertakes her current cognitive project via the process JMB, or she does independent justificatory work to come to hold a justified belief in the principles of an even stronger theory, like ZF. According to Fischer et al. (2021), the advantage of the former alternative is that Hypatia will not have to do any extra justificatory work – she gets by in that route on entitlements alone. Thus, Hypatia's current cognitive project is not undermined; in fact, it is an attractive alternative to justifying new mathematical beliefs via an independent route.

We highlight the key claim which has emerged at this point: that Hypatia has to carry out extra justificatory work if she arrives at newly justified beliefs in Predicative Analysis via an independently justified belief in the principles of ZF, but does not have to do so if she arrives at newly justified beliefs in Predicative Analysis starting out from a justified belief in the principles of EA and iterating the JMB process. If this claim is true (and if in addition there is a principled reason to think that Hypatia should avoid undertaking extra justificatory work if possible) then perhaps ZF is off the table after all. In this case, perhaps our justificatory attempts above which appeal to ZF are off the table too, so that the only hope Hypatia has of justifying p (the proposition asserting the validity of the rule $(\mathrm{r}_{\mathsf{TS}_0})$) is infinitely regressive in Wright's sense. If this claim is false though, then if what we have said above is correct, we seem to be left to wonder in what sense justifying p leads to justificatory regress.

So at this point, whether ZF is on the table for Hypatia as a means for providing justification hinges on the key claim above. In turn, whether any attempt to justify p leads to infinite justificatory regress in Wright's sense hinges on the key claim above. We will return to the key claim below. As a spoiler, we argue that the point about Hypatia having to undertake extra justificatory work is moot: we claim she has to do extra justificatory work *whichever* alternative route she proceeds by. This, we argue, places Hypatia's JMB process on an equal epistemic footing to her alternative route to a justified belief in Predicative Analysis via an independently justified belief in ZF. All things considered, we think this puts ZF back on the table for Hypatia as a means for providing justification. As a result, we think she is perfectly able to justify p in a way that does not meet Wright's clause 3.

Let's take stock of what we have said so far, before we move on.

4.1.3 Summary

Over sections 4.1.1 and 4.1.2, we have explored the extent to which we might think Hypatia's warrants for adopting the rules (T1) and (T2) for EA, and for adopting the rule ($\mathrm{r}_{\mathsf{TS}_0}$) for TS_0, are bona fide examples of entitlements of her cognitive project c.

That Hypatia has no sufficient reason to believe the sequents (T1) and (T2) are not valid seems plausible, and we attempted to tell a plausible story about why the proposition asserting the validity of (T1) and (T2) is a presupposition of c. Less plausible, we said, was the claim that any attempt to justify the validity of the sequents (T1) and (T2) commits Hypatia to infinite justificatory regress in Wright's sense. Our story was similar for reflection. That Hypatia has no sufficient reason to believe the rule ($\mathrm{r}_{\mathsf{TS}_0}$) is not valid seems plausible, and we also attempted to outline a plausible story about why the proposition asserting the validity of ($\mathrm{r}_{\mathsf{TS}_0}$) is a presupposition of c. During this discussion we reached an important juncture. We highlighted a fundamental difference between the reflection principle ($\mathrm{r}_{\mathsf{TS}_0}$), and the kind of Wrightean cornerstone propositions we encountered in section 3.2.1. The rule ($\mathrm{r}_{\mathsf{TS}_0}$), when added to Hypatia's current realm of means TS_0 for providing justification, yields a host of new mathematical consequences. But the Cartesian cornerstone proposition "I am not now in the midst of a lucid and persistent dream," considered in conjunction with one's current realm of means for providing justification in that scenario, yields no new empirical consequences. We left the following question open: what, then, is the epistemic status of these new consequences in Hypatia's scenario? We also left open the question of whether any attempt to justify the validity of the rule ($\mathrm{r}_{\mathsf{TS}_0}$) commits Hypatia to infinite justificatory regress in Wright's sense. This hinges on the key claim that she must do extra justificatory work if she is to arrive at newly justified mathematical beliefs by coming to hold an independently justified belief in the principles of ZF, but need not do any extra justificatory work if she arrives at newly justified mathematical beliefs via the JMB process.

With that, let us now turn to the validity of JMB. We are interested in the following question: if Hypatia's warrants for adopting the sequents (T1) and (T2), and for adopting the rule ($\mathrm{r}_{\mathsf{TS}_0}$) for TS_0, really are entitlements of cognitive project, *does* she thereby ultimately come to hold a *justified belief* in the principles of the theory $\mathsf{TS}_0 + (\mathrm{r}_{\mathsf{TS}_0})$? We argue the answer is no. Along the way, we also argue that the key claim above is false. Hypatia must do justificatory work either way. As a result, we are faced with a couple of problems: what, then, does Hypatia's warrant for adopting the principles (T1), (T2), and ($\mathrm{r}_{\mathsf{TS}_0}$), consist in? And what sort of warrant must this be, such that it can underwrite Hypatia's justified belief formation in JMB?

4.2 Validity of JMB

Let us refocus our discussion by recalling JMB, the first iteration of Hypatia's process:

(1) Hypatia starts out with a *justified belief* in the principles of the theory EA.

(2) Hypatia is *entitled* to adopt the rules (T1) and (T2) for EA.

(3) Hypatia thereby comes to *accept* the theory TS_0.

(4) Hypatia is thereby *justified in believing* the principles of the theory TS_0.

(5) Hypatia is *entitled* to adopt the rule ($\mathrm{r}_{\mathsf{TS}_0}$) for TS_0.

(6) Hypatia thereby comes to *accept* the theory $\mathsf{TS}_0 + (\mathrm{r}_{\mathsf{TS}_0})$.

(7) Hypatia is thereby *justified in believing* the principles of the theory $\mathsf{TS}_0 + (\mathrm{r}_{\mathsf{TS}_0})$.

We are interested in the conclusions (4) and (7). In particular, given what precedes those conclusions, we are interested in whether (4) and (7) follow. We approach this question by saying something about the epistemic terms appearing in JMB, which we have emphasized. These terms refer to the following kinds of epistemic attitude which Hypatia comes to adopt towards various corresponding theories: entitlements, acceptance, and justified belief.

The fundamental pattern that persists in JMB is this: from *justified belief* in one theory S, Hypatia *accepts* (via *entitlement*) an extension of S. Then, from *acceptance* of the extension of S, she comes to hold a *justified belief* in the extension of S. The process then repeats. Two things are worth pointing out about this pattern: first, that entitlements of cognitive project are entitlements to *accept*, and second, that acceptance is the sort of thing that must be capable of underwriting *justified belief*. So let us say something about what is meant by Hypatia's acceptance, in stages (3) and (6) above. If Hypatia's entitlements, as they appear in JMB, are bona fide examples of entitlements of cognitive project, then we can turn to (Wright, 2004) to learn more about what is meant by Hypatia's acceptance. In (Wright, 2004), acceptance is introduced

> as a more general attitude than belief, including belief as a subcase, which comes apart from belief in cases where one is warranted in acting on the assumption that P or taking it for granted that P or trusting that P for reasons that do not bear on the likely truth of P. (Wright, 2004, p. 177)

Acceptance of a proposition P understood in this general way is consistent with both agnosticism about P, and even with pessimism about the truth of P. But by the time we meet entitlements of cognitive project, this understanding of acceptance has shifted slightly. In particular, when P is an entitlement of some cognitive project, then if acceptance of P "is to be capable of underwriting rational belief in the things to which execution of the project leads, it has to be an attitude which *excludes doubt*" (Wright, 2004, p. 193). Thus:

> If there is entitlement of cognitive project, it has to be an entitlement not merely to act on the assumption that suitable presuppositions hold good, but to place trust in their doing so. (Wright, 2004, p. 193)

Why *trust*? Recall one of our salient remarks from the end of section 3.2.3: for entitlements of cognitive project to succeed, they have to be non-evidential – they cannot speak to the likely truth of the propositions they are supposed to warrant. But (Wright thinks that) it is precisely in the nature of trust that it gets by with little or no evidence (Wright, 2004, p. 194).[15] Thus entitlements of cognitive project are – *have to be* – entitlements to trust. So to say that we *accept* those propositions we are entitled to in the context of some cognitive project is just to say that we (rationally) *trust* those propositions. Transferring this understanding of acceptance to JMB, Hypatia's entitlements must also be entitlements to trust. So, at this point, we have fixed an understanding of all the epistemic terms appearing in JMB. Acceptance is the output of an entitlement, which amounts to a rational placing of trust. Since entitlements are non-evidential, on the basis of acceptance we can form *some* sort of belief.

But now we turn to our second observation about the epistemic pattern in JMB: that *acceptance* is the sort of thing that must be capable of underwriting *justified belief.* If entitlements as they appear in JMB are bona fide examples of entitlements of cognitive project, then we argue that JMB runs

[15] We note that this seems like a peculiar understanding of the term "trust." *Faith* might be more appropriate. Nonetheless, we'll continue to use "trust," since that is what is used by Wright.

into some serious trouble here. Here is the problem, in a nutshell: acceptance is understood as a rational entitlement to trust, capable of underwriting *some* sort of belief formation. But trust gets by with little or no evidence. Thus, acceptance is the sort of thing which has *little or no* evidence as input. But in JMB, acceptance is supposed to underwrite *justified* belief. We take it that *justified* belief is exactly the sort of belief which requires evidence as input. If acceptance is understood as the sort of thing which has little or no evidence as input, *how can it be said to underwrite the sort of epistemic attitude for which evidence is required*?

The problem, we argue, is particularly pertinent for stage (7) of JMB, where Hypatia arrives at justified belief in the principles of the theory $\mathsf{TS}_0 + (\mathrm{r}_{\mathsf{TS}_0})$. So, let us focus for the moment on stage (7). We will return to stage (4) shortly. Here we recall the upshot of one of our important junctures from section 4.1.2: that the source of Hypatia's warrant in general for the principles of $\mathsf{TS}_0 + (\mathrm{r}_{\mathsf{TS}_0})$, which are not derivable in TS_0 alone, lies with $(\mathrm{r}_{\mathsf{TS}_0})$. Suppose Hypatia has arrived at stage (6) of her process, whereby she has come to accept the theory $\mathsf{TS}_0 + (\mathrm{r}_{\mathsf{TS}_0})$. Consider her attitude towards the theory $\mathsf{TS}_0 + (\mathrm{r}_{\mathsf{TS}_0})$, strictly stronger than the theory EA which she starts out with (and strictly stronger than the theory TS_0, in which she currently holds a justified belief). Let us try and characterize the principles of $\mathsf{TS}_0 + (\mathrm{r}_{\mathsf{TS}_0})$ according to the kind of epistemic attitude held by Hypatia. On one hand, the principles of $\mathsf{TS}_0 + (\mathrm{r}_{\mathsf{TS}_0})$ which Hypatia holds a *justified* belief in by stage (6) of her process, is every member of the theory TS_0, apart from the sequents (T1) and (T2). On the other hand, the principles of $\mathsf{TS}_0 + (\mathrm{r}_{\mathsf{TS}_0})$ which Hypatia is merely *entitled* to believe by stage (6) of her process, are (T1), (T2), and (r_{TS_0}).

This leaves us to consider every consequence of $\mathsf{TS}_0 + (r_{\mathsf{TS}_0})$ not derivable in TS_0 alone. For argument's sake, let us consider only sentences in the language of EA derivable in $\mathsf{TS}_0 + (r_{\mathsf{TS}_0})$ but not TS_0 alone. (E.g., Con(EA).) What is Hypatia's attitude towards this class of principles of $\mathsf{TS}_0 + (r_{\mathsf{TS}_0})$? We have argued that the source of Hypatia's warrant for the principles of $\mathsf{TS}_0 + (r_{\mathsf{TS}_0})$, which are not derivable in TS_0 alone, lies with (r_{TS_0}). We conclude that in light of this, Hypatia's warrant for sentences in the language of EA, derivable in $\mathsf{TS}_0 + (r_{\mathsf{TS}_0})$ but not TS_0 alone, is also entitlement. In other words, she is merely *entitled to believe* sentences in the language of EA derivable in $\mathsf{TS}_0 + (r_{\mathsf{TS}_0})$, but not derivable in TS_0 alone. The point is that she is *not justified in believing* sentences in the language of EA derivable in $\mathsf{TS}_0 + (r_{\mathsf{TS}_0})$, but not TS_0 alone. She *cannot* be, if the kind of warrant Hypatia has for sentences in the language of EA derivable in $\mathsf{TS}_0 + (r_{\mathsf{TS}_0})$, but not TS_0 alone, *is not an evidential kind of warrant*. Thus, we claim that even if Hypatia's warrants for adopting the rules (T1), (T2), and (r_{TS_0}), are entitlements, it does not (*cannot*) follow that she comes to hold a justified belief in the principles of $\mathsf{TS}_0 + (r_{\mathsf{TS}_0})$.

This line of reasoning is closely related to what Wright calls the *leaching problem*:

> The general picture is that the cornerstones which sceptical doubt assails are to be held in place as things one may warrantedly trust without evidence. Thus at the foundation of all our cognitive procedures lie things we merely implicitly trust and take for granted, even though their being entitlements ensures that it is not irra-

> tional to do so. But in that case, what prevents this 'merely taken for granted' character from leaching upwards from the foundations, as it were like rising damp, to contaminate the products of genuine cognitive investigation? (Wright, 2004, p. 207)

Our claim, essentially, is that if Hypatia's warrant for the rule ($\mathrm{r}_{\mathsf{TS}_0}$) is mere entitlement, then her entitlement leaches, in Wright's sense, and is inherited by the consequences of $\mathsf{TS}_0 + (\mathrm{r}_{\mathsf{TS}_0})$ which cannot be derived in TS_0 alone. Thus, Hypatia does not come to hold newly justified mathematical beliefs in her move from (6) to (7). Rather, she only comes to hold newly entitled mathematical beliefs.

Now, in fact, Wright argues that leaching does *not* uniformly supplant the kind of ordinary evidential justification we take ourselves to have for our everyday beliefs. Our claim therefore requires some substantiation, for if leaching does not occur in Hypatia's context in the way that we have claimed it to occur, perhaps it can be said that she comes to hold a justified belief in the principles of the theory $\mathsf{TS}_0 + (\mathrm{r}_{\mathsf{TS}_0})$ after all. To make sense of why Wright thinks leaching is not a significant problem, consider again the Cartesian scenario from section 3.2.1:

(i′) My current experience is in all respects as if there is a hand in front of me.

(ii′) There is really a hand in front of me.

(iii′) I am not now in the midst of a lucid and persistent dream.

Here, the proposition "there is really a hand in front of me" in (ii′) is what we hold a justified belief in (by inference on the basis of (i′)). But the proposition "I am not now in the midst of a lucid and persistent dream" in (iii′) is what we are merely *entitled* to believe. To say that one's entitlement to (iii′) leaches, is just to say that one's justified belief in (ii′) is *replaced* by a mere entitlement to (ii′). This is precisely what Wright thinks does not happen. Rather, his claim is that entitlements *help recover* one's claim to hold a justified belief in (ii′). The point of entitlements is not to uniformly replace our justified beliefs, but to allow us to hold those justified beliefs by separating out cornerstones as propositions which are epistemically propped up, so to speak, in other ways.

One may well choose to press the veracity of Wright's claim, but that is of little concern to us here. Rather, (i′)–(iii′) serve to highlight a point of departure of entitlements of cognitive project as they occur in JMB, from the notion of entitlements of cognitive projects in Wright, 2004. Analogously, let us focus on the case of (r_{TS_0}), and ask: if Hypatia is entitled to adopt this rule, then what are the corresponding justified beliefs, which her entitlement recovers? Presumably, these include any theorem of TS_0. This aligns with Hypatia's I-III line of reasoning we discussed in section **??**:

(a) I have written down a proof of $\Gamma \Rightarrow \Delta$ from the axioms of TS_0.

(b) $\Gamma \Rightarrow \Delta$.

(c) (r_{TS_0}) is valid.

Hypatia comes to hold a justified belief in any such consequence $\Gamma \Rightarrow \Delta$ on the basis of her proof. But this requires, in turn, collateral warrant for the validity of ($\mathrm{r}_{\mathsf{TS}_0}$). If she is entitled to adopt that rule, then analogously to the story in Wright, 2004, it is her justified belief in any theorem of TS_0 which is recovered by her entitlement.

Here we recall another upshot of our earlier important juncture in section 4.1.2: the difference between (c) above, and the cornerstone proposition (iii$'$) above, is that the rule ($\mathrm{r}_{\mathsf{TS}_0}$), when added to Hypatia's current realm of means TS_0 for providing justification, yields a host of new mathematical consequences. But the cornerstone proposition "I am not now in the midst of a lucid and persistent dream," considered in conjunction with one's current realm of means for providing justification in the Cartesian scenario, which consists of our best empirical theories, yields *no* new empirical consequences.

With this in mind, consider Hypatia's epistemic attitude towards sentences in the language of EA derivable in $\mathsf{TS}_0 + (\mathrm{r}_{\mathsf{TS}_0})$, but not TS_0 alone. Observe that there are *no* analogs to these sentences in the story in the Cartesian scenario. Indeed, Hypatia's epistemic attitude toward these sentences does not even appear in her story until stage (6), where she comes to accept the theory $\mathsf{TS}_0 + (\mathrm{r}_{\mathsf{TS}_0})$. But even there, her attitude is only mere acceptance – it is certainly not the case that she holds a justified belief in these sentences by stage (6) of JMB. Thus, there is no sense in which Hypatia's entitlement to $\mathsf{TS}_0 + (\mathrm{r}_{\mathsf{TS}_0})$ *recovers* her justified belief in sentences in the language of EA derivable in $\mathsf{TS}_0 + (\mathrm{r}_{\mathsf{TS}_0})$, but not TS_0 alone. As a result, there is no reason to think that Hypatia's warrant for these sentences is anything other than the warrant she has for ($\mathrm{r}_{\mathsf{TS}_0}$), by virtue of which these sentences enter her story. That is, there is no reason to think that her warrant for these sentences is anything other than *entitled* belief. Overall, we think this the validity of Hypatia's move from (6)–(7) in JMB in serious doubt. Insofar as *arithmetical* theorems beyond those derivable in EA would count as a real success toward Hypatia's goal of justifying new *mathematical* beliefs, a general lack of justified belief in the principles of the theory $\mathsf{TS}_0+(\mathrm{r}_{\mathsf{TS}_0})$ severely damages the prospects for success of her cognitive project. Even at this early stage, the only new arithmetical (hence, mathematical) theorems she encounters are not such that she is able to acquire a justified belief in them.

We have learned the following from the preceding discussion: if Hypatia is to come to hold a *justified* belief in the theory $\mathsf{TS}_0 + (\mathrm{r}_{\mathsf{TS}_0})$, then the kind of warrant she has to have for adopting the rule $(\mathrm{r}_{\mathsf{TS}_0})$ must be capable of underwriting exactly that kind of belief – *justified* belief.

Here we recall the key claim which emerged from our discussion in section 4.1.2: that Hypatia must do extra justificatory work if she is to arrive at newly justified mathematical beliefs by coming to hold an independently justified belief in the principles of ZF, but she need not do any extra justificatory work if she arrives at newly justified mathematical beliefs via the JMB process.

Given what we said above, we think that this claim is false. For it follows from what we have said that Hypatia must absolutely carry out justificatory work in her JMB process, if she is to come to hold newly justified mathematical beliefs. In general, if all she has is entitled belief in reflection principles, like $(\mathrm{r}_{\mathsf{TS}_0})$, then *every* new arithmetical theorem she encounters in her JMB process inherits that entitled belief. So, when Hypatia carries out her JMB process, all she achieves is newly entitled mathematical beliefs. Put simply: to *justify* those beliefs, she must do justificatory work.

Recall also the idea from which this key claim emerged: Hypatia is perfectly able, in principle, to come to hold an independently (i.e. other than via a JMB-like process) justified belief in the principles of ZF. At this point, we have argued that the claim is false. This, we suggest, puts Hypatia's JMB process on an epistemic par with her alternative route to newly justified mathematical beliefs, by coming to independently hold a justified belief in the principles of ZF. Either way, to achieve newly justified mathematical beliefs, Hypatia has to do justificatory work. As we move forward into chapter 5, we will therefore take seriously, and make use of, Hypatia's route to an independently justified belief in the principles of ZF, and bring it to bear on her current cognitive project.

For now, let us round off our discussion of the validity of Hypatia's JMB process by considering stage (4), at which she arrives at a justified belief in the principles of the theory TS_0. Of course, our observations above also affirm our earlier suspicion that proposition asserting the validity of the sequents (T1) and (T2) is not a presupposition of Hypatia's cognitive project. Nonetheless, we are interested in the validity of the conclusion in stage (4), and so let us suppose for the moment, purely for argument's sake, that the validity of the sequents (T1) and (T2) is indeed an entitlement of Hypatia's cognitive project.

The first thing to note is that the theory TS_0 is conservative over EA (this was not the case for the theory $\mathsf{TS}_0 + (\mathrm{r}_{\mathsf{TS}_0})$ with respect to TS_0). That is, the only new consequences of TS_0 which are not derivable in EA alone are those in the expanded language, $\mathcal{L}_T$. We argue that a similar line of reasoning to the one we made above shows that Hypatia can only be merely entitled to believe the $\mathcal{L}_T$-consequences of TS_0. Hypatia's epistemic attitude toward the $\mathcal{L}_T$-consequences of TS_0 does not appear in her story until stage (3), where she comes to accept the theory TS_0. But even then, Hypatia's attitude is only mere acceptance – it is certainly not the case that she holds a justified belief in these sentences by stage (3) of JMB. So there is no sense in which Hypatia's entitlement to TS_0 can be said to recover her justified belief in the $\mathcal{L}_T$-consequences of TS_0. As a result, there is no reason to think that her warrant for these sentences is anything other than the warrant she has for the sequents (T1) and (T2), by virtue of which these sentences enter her picture. That is, her entitlement leaches, and there is no reason to think that Hypatia's warrant for $\mathcal{L}_T$-consequences of TS_0 is also anything other than *entitled* belief.

This, we think, is enough to undermine the conclusion in stage (4), that Hypatia arrives at a justified belief in the principles of the theory TS_0. We note, however, that in one sense, this is a less serious problem for Hypatia than the conclusion in stage (7) failing to hold. Insofar as *arithmetical* theorems beyond those derivable in EA would count substantially towards her goal of justifying new *mathematical* beliefs, presumably a general lack of justified belief in the principles of the theory TS_0 does not matter *too* much to Hypatia at this stage, because TS_0 yields no new *arithmetical* consequences when compared to EA itself. So, perhaps a general lack of justified belief in the principles of the theory TS_0 does not severely impair Hypatia's prospects, as we argued a general lack of justified belief in the principles of the theory $\mathsf{TS}_0 + (\mathrm{r}_{\mathsf{TS}_0})$ does. Nonetheless, we maintain that the conclusion in stage (4) is false.

So at this point, JMB looks to be in some trouble. There are a few problems. Hypatia's warrant for the principles (T1), (T2), and ($\mathrm{r}_{\mathsf{TS}_0}$), is not entitlement of her current cognitive project c, for we have argued that the propositions asserting the validity of these principles fail to meet Wright's clause 3. But her warrant for these principles is not ordinary mathematical justification, understood in the sense of derivability, either: Hypatia extends EA by the sequents (T1) and (T2), but EA can't derive (or interpret) those principles. So EA can't function as a means for justifying those sequents. Similarly, she extends TS_0 by the rule ($\mathrm{r}_{\mathsf{TS}_0}$), but TS_0 can't derive (or interpret) ($\mathrm{r}_{\mathsf{TS}_0}$). So TS_0 can't function as a means for justifying ($\mathrm{r}_{\mathsf{TS}_0}$). So what *is* Hypatia's warrant for these principles? And whatever the nature of this warrant, how can it underwrite justified belief in such a way that JMB succeeds after all? It is time to try and answer these questions using everything we have learned. While EA can't justify the sequents (T1) and (T2), and TS_0 can't justify the rule ($\mathrm{r}_{\mathsf{TS}_0}$), ZF can, and we have argued that ZF is back on the table for Hypatia. So let us see if we can fit these ideas together in a suitable way, and explore a way out of the problems we now face.

Chapter 5

A new kind of warrant

We propose a new kind of warrant that can stand in for Hypatia's warrant for the validity of the rules (T1), (T2), and ($\mathrm{r}_{\mathsf{TS}_0}$), in such a way that (we claim) justified belief is preserved in JMB. In section 5.1 we develop and put forward our proposal by modifying Wright's definition of entitlement of cognitive project in a few steps. In section 5.2 we argue that our new kind of warrant can underwrite justified belief.

We call our new kind of warrant *induced entitlement of cognitive project*. Before we define this new notion, let us front-load this chapter with the intuitive ideas behind induced entitlements of cognitive project, with a look ahead to chapter 6. Recall that *entitlements* of cognitive project exist as relations between propositions and cognitive projects. Aligning with this idea, first we formulate *induced* entitlements of cognitive project as relations between specific kinds of cognitive projects (encompassing Hypatia's cognitive project JMB), and specific kinds of propositions. The kinds of cognitive projects we define capture the idea that we start out with a justified belief in a theory of arithmetic S, and that our cognitive goal is to arrive in a justified belief in a suitable extension of S. The kinds of propositions we define assert the validity of various principles (P), which correspond to the principles we extend theories by, during a JMB process (disquotational truth sequents, and reflection principles).

Retaining the idea behind Wright's clause 1, these propositions are presuppositions of the specific kinds of cognitive project we define. But our idea is to modify Wright's clause 3. In particular, purely from the *internal* point view of such a cognitive project, any attempt to justify these propositions looks infinitely regressive in Wright's sense. So purely from the internal point of view of one's cognitive project, induced entitlements of cognitive project look like entitlements of cognitive project. However, from an *external* point of view of such a cognitive project, induced entitlements of cognitive project do not look like entitlements of cognitive project, because they can be justified in non-regressive ways. So, if a proposition p is an induced entitlement of such a cognitive project c, where p asserts the validity of some principle (P), then the success of c rationally requires us to not doubt p (i.e., the validity of principle (P)). And in attempting to justify induced entitlements of cognitive project, we rely at most on further propositions asserting the validity of principles of the same kind, but these further propositions are not presuppositions of c. Thus, the success of our cognitive project does *not* rationally require us to not doubt these further propositions.

Finally, we argue that induced entitlements can underwrite the formation of (mathematically) justified beliefs. Suppose p is an induced entitlement of cognitive project c. Then p is a proposition asserting the validity of some principle (P). We define the notion of being *warranted* in extending a theory S by the principle (P), *by induced entitlement of cognitive project* c. This definition will be satisfied whenever the proposition p is an induced entitlement of cognitive project c, and *if* we held a justified belief in the theory S + (P), then we would succeed in achieving the cognitive goal of c. Thus, we make a subtle distinction between the warrant we have for the propositions asserting the validity of a principle (P), and the warrant we have for the principle (P) itself.

The key idea is that all of this warrant is "outsourced" from the cognitive project c. This is where our in principle ability to come to *independently* hold a justified belief in certain theories is brought to bear on our solution to the problems we have encountered. Aligning with everything we have said so far, we will suppose that we may come to independently hold a justified belief in the theory ZF. The theory ZF derives every principle (P) we meet below. So the warrant we have for the principle (P) itself is just ordinary mathematical justification, understood in the sense of derivability. The source of this ordinary mathematical justification for (P) is not the theory S itself, which we extend by (P). But when we are warranted in extending S by (P) by induced entitlement, *some* ordinary justificatory source for (P) is witnessed. We claim that this makes the warrant we have for extending S by (P) fundamentally (mathematically) justificatory.[1] Thus, beliefs we form on the basis of this

[1] In particular, we do not need to formulate any new notion of mathematical justification.

kind of warrant are (mathematically) justified beliefs.

This outsourcing of justification is an essential ingredient in our solution. The basic problem we set out to solve was this: if we hold a justified belief in a theory S, and on this basis we are warranted in some sense in extending S by a principle (P) not derivable in S, then what is the nature of that warrant, such that we thereby come to hold a justified belief in S? We have argued that mere entitlement cannot be the answer, and in general, justification cannot come for free. To justify beliefs, we must do justificatory work. The idea behind induced entitlements is simply to outsource that justificatory work, by framing the situation in terms of specific kinds of cognitive project.

We use the ideas above to articulate several aspects of the *force* underlying induced entitlements. The idea is that if p is an induced entitlement of cognitive project c, then we *should* believe p because otherwise we cannot succeed in undertaking c, and furthermore, we can justify p in a non-regressive manner relative to c. In particular, we have every reason to think that we will not err on the basis of inferences made using the principle (P). Finally, if we are warranted by extending a theory S by (P) by induced entitlement of cognitive project c, then we *should* believe (P) itself because there is an ordinary mathematical justificatory source for (P). When we finally return to the first-orderist's scenario in chapter 6, we will fit all of these ideas together. The various mathematical theories of implicit commitments we proposed in chapter 2 correspond to the goals of certain kinds of cognitive projects. The warrant the first-orderist has for the principles comprising these theories of implicit commitments is induced entitlement of the corresponding cognitive project. Thus, the force underlying induced entitlements is what underlies the *commitment* of implicit commitments.

5.1 Induced entitlements of cognitive project

Our first task is to formulate induced entitlements of cognitive project as relations between specific kinds of cognitive projects, and specific kinds of propositions. Recall the idea: purely from an *internal* point of view of one's cognitive project, induced entitlements of cognitive project look like entitlements of cognitive project. However, from an *external* point of view of such a cognitive project, induced entitlements of cognitive project do not look like entitlements of cognitive project, because they can be justified in non-regressive ways. So, our first task is to formulate the notion of an attempt to justify a proposition being internal/external to c as precisely as we can (of course, there are epistemic notions involved). Here is our strategy: first, we make a few preliminary technical remarks. Second, we offer a general notion of a JMB-like cognitive project. Third, we make precise the idea of the scope of such a project. Fourth, we say how we succeed in achieving the goal of such a cognitive project. Finally, using all of these ideas, we formulate our notion of an attempt to justify a proposition being internal/external to c, and propose our idea of induced entitlements of cognitive project.

Preliminaries: for technical reasons, we introduce from (Fischer et al., 2021) a stronger version of the reflection rules we have been considering so far. We have been considering forms of the following reflection principle for sequents of a theory S:

$$\frac{\Rightarrow \mathrm{Pr}_{\mathsf{S}}(\ulcorner\Gamma \Rightarrow \Delta\urcorner)}{\Gamma \Rightarrow \Delta}\ (\mathrm{r}_{\mathsf{S}})$$

To account for sequent chains of reasoning featuring embedded implications, a second reflection rule is introduced in (Fischer et al., 2021). The second rule employs a two-place provability predicate $\mathrm{Pr}^2_{\mathsf{S}}(\ulcorner\Gamma \Rightarrow \Delta\urcorner, \ulcorner\Theta \Rightarrow \Lambda\urcorner)$, representing the fact that it is admissible in S to infer $\Theta \Rightarrow \Lambda$ from $\Gamma \Rightarrow \Delta$. The *reflection principle for provably admissible rules in* S is the following rule:

$$\frac{\Rightarrow \mathrm{Pr}^2_{\mathsf{S}}(\ulcorner\Gamma \Rightarrow \Delta\urcorner, \ulcorner\Theta \Rightarrow \Lambda\urcorner) \qquad \Gamma \Rightarrow \Delta}{\Theta \Rightarrow \Lambda}\ (\mathrm{R}_{\mathsf{S}})$$

In the context of any reasonable theory S, the theory $\mathsf{S} + (\mathrm{r}_{\mathsf{S}})$ is a subtheory of $\mathsf{S} + (\mathrm{R}_{\mathsf{S}})$. If S is an axiomatizable theory, we define the *reflection on* S as the closure of S under the reflection rule $(\mathrm{R}_{\mathsf{S}})$ by:

$$\mathrm{R}(\mathsf{S}) = \mathsf{S} + (\mathrm{R}_{\mathsf{S}}).$$

We may then define iterations of reflection on S as before. For example, $\mathrm{R}(\mathrm{R}(\mathsf{S}))$ is the result of closing $\mathrm{R}(\mathsf{S})$ under the rule $(\mathrm{R}_{\mathrm{R}(\mathsf{S})})$. We denote $\mathrm{R}(\mathrm{R}(\mathsf{S}))$ by $\mathrm{R}^2(\mathsf{S})$. We note that the difference between $(\mathrm{r}_{\mathsf{S}})$ and $(\mathrm{R}_{\mathsf{S}})$ is a technical one, rather than a conceptual one, so that in what follows, we claim to have not shifted the goalposts in any significant way. The value in formulating our new notions for the rule $(\mathrm{R}_{\mathsf{S}})$ is that we may easily illustrate all our new concepts below using results in (Fischer et al., 2017), which feature this rule.

Next, let c be a cognitive project and S a suitable arithmetical theory. We want to capture the idea that S is the "starting point" of a JMB-like process, and our goal is to arrive at a justified belief in the theory $\mathrm{R}(\mathsf{TS}_0)$. (Recall: the theory TS_0 is the result of extending S by the disquotational truth sequents (T1) and (T2) for S. The theory $\mathrm{R}(\mathsf{TS}_0)$ is the result of closing S under the reflection rule $(\mathrm{R}_{\mathsf{TS}_0})$ for TS_0.)

So, let us call c a *JMB-project for* S if the undertaking of c is a JMB-like process, where one begins with a justified belief in the axioms of S, and the cognitive goal of which is to arrive at a justified belief in the theory $\mathrm{R}(\mathsf{TS}_0)$. That is, we call c a JMB-project for S if it has the following structure:

(1) We start out with a justified belief in the principles of the theory S.

($\vdots$) $\vdots$

(n) We are thereby justified in believing the principles of the theory $\mathsf{TS}_0 + (\mathrm{R}_{\mathsf{TS}_0}) = \mathrm{R}(\mathsf{TS}_0)$.

Next we want to capture the idea of iterating the process above. Let c be a JMB-project for S. For $\beta \in \mathrm{Ord}$, we call c an *iterated JMB-project for* S *with limit* β just in case it has the following structure:

(1) We start out with a justified belief in the principles of the theory S.

($\vdots$) $\vdots$

(n) We are thereby justified in believing the principles of the theory $\mathsf{TS}_0 + (\mathrm{R}_{\mathsf{TS}_0}) = \mathrm{R}(\mathsf{TS}_0)$.

($\vdots$) $\vdots$

(α) We are thereby justified in believing the principles of the theory $\mathrm{R}^{\beta}(\mathsf{TS}_0)$, the result of recursively closing the theory $\mathrm{R}(\mathsf{TS}_0)$ under reflection β times.

For example, during an iterated JMB-project for S with limit 1, we begin with a justified belief in the theory S, and ultimately we thereby arrive at a justified belief in the theory $\mathrm{R}^2(\mathsf{TS}_0)$.

Next, we would like to keep track of the theories we are aiming to arrive at a justified belief in, by the end of each stage of iteration of c.

Definition 17. Let c be an iterated JMB-project for S with limit β. Recursively define a sequence $\langle \mathsf{T}_\alpha : \alpha \leq \beta \rangle$ as follows:

$$\mathsf{T}_0 = \mathsf{TS}_0 + (\mathrm{R}_{\mathsf{TS}_0}) = \mathrm{R}(\mathsf{TS}_0)$$

$$\mathsf{T}_\xi = \mathrm{R}(\mathsf{T}_\alpha) \text{ if } \xi = \alpha + 1$$

$$\mathsf{T}_\xi = \bigcup_{\alpha < \xi} \mathsf{T}_\alpha \text{ if } \xi \text{ is a limit.}$$

We call the sequence

$$\mathcal{S}_c = \langle \mathsf{T}_\alpha : \alpha \leq \beta \rangle$$

defined in this manner the *scope* of c.

We also want to account for theories which do not belong to the scope of c itself, but rather are “absorbed” as one carries out an iterated JMB-project.

Definition 18. Let T be a theory and c be an iterated JMB-project for S

with limit β. We say that T *is embedded in the scope of* c just in case there exists a theory T_α in the scope of c such that $\mathsf{T}_\alpha \vdash \mathsf{T}$.

Thus, the scope of the following iterated JMB-project for S with limit 0 is $\{\mathrm{R}(\mathsf{TS}_0)\}$:

(1) We start out with a justified belief in the principles of the theory S.

($\vdots$) $\vdots$

(n) We are thereby justified in believing the principles of the theory $\mathsf{TS}_0 + (\mathrm{R}_{\mathsf{TS}_0}) = \mathrm{R}(\mathsf{TS}_0)$.

The theory TS_0 is embedded in the scope of this cognitive project.

As we will see below, there are other examples of natural theories T properly embedded in the scope of Hypatia's cognitive project, which we will reconstruct shortly (i.e. T is embedded in the scope of Hypatia's cognitive project, but does not itself belong to the scope of her cognitive project).

Finally, we offer a very general notion of how we succeed in achieving the kind of cognitive goal we are interested in. Let S be a suitable arithmetic theory. We say that a theory T *witnesses success over* S just in case $\mathsf{T} \supsetneq \mathsf{S}$ and one holds a justified belief in T. Thus, if c is an iterated JMB-project for S with limit β, and $\mathcal{S}_c = \langle \mathsf{T}_\alpha : \alpha \leq \beta \rangle$ is the scope of c, then if justified belief is preserved throughout the corresponding JMB process, every theory T embedded in $\mathcal{S}_c$ such that $\mathsf{T} \supsetneq \mathsf{S}$ witnesses success over S. In such cases (i.e., where an iterated JMB-project c for S with limit β is fixed, and one holds a justified belief in T, for some T embedded in $\mathcal{S}_c$ with $\mathsf{T} \supsetneq \mathsf{S}$) we say that T

witnesses the success of c. But we also allow for other ways of succeeding in justifying new mathematical beliefs, independently of a fixed cognitive project c.

Let us pause and get an example on the table to illustrate all this new terminology. In fact, we will give a particular reconstruction of Hypatia's cognitive project. Structurally, we may write her cognitive project (denote it by c as usual) in the following way:

(1) We start out with a justified belief in the principles of the theory EA.

($\vdots$) $\vdots$

(n_1) We are thereby justified in believing the principles of the theory $\mathsf{TS}_0 = \mathsf{EA} + (\mathrm{T1}) + (\mathrm{T2})$.

($\vdots$) $\vdots$

(n_2) We are thereby justified in believing the principles of the theory $\mathsf{TS}_0 + (\mathrm{R}_{\mathsf{TS}_0}) = \mathrm{R}(\mathsf{TS}_0)$.

($\vdots$) $\vdots$

(ω) We are thereby justified in believing the principles of the theory $\mathrm{R}^\omega(\mathsf{TS}_0)$.

In particular, c is an iterated JMB-project with limit ω. The scope $\mathcal{S}_c$ of c is $\langle \mathrm{R}^n(\mathsf{TS}_0) : n \leq \omega \rangle$. The strongest theory belonging to $\mathcal{S}_c$ is the theory $\mathrm{R}^\omega(\mathsf{TS}_0)$. Moreover, the theory $\mathrm{R}^\omega(\mathsf{TS}_0)$ can define ramified truth predicates indexed by all ordinals $\omega^{\omega \times n}$ for all natural numbers n.[2] Thus, if justified

[2] See (Fischer, Nicolai, & Horsten, 2017, Corollary 3).

belief is preserved throughout the process above, we capture the idea that Hypatia succeeds in her undertaking, for the theory $\mathrm{R}^{\omega}(\mathsf{TS}_0)$ witnesses the success of c. (Again, preservation of justified belief depends on what stands in for the warrant in this process. What we've argued so far is that if entitlement of cognitive project stands in for the warrant in this process, then Hypatia's justified belief is not preserved.)

Clearly TS_0 is embedded in the scope $\mathcal{S}_c$ of c: it is a subtheory of $\mathrm{R}(\mathsf{TS}_0)$. But there is also another natural example of a theory T such that: (1) there exists a theory T^* in the scope of Hypatia's cognitive project such that $\mathsf{T}^* \vdash \mathsf{T}$, and (2) T does not belong to the scope of c itself. This theory is PKF, an internal axiomatization of Kripke's theory of truth (Halbach & Horsten, 2006). Let $\mathcal{S}_c = \langle \mathsf{T}_\alpha : \alpha \leq \omega \rangle$ be the scope of c. Fischer et al. (2017) show the following:

$$\mathsf{TS}_0 \subsetneq_1 \mathrm{R}(\mathsf{TS}_0) \subsetneq_2 \mathsf{PKF} \subsetneq_3 \mathrm{R}^2(\mathsf{TS}_0).$$

$\subsetneq_1$ is trivial. $\subsetneq_2$ follows from Proposition 2 and Lemma 4 of (Fischer et al., 2017). $\subsetneq_3$ follows from Corollary 1 and Proposition 3 of (Fischer et al., 2017). The conceptual point is that PKF is embedded in the scope of c, as witnessed by the theory $\mathrm{R}^2(\mathsf{TS}_0) \in \mathcal{S}_c$, but PKF does not belong to the scope of c itself.[3] Rather, PKF is absorbed between Hypatia's first and second iterations of the process above.

With an example on the table, next, we want to start filling in the $\vdots$. Of

[3] PKF is used in (Fischer et al., 2017) as a means of comparison for theories of iterated reflection using proof-theoretic ordinal analysis. For example, PKF proves arithmetical transfinite induction up to the ordinal $\varphi_\omega 0$ (Halbach & Horsten, 2006).

course, *one* way to structure a JMB-project for S is to extend S in just the way Hypatia does, during her cognitive project. So let us give the principles a name, by which she structures her project. Here, we only consider the disquotational truth sequents for S, and reflection rules. But we might hope to generalize the following definition to include other principles, and this is just what we do in chapter 6, when we return to the first-orderist's scenario.

Let T be any suitable theory. We say that a proposition p is an *extension validator for* T just in case p asserts either:

(a) that the sequents $\varphi \Rightarrow T(\ulcorner \varphi \urcorner)$ and $T(\ulcorner \varphi \urcorner) \Rightarrow \varphi$ are valid, for some $\mathcal{L}_T$-formula φ, or

(b) that the rule (R_{T}) is valid.

Now we are ready to formulate the idea of an attempt to justify a certain kind of proposition being internal/external to an iterated JMB-project. Essentially, we will modify Wrights clause 3: any attempt to justify p would involve further presuppositions of c in turn of no more secure a prior understanding... and so on without limit; so that if we accepted that there is nevertheless an onus to justify p, we would implicitly undertake a commitment to an infinite regress of justificatory projects, each concerned to vindicate the presuppositions of its predecessors. Wright's clause 3 is *negative.* It characterizes principles p such that one *cannot* provide evidence for p in a non-regressive manner. We propose to modify clause 3 by turning things around, and drawing on some of our observations from chapter 4.

Let us recall those observations. Earlier, we argued that if ZF is on the table for Hypatia, then the proposition asserting the validity of the rule ($\mathrm{R}_{\mathsf{TS}_0}$) fails to meet Wright's clause 3. She is perfectly able to justify the validity of this rule without relying on a further presupposition of her cognitive project. Recall the idea: Hypatia *proves* that $\mathcal{M} \models \varphi$ in order to conclude that $\mathcal{M} \models \varphi$. Thus, in attempting to justify the validity of the rule ($\mathrm{R}_{\mathsf{TS}_0}$), she relies on the validity of *some* kind of inference rule (R_{S}), of a piece with the rule ($\mathrm{R}_{\mathsf{TS}_0}$) whose validity she set out to justify in the first place. But S must be stronger than TS_0 itself, and this opens the door: if S in fact lies beyond the scope of Hypatia's cognitive project c, we argued that there is no reason to think (R_{S}) is a presupposition of c. Since Hypatia is perfectly able to independently come to hold a justified belief in the principles of ZF, she can justify the validity of the rule ($\mathrm{R}_{\mathsf{TS}_0}$) from the external (to c) point of view of ZF, and only has to rely on the validity of (R_{ZF}).

So now what we want to do is to formulate the idea of Hypatia's attempted justification of p as "involving some version of the same principle," in such a way that the version of the principle she relies on in her attempted justification of p is formulated for a theory which lies outside the scope of c. This was the purpose of our defining the notion of extension validator propositions.

Let T be any suitable theory and let p_{T} be an extension validator for T. We say that an attempt to justify p_{T} is *schematic* just in case the attempt to justify p_{T} involves relying on an extension validator p_{S} for some theory $\mathsf{S} \supseteq \mathsf{T}$. We denote by $\mathcal{J}_{\mathsf{T}}$ the class of theories $\mathsf{S} \supseteq \mathsf{T}$ such that any attempt to justify p_{T} involves relying on an extension validator p_{S} for S. We call the members of the class $\mathcal{J}_{\mathsf{T}}$ the *bases* of the corresponding attempt to justify p_{T} which is schematic.

Thus, Hypatia's attempt to justify the validity of the sequent $T(\ulcorner\varphi\urcorner) \models \varphi$ is not schematic. We argued earlier that she is perfectly able to justify $\varphi \models T(\ulcorner\varphi\urcorner)$ and $T(\ulcorner\varphi\urcorner) \models \varphi$, and in making those justificatory attempts, she need not rely on any further extension validators, for any other theories.[4] On the other hand, Hypatia's attempts to justify the validity of the rule $(\mathrm{R}_{\mathsf{TS}_0})$, which we described earlier, are schematic. We outlined a few justificatory attempts, in which she relied on the extension validators for a range of theories S (those she meets during her cognitive project, and ZF). Those extension validators were the propositions asserting the validity of the corresponding rule $(\mathrm{R}_{\mathsf{S}})$.

Of course, the notion of a justificatory attempt being schematic is vague: it involves the notion of an attempt to justify a proposition. We won't attempt to make this notion any more precise than we have here, but we have in mind any of our previous justificatory attempts. The notion of an attempt to justify p_{T} being schematic is intended to encompass the idea that our understanding of the corresponding principle talked about by the proposition p_{S}, upon which we rely in attempting to make such a justificatory attempt, is of no more secure a prior understanding than the principle talked about by the proposition p_{T} itself. Thus, if one were to sincerely accept an onus to justify p_{T}, one would thereby implicitly undertake a commitment to justify p_{S}. However, it need not be the case that p_{S} is a presupposition of c.

Definition 19. Let:

- c be an iterated JMB-project for S with limit β,

[4] For $\mathcal{M} \models \varphi$ just *means* the same thing as $\mathcal{M} \models T(\ulcorner\varphi\urcorner)$.

- T be any theory, and
- p_{T} be an extension validator for T.

We say that an attempt to justify p_{T} is *external to c* just in case:

- if the attempt to justify p_{T} is schematic, then there is a basis $\mathsf{B} \in \mathscr{J}_{\mathsf{T}}$ of the justificatory attempt such that: (1) we hold a justified belief in B, and (2) B is not embedded in the scope of c.

We say that an attempt to justify p_{T} is *internal to c* just in case:

- the attempt to justify p_{T} is schematic, and every basis $\mathsf{B} \in \mathscr{J}_{\mathsf{T}}$ of the justificatory attempt in which we hold a justified belief is also embedded in the scope of c.

Notice that if an attempt to justify p_{T} is *not* schematic, then that attempt to justify p_{T} is external to c. Our intention is to capture our earlier attempts to justify the validity of the disquotational truth sequents, during which we relied on no extension validators for any other theories. The idea is that we can make these justificatory attempts "from any perspective." Furthermore, if an attempt to justify p_{T} is external to c, then there is a basis B of that justificatory attempt such that we hold a justified belief in B, and B is not embedded in the scope of c. Thus, if p_{T} asserts the validity of principle (P), then a source of justification for the validity of the principle (P) is witnessed. In this way, our justification for the validity of the principle (P) is outsourced from c.

The notion of an attempt to justify p_{T}'s being external to c differs from Wright's notion of justificatory regress. Recall that Wright's notion of justificatory regress is negative: it characterizes propositions p such that one *cannot* justify p in a non-regressive manner. But the notion of p_{T}'s being external to c characterizes propositions such that one *can* justify p in a non-regressive manner *relative to c*. This is where we have turned things around. On the other hand, the notion of an attempt to justify p_{T} being internal to c is where things start to line up with clause 3 of Wright's definition. In particular, if an attempt to justify p_{T} is internal to c, then it is schematic, and every basis S of that justificatory attempt in which we hold a justified belief is embedded in the scope of c. Thus, the corresponding extension validators p_{S} on which we rely in attempting to give such a justification are presuppositions of c. Since the principles talked about by p_{S} are of no more secure a prior understanding than the principles talked about by p_{T} itself, and if we were to sincerely accept an onus to justify p_{T}, we would thereby implicitly undertake a commitment to justify each corresponding p_{S}, we have met the conditions of infinite justificatory regress in Wright's sense for that justificatory attempt. If this scenario plays out for *any* attempt to justify p_{T}, we have thereby met the conditions of Wright's clause 3.

The example we gave earlier, whereby Hypatia boosted her initial theory with reflection, in order to try and justify the validity of the reflection rule ($\mathrm{R}_{\mathsf{TS}_0}$), is an attempt to justify an extension validator for TS_0 which is internal to c. We said that this justificatory strategy never exceeds the limits of Hypatia's current cognitive project c, since transfinitely iterating the reflection operation is (in part) how her JMB process is structured. But the notion of an attempt to justify p_{T}'s being external to c does not require this. It requires only that Hypatia *can* step far enough beyond the scope of her cognitive project to avoid justificatory regress relative to the project itself.

At this point, we have everything on the table. So let us propose the following kind of warrant for propositions of the sort we are interested in:

Definition 20. Let c be an iterated JMB-project for S with limit β. Let T be any theory and let p_{T} be an extension validator for T. We say that p_{T} is an *induced entitlement of cognitive project c* just in case the following conditions hold:

(i) p_{T} is a presupposition of c.

(ii) There exists an attempt to justify p_{T} which is external to c.

Definition 20 is narrow. It is formulated only for principles of the sort that we have been interested in so far. We do not think this is necessarily a drawback of our notion of induced entitlements.[5] For one of the morals of our discussion so far is that the nature of the cognitive project one is engaged in is essential for making sense of what our warrant for principles of the form p_{T} consists in. Definition 20 focuses on cognitive projects of the sort we have been interested in so far. Shortly, we will generalize definition 20, so that it encompasses cognitive projects of other sorts.

Notice also that we do not need any counterpart to Wright's clause 2, which required us to have no sufficient reason for believing a proposition to be false.

[5] We could in principle expand our list of extension principles to include propositions asserting the validity of other kinds of sequent or rule.

For if p_{T} is an induced entitlement of cognitive project c, then there exists an attempt to justify p_{T} which is external to c. The point is that if this is the case, we need not assure ourselves that we *lack* (mathematical) evidence *against* p_{T}: in attempting to justify p_{T}, we assure ourselves that in principle, we *have* (mathematical) evidence *for* p_{T}.

Finally, we would like to ensure that our notion of induced entitlement of cognitive project (which is a relation between propositions and cognitive projects) aligns with the principles themselves which extension validators assert the validity of.

Definition 21. Let c be an iterated JMB-project for S with limit β. Let T be any theory embedded in the scope of c and let p_{T} be an extension validator for T which asserts the validity of some principle (P). We say that we are *warranted in extending* T *by* (P) *by induced entitlement of cognitive project* c just in case:

(i) p_{T} is an induced entitlement of cognitive project c.

(ii) If we held a justified belief in T + (P), then T + (P) would witness the success of c.

Here is the idea. Let c be an iterated JMB-project for S with limit β. Let T be any theory embedded in the scope of c and let p_{T} be an extension validator for T which asserts the validity of some principle (P). If p_{T} is an induced entitlement of cognitive project c, this means that when we try to justify p_{T}, we rely at most on extension validators p_{S} for some basis $\mathsf{S} \in \mathcal{J}_{\mathsf{T}}$ of our justificatory attempt. But if we do end up relying on extension validators in this way, then there is a particular basis $\mathsf{B} \in \mathcal{J}_{\mathsf{T}}$ of the justificatory attempt such that: (1) we hold a justified belief in B, and (2) B is not embedded in the scope of c. By looking back at our project from the point of view of B, we are assured that (P) is a valid principle. Furthermore, extending T by (P) would put us in a place where we are able to achieve the goal of c. So, from an internal perspective, why would we *not* be warranted in extending T by (P)?

This is the first stage at which we can say something about the force underlying induced entitlements. We outline the basic idea here, and return to it at the end of chapter 6. Suppose we are warranted in extending a theory T by some principle (P) by induced entitlement of an iterated JMB-project c. This means two things: (i) the proposition asserting the validity of (P) is an induced entitlement of c, and (ii) if we held a justified belief in $\mathsf{S} + $(P), then $\mathsf{S} + $(P) would witness the success of c. If we hold a justified belief in T, why *should* we also believe (P)?

At this point, we can put our finger on three such reasons. The first two reasons derive from condition (i) above, and concern the validity of (P). Here is the idea. In undertaking c, we are hoping to achieve a justified belief in the principles of the resulting theory $\mathsf{S} + $(P). So, we want to be sure that the principle (P), whose validity we *have to* rely on in undertaking c, and via which we hope to arrive at this target justified belief, will not result in belief in falsities. In particular: if we doubted the validity of (P), we cannot rationally maintain that our project is still significant or competent. Thus, the validity of (P) is necessary for achieving our cognitive goal. This is one reason why we should believe (P). But furthermore, induced entitlements assure us that we will not err on the basis of inferences made using the principle (P), because induced entitlements witness a justification of the validity of (P). Thus, not only should we believe that (P) is valid because undertaking c would be impossible otherwise, but we should also believe (P) because we have every reason to believe that (P) *is* in fact valid.

The third reason derives from condition (ii) above: if we held a justified belief in $\mathsf{S}+$(P), then $\mathsf{S}+$(P) would witness the success of c. So, in particular, we should believe (P) because doing so puts us in a position where we are able to achieve our cognitive goal. If we believe (P), then we believe $\mathsf{S}+$ (P). So all that is left to do is say why that belief is justified. We make this argument in section 5.2 below. For now, let us tie together the notions we have defined in chapter 5, and answer one of the questions we were left with at the end of chapter 4: what is Hypatia's warrant for the principles (T1), (T2), and ($\mathrm{R}_{\mathsf{TS}_0}$)?

Let c be the following iterated JMB-project for EA with limit 0, essentially the first iteration of Hypatia's cognitive project:

(1) We start out with a justified belief in the principles of the theory EA.

($\vdots$) $\vdots$

(n_1) We are thereby justified in believing the principles of the theory $\mathsf{TS}_0 = \mathsf{EA} + (\mathrm{T1}) + (\mathrm{T2})$.

($\vdots$) $\vdots$

(n_2) We are thereby justified in believing the principles of the theory $\mathsf{TS}_0 + (\mathrm{R}_{\mathsf{TS}_0}) = \mathrm{R}(\mathsf{TS}_0)$.

Clearly EA is embedded in the scope of c. The proposition p_{EA} asserting the validity of the disquotational truth sequents (T1) and (T2) for EA is an extension validator for EA. Furthermore, p_{EA} is an induced entitlement of the cognitive project c: we argued earlier that p_{EA} is a presupposition of c, and we also exhibited an attempt to justify p_{EA} which was external to c, for that justificatory attempt was not schematic. Then we are warranted by induced entitlement in extending EA by (T1) and (T2).

The theory TS_0 is also embedded in the scope of c. The proposition p_{TS_0} asserting the validity of the rule $(\mathrm{R}_{\mathsf{TS}_0})$ is an extension validator for TS_0. Furthermore, p_{TS_0} is an induced entitlement of the cognitive project c: we argued earlier that p_{TS_0} is a presupposition of c, and we also exhibited an attempt to justify p_{TS_0}, in which we relied on an extension validator for the theory ZF. Thus, *if we hold a justified belief in* ZF, then we are warranted by induced entitlement of c in extending TS_0 by $(\mathrm{R}_{\mathsf{TS}_0})$.

Now consider Hypatia's more ambitious cognitive project. Let c be the following iterated JMB-project for EA with limit ω as above:

(1) We start out with a justified belief in the principles of the theory EA.

($\vdots$) $\vdots$

(n_1) We are thereby justified in believing the principles of the theory $\mathsf{TS}_0^{\mathsf{EA}} = \mathsf{EA} + (\mathrm{T1}) + (\mathrm{T2})$.

($\vdots$) $\vdots$

(n_2) We are thereby justified in believing the principles of the theory $\mathsf{TS}_0^{\mathsf{EA}} + (\mathrm{R}_{\mathsf{TS}_0^{\mathsf{EA}}}) = \mathrm{R}(\mathsf{TS}_0^{\mathsf{EA}})$.

($\vdots$) $\vdots$

(ω) We are thereby justified in believing the principles of the theory $\mathrm{R}^{\omega}(\mathsf{TS}_0^{\mathsf{EA}})$.

Recall that the scope of c is the sequence $\mathcal{S}_c = \langle \mathrm{R}^n(\mathsf{TS}_0) : n \leq \omega \rangle$. All the remarks from the two preceding paragraphs hold for this cognitive project c. But we also have the following. Let $\omega \geq n \geq 2$ be arbitrary. The theory $\mathrm{R}^n(\mathsf{TS}_0)$ is embedded in the scope of c. Let T be the theory $\mathrm{R}^{n-1}(\mathsf{TS}_0)$. Then the proposition p_{T} asserting the validity of the rule $(\mathrm{R}_{\mathrm{R}^{n-1}(\mathsf{TS}_0)})$ is an extension validator for the theory T. Furthermore, p_{T} is an induced entitlement of the cognitive project c: we argued earlier that p_{T} is a presupposition of c, and we also argued that we can attempt to justify p_{T} by relying on an extension validator for the theory ZF. Thus, *if we hold a justified belief in* ZF, then we are warranted by induced entitlement of c in extending $\mathrm{R}^{n-1}(\mathsf{TS}_0)$ by $(\mathrm{R}_{\mathrm{R}^{n-1}(\mathsf{TS}_0)})$.

So, at this point, we have made sense of the following particular reconstruction of Hypatia's cognitive project:

(1) We start out with a justified belief in the principles of the theory EA.

(2) We are warranted in extending EA by the sequents (T1) and (T2) for EA by induced entitlement of cognitive project.

(3) We are thereby justified in believing the principles of the theory $\mathsf{TS}_0 = \mathsf{EA} + (\mathrm{T1}) + (\mathrm{T2})$.

(4) We are warranted in extending TS_0 by the rule $(\mathrm{R}_{\mathsf{TS}_0})$ by induced entitlement of cognitive project.

(5) We are thereby justified in believing the principles of the theory $\mathsf{TS}_0 + (\mathrm{R}_{\mathsf{TS}_0}) = \mathrm{R}(\mathsf{TS}_0)$ by induced entitlement of cognitive project.

(6) We are warranted in extending $\mathrm{R}(\mathsf{TS}_0)$ by the rule $(\mathrm{R}_{\mathrm{R}(\mathsf{TS}_0)})$ by induced entitlement of cognitive project.

(7) We are thereby justified in believing the principles of the theory $\mathrm{R}^2(\mathsf{TS}_0)$.

$(\vdots)$ $\vdots$

(ω) We are thereby justified in believing the principles of the theory $\mathrm{R}^\omega(\mathsf{TS}_0)$.

This is an iterated JMB-project for EA with limit ω, and we have formulated a particular notion of warrant which can fill in the $\vdots$ from before. So, in particular, we have addressed one of the issues we were left with at the end of chapter 4. There, we said that Hypatia's warrant for the principles (T1), (T2), and $(\mathrm{R}_{\mathsf{TS}_0})$, is not entitlement of her current cognitive project c. But her warrant for these principles also cannot simply be ordinary mathematical justification (understood as derivability). We asked: what *is* Hypatia's warrant for these principles? We have formulated the following answer: if Hypatia can come to independently hold a justified belief in the principles of ZF (and if she can, then there is no reason why she *shouldn't*, for on the face of things there is just as much justificatory work involved in this as there is in undertaking her JMB process), then the warrant she has for these principles is induced entitlement of the cognitive project c. In fact, we think this puts Hypatia in an economical position: the *only* justificatory work she has to do is to arrive at a justified belief in ZF. Induced entitlements make full use of that justificatory work, and don't require her to do anything else.

But what we do not yet know is whether this reconstruction of Hypatia's cognitive project is *successful*. So the final thing we need to do is provide an answer to the other question we were left with at the end of chapter 4: is the nature of this warrant such that it can underwrite justified belief? If induced entitlements of cognitive project can underwrite justified belief, then we will have shown that Hypatia can succeed in carrying out her JMB process after all. Next, we argue that induced entitlements of cognitive project can do exactly this.

5.2 Underwriting justified belief

We argue that induced entitlements are fundamentally (mathematically) justificatory (unlike entitlements of cognitive project per Wright). For this reason, they can underwrite (mathematical) justified belief. Thus, induced entitlements of cognitive project can stand in for the warrant in Hypatia's iterated JMB-project with limit ω above, in such a way that she does thereby succeed in arriving at justified beliefs in the principles of Predicative Analysis.

Here is the basic idea. Let T be a theory and Φ some (set of) principle(s) formulated in the language of T. For such Φ, let us say that T is a *source of justification for* Φ just in case: (1) we hold a justified belief in the axioms of T, and (2) T derives Φ. Now let us write $\mathsf{T} \mapsto \mathsf{T} + \Phi$ to mean "we are warranted by $\mapsto$ in extending T by Φ." Call T the *source of the warrant* $\mapsto$. So far, we have adopted the following picture: if we hold a justified belief in T, and T derives Φ, then it makes sense to think that we are also justified in believing

Φ. Thus, it makes sense to think of $\mapsto$ as "justification." This, essentially, is how we argued that one's justified belief in the axioms of a theory propagate to the theory itself. The key observation is that the source of justification for Φ *is the same as* the source of the warrant "justification."

Now let S be a theory and Ψ some (set of) principle(s) formulated in a language extending that of S such that: (1) we hold a justified belief in S, and (2) S does not derive Ψ. Suppose $\mathsf{S} \mapsto \mathsf{S} + \Psi$. Then it no longer makes sense to think of the warrant $\mapsto$ as "justification," because S does not derive Ψ. But if there is *some* theory S^* in which we hold a justified belief which *does* derive Ψ, then there *is* a source of justification for Ψ. And we formulated induced entitlements of cognitive project in such a way that this ordinary justificatory source is witnessed. For in all the examples we have seen, the basis B of our attempt to justify the validity of the various principles we have considered, is also such that B derives those principles outright. So while the source of justification for Ψ *is not the same as* the source of the warrant "induced entitlement of cognitive project," induced entitlements of cognitive project are fundamentally (mathematically) justificatory simply because they witness *some* source of (mathematical) justification for Ψ. Because they are fundamentally (mathematically) justificatory, any belief we form on the basis of an induced entitlement is also (mathematically) justified. Thus, induced entitlements of cognitive project can underwrite the formation of (mathematically) justified beliefs. Moreover, the kind of justified belief they underwrite is ordinary mathematical justified belief, understood in the sense of derivability.

Now let us spell out the argument by way of an example. Consider the first iteration of Hypatia's cognitive project, the following iterated JMB-project c with limit 0 above:

(1) We start out with a justified belief in the principles of the theory EA.

(2) We are warranted in extending EA by the sequents (T1) and (T2) by induced entitlement of cognitive project.

(3) We are thereby justified in believing the principles of the theory $\mathsf{TS}_0 = \mathsf{EA} + (\mathrm{T1}) + (\mathrm{T2})$.

(4) We are warranted in extending TS_0 by the rule $(\mathrm{R}_{\mathsf{TS}_0})$ by induced entitlement of cognitive project.

(5) We are thereby justified in believing the principles of the theory $\mathsf{TS}_0 + (\mathrm{R}_{\mathsf{TS}_0}) = \mathrm{R}(\mathsf{TS}_0)$ by induced entitlement of cognitive project.

Let us suppose that we have independently come to hold a justified belief in ZF, and first let us focus on the move from (1) to (3). We focus on sequents of the form (T1). No instance of (T1) is derivable in EA, so it does not make sense to think of "warranted" in (2) as "justified." But we have argued that the proposition asserting the validity of any instance of (T1) is an induced entitlement of cognitive project c. In particular, we can justify the validity of any instance of (T1) in a way that meets the conditions for being external to c. So all we need now is an ordinary source of mathematical justification for (T1). But we hold a justified belief in ZF, so we have one. That is, $\mathsf{ZF} \vdash \mathsf{EA} + (\mathrm{T1})$, so ZF itself is an ordinary source of mathematical justification for (T1).[6] So

[6] In fact, our ordinary source of mathematical justification need not be ZF itself. It could

while the source of justification ZF for (T1) is not the same as the source EA of our induced entitlement for (T1), there *is* an ordinary source of mathematical justification for (T1). So (we claim) our induced entitlement to extend EA by (T1) is fundamentally mathematically justificatory. As a result, we are able to form a justified belief about any instance of (T1) on the basis of our induced entitlement. The intuition is this: we are already assured (T1) is a valid principle, and by looking back at our project from the external point of view of ZF, we can justify (T1) itself in the perfectly ordinary sense of derivability. So, from the internal point of view of c, why would we *not* think that our induced entitlement to extend EA by (T1) is fundamentally justificatory?

We reason similarly for the move from (3) to (5). The source of our justification for ($\mathrm{R}_{\mathsf{TS}_0}$) cannot be TS_0 itself, by the usual Gödelian considerations. So it does not make sense to think of "warranted" in (4) as "justified." But we have argued that the proposition asserting the validity of the rule ($\mathrm{R}_{\mathsf{TS}_0}$) is an induced entitlement of cognitive project c. In particular, we can justify the validity of ($\mathrm{R}_{\mathsf{TS}_0}$) using ZF as a basis for our justification. But furthermore, $\mathsf{ZF} \vdash (\mathrm{R}_{\mathsf{TS}_0})$, so ZF itself is a perfectly ordinary source of mathematical justification for ($\mathrm{R}_{\mathsf{TS}_0}$). Again, while the source of justification ZF for ($\mathrm{R}_{\mathsf{TS}_0}$) is not the same as the source TS_0 of our induced entitlement for ($\mathrm{R}_{\mathsf{TS}_0}$), there *is* a source of justification for ($\mathrm{R}_{\mathsf{TS}_0}$). So, our induced entitlement to extend TS_0 by ($\mathrm{R}_{\mathsf{TS}_0}$) is fundamentally justificatory. As a result, we are able to form a justified belief about ($\mathrm{R}_{\mathsf{TS}_0}$) on the basis of our induced entitlement.

Here are some pictures to illustrate the above. Let

be any theory which derives (T1) and (T2). But our choice of ZF aligns things nicely with reflection below.

$$\longrightarrow$$

stand for any justificatory kind of warrant, so we read

$$\longrightarrow \mathsf{S}$$

as “we hold a justified belief in S.”[7] Let $\mapsto$ stand for induced entitlement. For stages (1)–(3) of the cognitive project above, we have argued that if the following arrows exist:

$$\begin{tikzcd}
 & {} \arrow[d] \\
 & \mathsf{ZF} \arrow[d, bend right] \arrow[d, bend left] \\
\longrightarrow \mathsf{EA} & \mathsf{EA} + (\mathrm{T1}) + (\mathrm{T2})
\end{tikzcd}$$

Figure 5.1: Justification diagram 1

then the following induced arrow:

$$\mathsf{EA} \;\mapsto\; \mathsf{EA} + (\mathrm{T1}) + (\mathrm{T2})$$

Figure 5.2: Induced entitlement diagram 1

which we have already argued exists, is also justificatory. The key idea is that a source of justification for the sequents (T1) and (T2) is witnessed. In particular, a source of ordinary mathematical justification for the sequents

[7] This is somewhat imprecise, since S might be a theory or a principle. But this will illustrate the idea.

(T1) and (T2), understood in the sense of derivability, is witnessed. In this way, our justification for the sequents (T1) and (T2) itself is outsourced from c.

Continuing: since we have argued that the arrows of Figure 5.1 *do* exist, now we hold a justified belief in $\mathsf{EA} + (\mathrm{T1}) + (\mathrm{T2}) = \mathsf{TS}_0$. Moving on to stages (3)–(5), now the following justificatory arrows exist:

$$\begin{tikzcd}
& & {} \arrow[d] \\
& & \mathsf{ZF} \arrow[d] \\
{} \arrow[r] & \mathsf{TS}_0 & \mathsf{TS}_0 + (\mathrm{R}_{\mathsf{TS}_0})
\end{tikzcd}$$

Figure 5.3: Justification diagram 2

Hence the following induced arrow:

$$\mathsf{TS}_0 \quad \mapsto \quad \mathsf{TS}_0 + (\mathrm{R}_{\mathsf{TS}_0})$$

Figure 5.4: Induced entitlement diagram 2

which we have already argued exists, is also justificatory. Again, the idea is that a source of ordinary mathematical justification for the rule $(\mathrm{R}_{\mathsf{TS}_0})$ is witnessed. In this way, our justification for the rule $(\mathrm{R}_{\mathsf{TS}_0})$ is outsourced from c.

This is the second stage at which we can say something about the force underlying induced entitlements. Suppose in general that we are warranted in extending a theory T by some principle (P) by induced entitlement of an iterated JMB-project c. If we hold a justified belief in T, why *should* we also believe (P)? Above, we gave two reasons why we should believe that (P) is *valid*: if we doubted the validity of (P), we could not rationally maintain that our project is still significant or competent, and we are assured that we will not err on the basis of inferences made using the principle (P), because induced entitlements witness a justification of the validity of (P). Furthermore, we said that we should believe (P) because doing so puts us in a position to achieve our cognitive goal. At this point we can give one final reason why we should believe (P): if there is a theory in which we hold a justified belief which derives the principle (P) outright, then we have an ordinary source of mathematical justification for (P) itself. So, we should believe (P) for this reason too.

In fact, in the context of Hypatia's iterated JMB-project for EA with limit ω, we have seen that our source of justification for the *validity* of the rule $(\mathrm{R}_{\mathsf{TS}_0})$ aligns with our ordinary source of mathematical justification $(\mathrm{R}_{\mathsf{TS}_0})$ itself. We may justify the validity of each of $(\mathrm{R}_{\mathsf{TS}_0})$ using ZF as a basis, and ZF derives $(\mathrm{R}_{\mathsf{TS}_0})$ outright.[8] So *all* of the justificatory work we have to do, to justify our belief in $(\mathrm{R}_{\mathsf{TS}_0})$, lies in independently justifying ZF. Not only have we outsourced all of the requisite justificatory work, but we have outsourced as little justificatory work as possible.

Finally, we note that extensions by reflection were the root of a real problem, if all the warrant we have for the reflection principle is entitlement of cognitive project in Wright's sense. For we argued that if $(\mathrm{R}_{\mathsf{TS}_0})$ is warranted

[8] This is also the case for the iterated reflection rules.

merely by entitlement of cognitive project, there is no independent *justificatory* source for the resulting new $\mathcal{L}_{\mathsf{EA}}$-consequences of the theory $\mathrm{R}(\mathsf{TS}_0)$. In this case, our entitlements leach, and we lose any hope of preserving justified belief. But now things are different. If $(\mathrm{R}_{\mathsf{TS}_0})$ is warranted by induced entitlement of cognitive project, then there is an independent (from c) justificatory source for the resulting new $\mathcal{L}_{\mathsf{EA}}$-consequences of the theory $\mathrm{R}(\mathsf{TS}_0)$. It is the same source of the justification we have for $(\mathrm{R}_{\mathsf{TS}_0})$ itself. If what we have said is correct, we thereby *do* arrive at a justified belief in all the new $\mathcal{L}_{\mathsf{EA}}$-consequences of the theory $\mathrm{R}(\mathsf{TS}_0)$. In this way, we arrive at newly justified *arithmetical* beliefs.

5.3 Concluding remarks

We formulated induced entitlements of cognitive project to answer the following two questions: what is Hypatia's warrant for the principles (T1), (T2), and $(\mathrm{R}_{\mathsf{TS}_0})$? Whatever this warrant is, how can it underwrite justified belief in such a way that her JMB process succeeds after all? Induced entitlements of cognitive project keep the scope of Hypatia's cognitive project in sharp focus. They are such that from an internal perspective, those propositions look like entitlements in Wright's sense. But we also allow for the possibility of justifying these propositions from an external perspective. Furthermore, induced entitlements witness a justificatory source for the principles which are warranted by induced entitlement. This witness, we argued, makes induced entitlements of cognitive project fundamentally justificatory. Thus, we can substitute “warranted by induced entitlement of cognitive project” for “warranted” in JMB processes, in such a way that they succeed.

Before we finally return to the first-orderist’s scenario, let us offer a speculative comparison between induced entitlements of cognitive project, and entitlements of cognitive project in Wright’s sense. In contrast to our notion of induced entitlements, and in light of our remarks above, we may reformulate a narrow version of Wright’s definition of entitlement of cognitive project as it applies to extension validator propositions p_{T} of iterated JMB-projects:

Definition 22. Let c be an iterated JMB-project for S with limit β. Let T be any theory and let p_{T} be an extension validator for T. We say that p_{T} is an *entitlement of cognitive project* c just in case the following conditions hold:

(i) p_{T} is a presupposition of c.

(ii) We have no sufficient reason to believe that p_{T} is false.

(ii) Any attempt to justify p_{T} is internal to c.

We suggest that entitlements of cognitive project in Wright’s sense are a limiting case of induced entitlements of cognitive project as we have formulated the notion, where limit is understood in an epistemic sense.

One of the ideas behind induced entitlements of cognitive project is that there is a justificatory witness (which lies beyond the scope of c) to the principles which extension validator propositions assert the validity of. What happens, though, when we consider extending the *strongest* theory T in which we currently hold a justified belief, by the corresponding truth sequents and reflection rules? In this case, there can be *no* external justificatory witness. If we don't hold a justified belief in *any* theory beyond T, there cannot be a theory which lies beyond the scope of c in which we hold a justified belief. In such a case, when T is the mathematical "ceiling" of our justified beliefs, the corresponding truth sequents and reflection rules cannot be induced entitlements of any such cognitive project. Does it make sense now to think that these principles are warranted by entitlement of cognitive project in Wright's sense?

Possibly. Let us focus on reflection. On the story we have put forward, for our justified beliefs to propagate from the axioms of T to the theory of T, we presuppose the validity of the corresponding reflection rule (R_{T}). In particular, the validity of (R_{T}) is a presupposition of c, and let us suppose further that we have no reason to believe that (R_{T}) is not valid. Then clauses 1 and 2 of definition 22 are met.

But it is not clear to us how one would show that the validity of (R_{T}) meets clause 3 of Wright's definition. For example: to attempt to justify (R_{EA}), we moved to a theory strong enough to prove the soundness of EA, and appealed to the corresponding reflection principle for that theory. But we leveraged the idea that we also held a justified belief in that theory. If we don't hold a justified belief in any theory stronger than T, how would we even *articulate* an attempt to justify (R_{T}), never mind show that any such attempt was internal to c? Perhaps the details would be inductive: we would say something like "I came to hold a justified belief in (R_{S}) for some weaker-than-T theory S. There is no fundamental difference between (R_{T}) and (R_{S}). So, inductively, I should believe (R_{T})." In this case, our tentative suggestion is that the validity of (R_{T}) then meets the conditions of definition 22.

Nonetheless, one thing should be clear: we would not succeed in achieving the goal of such a cognitive project c, if our goal was to arrive at newly *justified* mathematical beliefs. For we could not succeed in *justifying* beliefs in consequences of extending T by principles which we are merely *entitled* to believe.

We do not think that this is necessarily a problem, in and of itself. For example, returning to Hypatia's context, let us ignore the possibility of independently justifying ZF, so that (for example) EA *is* the ceiling of our mathematically justified beliefs, and suppose that all we have beyond EA are entitled beliefs. For instance, aligning with our remarks in chapter 4, for our justified belief in the axioms of EA to transmit to the general theory of EA, then we are entitled (in the context of some cognitive project) to believe that the reflection principle (R_{EA}) for EA is valid. If we are thereby warranted in extending EA by (R_{EA}) by entitlement, then we arrive in general at entitled belief in the theory $\mathsf{EA} + (\mathrm{R}_{\mathsf{EA}}) \equiv \mathsf{PA}$.

But so what? Perhaps it does not really matter what the general epistemic status of PA is. For how would we differentiate two individuals, one of whom claimed to hold a justified belief in the principles of PA, and one of whom claimed to hold an entitled belief in the principles of PA? Perhaps we could just ask these individuals what kind of belief they take themselves to hold in the principles of PA. But (assuming these individuals understood our question) how would we be able to tell that this difference was not merely terminological?

It seems difficult to ascertain that the individual who claims to hold an entitled belief in the principles of PA, would be worse off than the individual who claims to hold a justified belief in the principles of PA.

As a contrast case, consider the following scenario: you and I are going to start flipping an unbiased coin. Suppose that I start out with a justified belief that the coin will turn up heads almost every time, but you start out with an entitled belief that the coin will turn up heads almost every time. We start flipping the coin, and both of us try and predict the outcome of each coin flip. Presumably, a neutral observer would be able to tell which of us started out with a justified belief that the coin will turn up heads almost every time, and which of us started out with an entitled belief that the coin will turn up heads almost every time. For if the coin was unbiased, and we flipped it a large number of times, then overall I would predict heads fewer times than you would (you would predict heads almost every time, for you cannot but take it for granted that the coin will turn up heads almost every time). And so, if we started betting on the coin outcomes, you would end up worse off than I do. In this scenario, I would much rather justify my belief about the bias of the coin, than merely take it for granted that the coin would turn up heads almost every time.

But in what sense would one rather hold a justified belief in the principles of PA, than merely take the principles of PA for granted? This seems to boil down to an issue we raised in chapter 3. If one really would rather hold a justified belief in the principles of PA, than merely take the principles of PA for granted, it seems to be because one thinks that there is some underlying normative force to justificatory methods in mathematics. But if there is no such underlying normative force, then if all we have are entitled beliefs in mathematical contexts, we do not think that this is a *bad* thing.

We end this note by pointing out that we do not think anything we have said hangs on this issue. Whatever the force (if there is any) underlying justificatory methods in mathematics consists of, the point of our discussion over chapters 3–5 has been to make sense of a *justificatory* epistemic route from theories of *arithmetic* (like EA) to theories in the realm of Predicative Analysis, where justification is understood in the sense of independent methods of axiom justification, or derivability. And to make sense of such a justificatory epistemic route, we have argued that justification must be outsourced *somewhere*.

It is finally time to return to the question we set off to investigate at the beginning of chapter 3. There, we asked what kind of warrant $\mapsto$ could be, such that it would make sense of the following scenario:

$$\begin{aligned}
\text{justified belief in } \mathsf{PA} &\mapsto \text{justified belief in } \mathsf{PA}^{T}_{\mathrm{Ax}_{\mathsf{PA}}}\\
\text{justified belief in } \mathsf{PA} &\mapsto \text{justified belief in } (\mathsf{PA}^{U}_{\mathrm{Ax}_{\mathsf{PA}}})_{\omega}\\
\text{justified belief in } \mathsf{PA} &\not\mapsto \text{justified belief in } (\mathsf{PA}^{T}_{\mathrm{Ax}_{\mathsf{PA}}})_{\omega}
\end{aligned}$$

The hasty suggestion was that maybe $\mapsto$ could be entitlement of some cognitive project. We have not even spoken about a cognitive project for the first-orderist, but even if we were to formulate one at this stage, it is clear why entitlement of cognitive project cannot stand in for $\mapsto$. For entitlements cannot underwrite justified belief. However, over the course of our investigation, we have also formulated the notion of induced entitlements of cognitive project, which we argued can underwrite justified belief. So let us see if our new notion can stand in for $\mapsto$.

Chapter 6

The commitment of implicit commitments

Our goal in this last chapter is to tie together everything we have learned. Let S be a suitable arithmetical theory in which we hold a (mathematically) justified belief, and let $\mathsf{I}(\mathsf{S})$ be a theory of implicit commitments in justifiably believing S. By the end of this chapter, we will answer the following question: if we hold a belief in the principles of S, why *should* we also believe the principles of $\mathsf{I}(\mathsf{S})$?

To arrive at our answer, first we argue that induced entitlements of cognitive project can make sense of the first-orderist's scenario from chapter 3. This will answer the question we set off to investigate in chapter 3: what sort of warrant $\mapsto$ makes sense of the following situation?

$$\begin{aligned}
\text{justified belief in } \mathsf{PA} &\mapsto \text{ justified belief in } \mathsf{PA}^{T}_{\mathrm{Ax}_{\mathsf{PA}}} \\
\text{justified belief in } \mathsf{PA} &\mapsto \text{ justified belief in } (\mathsf{PA}^{U}_{\mathrm{Ax}_{\mathsf{PA}}})_{\omega} \\
\text{justified belief in } \mathsf{PA} &\not\mapsto \text{ justified belief in } (\mathsf{PA}^{T}_{\mathrm{Ax}_{\mathsf{PA}}})_{\omega}
\end{aligned}$$

Second, we will say what underlies the force of *should* above, by tying together the theories of implicit commitments we identified in chapter 2, with the goals of certain kinds of cognitive projects.

It should be clear that there are a few obstacles in our way. The first is to formulate a cognitive project for the first-orderist. This kind of cognitive project obviously cannot structurally resemble Hypatia's iterated JMB-project. One reason is that reflection principles are the very thing that the first-orderist wants to avoid. Another reason is that the notion of truth featuring in iterated JMB-projects is type-free, but our notion of truth from chapter 2 is typed. So any such cognitive project will not be iterable. So, we need to formulate a different kind of cognitive project. Second, and in turn, this will require us to generalize our notion of induced entitlements of cognitive project, because we defined induced entitlements of cognitive project as a relation which holds between propositions of a very specific sort, and cognitive projects of a very specific sort. If we are going to generalize our notion to apply to other sorts of cognitive projects, then we will also need to generalize the kind of propositions which induced entitlements are defined for. We address these obstacles in turn.

6.1 Stability-projects

We formulate a kind of cognitive project which respects the epistemic stability associated with first-orderism. We call this kind of cognitive project a stability-project. To motivate stability-projects, first we need to think about the principles by which PA is extended, in the first-orderist's scenario. Recall (from chapter 2) that we considered extensions of PA by various semantic and schematic principles. We formulated these extensions of PA as sequent calculi. In particular, we considered various extensions of PA by combinations of appropriate versions of the general sequents and rules below, formulated for a suitable theory T.[1]

Uniform disquotational truth sequents.

(UT1) $\varphi(\underline{x}) \Rightarrow T(\ulcorner \varphi(\underline{x}) \urcorner)$

(UT2) $T(\ulcorner \varphi(\underline{x}) \urcorner) \Rightarrow \varphi(\underline{x})$

where x is arbitrary and $\varphi(v)$ is any $\mathcal{L}_{\mathsf{T}}$-formula.

Induction rule.

$$\frac{\varphi(x) \Rightarrow \varphi(x+1)}{\varphi(\underline{0}) \Rightarrow \varphi(t)}\,(\mathrm{Ind}_{\mathcal{L}_{\mathsf{T}}})$$

where x is not free in the lower sequent, t is an arbitrary term, and $\varphi(v)$ is any formula in the language $\mathcal{L}_{\mathsf{T}}$ of T.

[1] Recall also that our notion of truth is typed in this setting, unlike the setting in (Fischer et al., 2021).

Axiom soundness rule.

$$\frac{\Gamma \Rightarrow \Delta, D(s)}{\Gamma \Rightarrow \Delta, T(s)}\ (\mathrm{D})$$

where $D(x)$ is a T-schema,[2] and the case of interest is where $D(x)$ is $\mathrm{Ax}_{\mathsf{T}}(x)$. We continue to write (D) for the axiom soundness rule.

In what follows, we will also consider the following sequents,[3] which correspond to the compositional clauses governing the behavior of our typed truth predicate.

Compositional truth sequents.

$(T =_1)$ $\mathrm{ct}(x), \mathrm{ct}(y), \mathrm{val}(x) = \mathrm{val}(y) \Rightarrow T(x \mathbin{\underset{\cdot}{=}} y)$.

$(T =_2)$ $\mathrm{ct}(x), \mathrm{ct}(y), \mathrm{val}(x) = \mathrm{val}(y), T(x \mathbin{\underset{\cdot}{=}} y) \Rightarrow \mathrm{val}(x) = \mathrm{val}(y)$.

$(T\wedge_1)$ $\mathrm{Sent}_{\mathcal{L}_{\mathsf{T}}}(x \mathbin{\underset{\cdot}{\wedge}} y), T(x) \wedge T(y) \Rightarrow T(x \mathbin{\underset{\cdot}{\wedge}} y)$.

$(T\wedge_2)$ $\mathrm{Sent}_{\mathcal{L}_{\mathsf{T}}}(x \mathbin{\underset{\cdot}{\wedge}} y), T(x \mathbin{\underset{\cdot}{\wedge}} y) \Rightarrow T(x) \wedge T(y)$.

$(T\neg_1)$ $\mathrm{Sent}_{\mathcal{L}_{\mathsf{T}}}(x), T(\underset{\cdot}{\neg} x) \Rightarrow \neg T(x)$.

$(T\neg_2)$ $\mathrm{Sent}_{\mathcal{L}_{\mathsf{T}}}(x), \neg T(x) \Rightarrow T(\underset{\cdot}{\neg} x)$.

$(T\forall_1)$ $\mathrm{Sent}_{\mathcal{L}_{\mathsf{T}}}(\underset{\cdot}{\forall} y x), \forall y T(x[y/v]) \Rightarrow T(\underset{\cdot}{\forall} y x)$.

$(T\forall_2)$ $\mathrm{Sent}_{\mathcal{L}_{\mathsf{T}}}(\underset{\cdot}{\forall} y x), T(\underset{\cdot}{\forall} y x) \Rightarrow \forall y T(x[y/v])$.

[2] Defined in section 2.4.

[3] Which we did not explicitly write down in chapter 2.

For a suitable theory T, let SC[4] be the collection of all principles of the following form: $(T =_1)$, $(T =_2)$, $(T\wedge_1)$, $(T\wedge_2)$, $(T\neg_1)$, $(T\neg_2)$, $(T\forall_1)$, $(T\forall_2)$, and (D). Let IC[5] be the collection of all principles of the following form (UT1), (UT2), $(\mathrm{Ind}_{\mathcal{L}_\mathsf{T}})$, and (D). We are interested in extensions of T by subsets of the collection:

$$\mathsf{SC} \cup \mathsf{IC}.$$

We want to capture the idea that the first-orderist starts out with a justified belief in the principles of PA, and thereby arrives at new *mathematical* beliefs (broadly understood), but not new *arithmetical* beliefs. This aligns with the idea of epistemic stability. We also want our semantic and schematic components to feature centrally in this idea, so we will make use of the collection $\mathsf{SC} \cup \mathsf{IC}$ we defined above.

We give a general definition of such a project. Let S be a suitable arithmetic theory. We say that a cognitive project c is a *stability-project for* S just in case it has the following structure:

(1) We start out with a justified belief in the principles of the theory S.

($\vdots$) $\vdots$

(n) We are thereby justified in believing the principles of $\mathsf{S} + \mathsf{E}$, where $\mathsf{E} \subseteq \mathsf{SC} \cup \mathsf{IC}$ and $\mathsf{S} + \mathsf{E}$ is conservative over S.

When we enumerated the stages of Hypatia's process, we collected up all the theories she reaches at the end of each iteration, into what we called

[4] For "semantic core."

[5] For "inductive core."

the scope of her iterated JMB-project. The purpose of defining the scope of iterated JMB-projects was to differentiate theories that are internal to those cognitive projects, and theories that are external to those cognitive projects. We want to do something similar for stability-projects, but it makes no sense to define the scope of stability-projects as we did before, because they are not iterative. However, things are rather simpler for us now: we want the scope of a stability-project to consist of all conservative extensions of S by a subset of principles in the collection $\mathsf{SC} \cup \mathsf{IC}$.

Let S be a suitable theory and c be a stability-project for S. Define:

$$\mathcal{E}_c = \{\mathsf{S}^* \supseteq \mathsf{S} : \mathsf{S}^* = \mathsf{S} + \mathsf{E} \text{ for some subset } \mathsf{E} \subseteq \mathsf{SC} \cup \mathsf{IC}\}.$$

For a stability-project c for S, the collection $\mathcal{E}_c$ is just all the possible extensions of S by our semantic and schematic components. Thus, there are $\mathsf{S}^* \in \mathcal{E}_c$ which are not conservative over S.

Definition 23. Let S be a suitable theory and c be a stability-project for S. We call the collection:

$$\mathcal{S}_c = \{\mathsf{S}^* \supseteq \mathsf{S} : \mathsf{S}^* \in \mathcal{E}_c \text{ and } \mathsf{S}^* \text{ is a conservative extension of } \mathsf{S}\}$$

the *scope* of c.

This is the idea which will ensure our success. By defining the notion of the scope of the right kind of cognitive project, we will be able to make sense of the first-orderist's scenario.

To align with our work in chapter 5, we account for theories which do not belong to the scope of c itself, but rather are "absorbed" as one carries out a stability-project.

Definition 24. Let T be a theory and c be a stability-project for S. We say that T *is embedded in the scope of* c just in case there exists a theory S^* in the scope of c such that $\mathsf{S}^* \vdash \mathsf{T}$.

Our notion of *success* is as before. Recall: let S be a suitable arithmetic theory. We say that a theory T *witnesses success* over S just in case $\mathsf{T} \supsetneq \mathsf{S}$ and one holds a justified belief in T. Thus, if c is a stability-project for S and $\mathcal{S}_c$ is the scope of c, then if justified belief is preserved throughout the corresponding epistemic process, every theory embedded in $\mathcal{S}_c$ which properly extends S witnesses success over S. In such cases we say that T *witnesses the success of c*. The point is that if T witnesses the success of a stability-project c, then T is conservative over S.[6]

Now we want to fill in the $\vdots$. We offer a generalized typed version of our list of extension validators from chapter 5. Let T be any suitable theory. We say that a proposition p is an *extension validator for* T just in case p asserts any of the following:[7]

[6] But as before, we also allow for other ways of succeeding in achieving newly justified beliefs, even in conservative extensions of S. For instance, consider the following example from (Fischer et al., 2021). Let S be an $\mathcal{L}_T$-theory consisting of the axioms of PA (where truth does not appear in instances of PA's induction schema), the fully compositional truth axioms, and the axiom $M \vee \exists\varphi\neg\mathrm{Ind}(T\varphi)$, where M asserts the consistency of ZFC plus some large cardinal axiom, and $\mathrm{Ind}(T\varphi)$ is the instance of the induction scheme for $T\varphi$ with φ a code of a $\mathcal{L}_A$-formula with one free variable. Then S is semantically conservative over PA, hence proof theoretically conservative over PA.

[7] The uniform disquotational sequents are typed counterparts encompassing their type-free sentential versions (T1) and (T2) from chapter 5.

(a) that the rule $(\mathrm{R}_{\mathsf{T}})$ is valid, or

(b) that for arbitrary x, the sequents $\varphi(\underline{x}) \Rightarrow T(\ulcorner\varphi(\underline{x})\urcorner)$ and $T(\ulcorner\varphi(\underline{x})\urcorner) \Rightarrow \varphi(\underline{x})$ are valid, for some formula $\varphi(v)$ in the language of T restricted to its arithmetical part, or

(c) that the sequents $(T =_1)$, $(T =_2)$, $(T\wedge_1)$, $(T\wedge_2)$, $(T\neg_1)$, $(T\neg_2)$, $(T\forall_1)$, and $(T\forall_2)$ are valid, or

(d) that the rule $(\mathrm{Ind}_{\mathcal{L}_{\mathsf{T}}})$ is valid.

We note that we did not formulate an extension validator proposition for the axiom soundness rule (D). The reason is that we will not need to consider extensions of PA by the rule (D). Rather, (D) occurs as a result of extending PA by either the uniform disquotational truth sequents, or the fully compositional truth sequents. We say more about this below.

We formulate the notions of an attempt to justify extension validators being internal/external to a stability-project as before.[8] Let T be any suitable theory and let p_{T} be an extension validator for T. We say that an attempt to justify p_{T} is *schematic* just in case the attempt to justify p_{T} involves relying on an extension validator p_{S} for some theory $\mathsf{S} \supseteq \mathsf{T}$. We denote by $\mathcal{J}_{\mathsf{T}}$ the class of theories $\mathsf{S} \supseteq \mathsf{T}$ such that any attempt to justify p_{T} involves relying on an extension validator p_{S} for S. We call the members of the class $\mathcal{J}_{\mathsf{T}}$ the *bases* of the corresponding attempt to justify p_{T} which is schematic.

Definition 25. Let:

[8] Really we should distinguish these definitions from those in the type-free setting, but we take it that the context is clear enough.

- c be a stability-project for an arithmetical theory S,
- T be any theory, and
- p_{T} be an extension validator for T.

We say that an attempt to justify p_{T} is *external to c* just in case:

- if the attempt to justify p_{T} is schematic, then there is a basis $\mathsf{B} \in \mathscr{J}_{\mathsf{T}}$ of the justificatory attempt such that: (1) we hold a justified belief in B, and (2) B is not embedded in the scope of c.

We say that an attempt to justify p_{T} is *internal to c* just in case:

- the attempt to justify p_{T} is schematic, and every basis $\mathsf{B} \in \mathscr{J}_{\mathsf{T}}$ of the justificatory attempt in which we hold a justified belief is also embedded in the scope of c.

If an attempt to justify p_{T} is not schematic, then that attempt to justify p_{T} is external to c. Our intention behind this is to capture our attempts to justify the validity of the uniform disquotational truth sequents, and the fully compositional truth sequents, where (as we will see) we rely on no extension validators for any other theories. We can make these justificatory attempts "from any perspective."

Next, we generalize our notion of induced entitlements of cognitive project to apply to stability-projects:

Definition 26. Let c be stability-project for S. Let T be any theory and let p_{T} be an extension validator for T. We say that p_{T} is an *induced entitlement of cognitive project c* just in case the following conditions hold:

(i) p_{T} is a presupposition of c.

(ii) There exists an attempt to justify p_{T} which is external to c.

Finally, we ensure that our notion of induced entitlement of stability-projects aligns with the principles themselves which extension validators assert the validity of.

Definition 27. Let c be a stability-project for an arithmetical theory S. Let T be any theory embedded in the scope of c and let p_{T} be an extension validator for T which asserts the validity of some principle (P). We say that we are *warranted in extending* T *by* (P) *by induced entitlement of cognitive project* c just in case:

(i) p_{T} is an induced entitlement of cognitive project c.

(ii) If we held a justified belief in T + (P), then T + (P) would witness the success of c.

Let us recapitulate the idea behind this definition. Let c be a stability-project for an arithmetical theory S. Let T be any arithmetical theory embedded in the scope of c and let p_{T} be an extension validator for T which asserts the validity of some principle (P). If p_{T} is an induced entitlement of cognitive project c, this means that when we try to justify p_{T}, we rely at most on extension validators p_{S} for some basis $\mathsf{S} \in \mathcal{J}_{\mathsf{T}}$ of our justificatory attempt. But if we do end up relying on extension validators in this way, then there is a particular basis $\mathsf{B} \in \mathcal{J}_{\mathsf{T}}$ of the justificatory attempt such that: (1) we hold a justified belief in B, and (2) B is not embedded in the scope of c. So, by looking back at our project from the point of view of B, we are assured that (P) is a valid principle. Furthermore, extending T by (P) would put us in a position to achieve the goal of c. So, from an internal perspective, why would we *not* be warranted in extending T by (P)?

So, now our notion of induced entitlement of cognitive project is defined for two kinds of cognitive project: iterated JMB-projects in the type-free setting, and stability-projects in the typed setting. We argued that induced entitlements of iterated JMB-projects solved Hypatia's problem earlier. Our suggestion now is that *if the first-orderist independently holds a justified belief in* ZF, then induced entitlements of stability-projects offer a solution to the problem at the end of chapter 2: on the basis of their justified belief in PA, how can the first-orderist claim to thereby hold a justified belief in the theories $(\mathsf{PA}^U_{\mathrm{Ax}_{\mathsf{PA}}})_\omega$ and $\mathsf{PA}^T_{\mathrm{Ax}_{\mathsf{PA}}}$ considered individually, but claim to not thereby hold a justified belief in both of these theories at the same time?

To show this, we have a little work to do. We will reconstruct two stability-projects c for PA. The first will correspond to the idea that the first-orderist, from a justified belief in PA, may coherently also claim to hold a justified belief in the principles of the theory $(\mathsf{PA}^U_{\mathrm{Ax}_{\mathsf{PA}}})_\omega$. The second will correspond to the idea that the first-orderist, from a justified belief in PA, may coherently also claim to hold a justified belief in the principles of the theory $\mathsf{PA}^T_{\mathrm{Ax}_{\mathsf{PA}}}$.

6.1.1 Justified belief in $(\mathsf{PA}^U_{\mathrm{Ax}_{\mathsf{PA}}})_\omega$

Consider the following stability-project c for PA, a reconstruction of the idea that the first-orderist, from a justified belief in PA, may coherently also claim to hold a justified belief in the principles of the theory $(\mathsf{PA}^U_{\mathrm{Ax}_{\mathsf{PA}}})_\omega$:

(1) We start out with a justified belief in the principles of the theory PA.

$(\vdots)$ $\vdots$

(n) We are thereby justified in believing the principles of $(\mathsf{PA}^U_{\mathrm{Ax}_{\mathsf{PA}}})_\omega$.

We argue that it is possible to fill in the $\vdots$ in this process in the following way:

(1) We start out with a justified belief in the principles of the theory PA.

(2) We are warranted in extending PA by the sequents (UT1) and (UT2) by induced entitlement of this cognitive project.

(3) We are thereby justified in believing the principles of PA^U.

(4) We are warranted in extending PA^U by the rule $(\mathrm{Ind}_{\mathcal{L}_{\mathsf{PA}^U}}) = (\mathrm{Ind}_T)$ by induced entitlement of this cognitive project.

(5) We are thereby justified in believing the principles of $(\mathsf{PA}^U)_\omega$.

(6) We are thereby justified in believing the principles of $(\mathsf{PA}^U_{\mathrm{Ax}_{\mathsf{PA}}})_\omega$.

Notice that the move from step (5) to (6) is legitimized by our Theorem 3 from chapter 2: we showed that the theory $(\mathsf{PA}^U)_\omega$ can interpret the rule (D).[9] Thus, if we hold a justified belief in the axioms of $(\mathsf{PA}^U)_\omega$, then relying only on the validity of the corresponding reflection rule $(\mathrm{R}_{(\mathsf{PA}^U)_\omega})$, the rule (D) inherits justified belief in the ordinary sense of derivability. This explains why

[9] In fact, PA^U alone can do this. Fully extended induction plays no essential role in interpreting (D) in the proof of Theorem 3. Neither does fully compositional truth play any essential role in interpreting (D) in Leigh's Theorem 1.

we did not formulate an extension validator proposition for the rule (D). To acquire a justified belief in axiom soundness, we need only require a warrant for extending our base theory by (at least) the uniform truth sequents (UT1) and (UT2).

So it remains to argue that we are warranted in:

- extending PA by the sequents (UT1) and (UT2), and
- extending PA^U by the rule $(\mathrm{Ind}_{\mathcal{L}_{\mathsf{PA}^U}}) = (\mathrm{Ind}_T)$,

by induced entitlement of c. We address each of these in turn.

Let p_{PA} be the extension validator asserting the validity of (UT1) (we remark on the argument for (UT2) afterwards). Suppose we hold a justified belief in ZF. We argue that p_{PA} is an induced entitlement of cognitive project c. To establish that p_{PA} is an induced entitlement of cognitive project c, we need to show two things: (i) that p_{PA} is a presupposition of c, and (ii) that there exists an attempt to justify p_{PA} which is external to c.

We argue that p_{PA} is a presupposition of c in a similar way to before. Fix arbitrary x. The idea is to distinguish between our justified belief in the sentence $\varphi(x)$ in the language $\mathcal{L}_{\mathsf{PA}}$ of PA, and our justified belief in the corresponding sentence $T(\ulcorner\varphi(\underline{x})\urcorner)$ in the expanded language $\mathcal{L}_T$. The way in which our justified belief in $\varphi(x)$ propagates to $T(\ulcorner\varphi(\underline{x})\urcorner)$ presupposes the validity of the sequent (UT1). But this propagation of beliefs, from $\varphi(x)$ to $T(\ulcorner\varphi(\underline{x})\urcorner)$, must occur at some point during our stability-project c. Thus, if we were to doubt the validity of (UT1), we would also doubt consequences of the form $T(\ulcorner\varphi(\underline{x})\urcorner)$. Hence, we would also doubt the overall significance or competence of c.

For suppose we hold a justified belief in $\varphi(x)$. How would we come to hold a justified belief in the $\mathcal{L}_T$-sentence $T(\ulcorner\varphi(\underline{x})\urcorner)$? We might try to exhibit a proof of $T(\ulcorner\varphi(\underline{x})\urcorner)$ from the axioms of PA, and assure ourselves that justified belief is preserved throughout this process. But this isn't possible: $T(\ulcorner\varphi(\underline{x})\urcorner)$ is not formulated in the language of PA, and PA can't interpret $T(\ulcorner\varphi(\underline{x})\urcorner)$ uniformly for all φ.

How else might we come to hold a justified belief in $T(\ulcorner\varphi(\underline{x})\urcorner)$? Analogously to section 4.1.1, we suggest our remaining option is this: we argue that since the truth predicate T is disquotational, then $\varphi(x)$ *means the same thing* as $T(\ulcorner\varphi(\underline{x})\urcorner)$. Thus, since we hold a justified belief in $\varphi(x)$, our justified belief is preserved in virtue of the identical meaning of these two sentences. Insofar as the meanings of φ and $T(\ulcorner\varphi(\underline{x})\urcorner)$ consist in their truth conditions, then we infer a justified belief in $T(\ulcorner\varphi(\underline{x})\urcorner)$ on the basis of our justified belief in $\varphi(x)$, and the following principle:

(†) Whenever $\varphi(x)$ is true, $T(\ulcorner\varphi(\underline{x})\urcorner)$ is true.

The point is that for our justified belief in the arithmetical consequences $\varphi(x)$ of PA to transfer to the corresponding $\mathcal{L}_T$-sentences $T(\ulcorner\varphi(\underline{x})\urcorner)$, we require collateral warrant for the principle (†).

But the principle (†) for which we require collateral warrant *is* just the proposition asserting the validity of the corresponding instance of the sequent (T1): that whenever $\mathcal{M} \models \varphi(x)$, we have $\mathcal{M} \models T(\ulcorner\varphi(\underline{x})\urcorner)$. Thus, if we infer justified belief in $T(\ulcorner\varphi(\underline{x})\urcorner)$ on the basis of our justified belief in $\varphi(x)$ and the principle (†), then we presuppose the validity of the sequent $\varphi(x) \Rightarrow$

$T(\ulcorner\varphi(\underline{x})\urcorner)$.

So if we were to doubt the proposition asserting the validity of instances of the form $\varphi(x) \Rightarrow T(\ulcorner\varphi(\underline{x})\urcorner)$ (where doubt might include even agnosticism about the validity of instances of the form $\varphi(x) \Rightarrow T(\ulcorner\varphi(\underline{x})\urcorner)$), we would also in general doubt the significance or competence of our cognitive project c. For the success of c requires that we are able to arrive at a *justified belief* in any $\mathcal{L}_T$-sentence $T(\ulcorner\varphi(\underline{x})\urcorner)$, where $\varphi(x)$ is an arithmetical consequence of PA. And justified belief excludes doubt.

Let us remark on the argument for the validity of the sequent (UT2). To make a similar argument, we would want to identify a stage during c during which our justified beliefs in $\mathcal{L}_T$-sentences $T(\ulcorner\varphi(\underline{x})\urcorner)$ propagate to the corresponding arithmetical consequences $\varphi(x)$ of PA. Analogously to section 4.1.1, we admit that it seems a little strange to think of this as happening during c: we start out with justified beliefs in purely arithmetical sentences, and after that, our beliefs are supposed to propagate to sentences in the expanded language, rather than the other way around. But as before, the validity of the sequent (UT2) seems to ensure a certain *coherence* between our justified beliefs in arithmetical sentences $\varphi(x)$, and the corresponding $\mathcal{L}_T$-sentences $T(\ulcorner\varphi(\underline{x})\urcorner)$. It would presumably be rather strange if we were able to infer a justified belief in $\mathcal{L}_T$-sentences $T(\ulcorner\varphi(\underline{x})\urcorner)$ on the basis that we hold a justified belief in $\varphi(x)$ and the formal counterpart to the principle (†) above, but *not* have this cohere with the idea that if we hold a justified belief in an arbitrary $\mathcal{L}_T$-sentence $T(\ulcorner\varphi(\underline{x})\urcorner)$, we can also infer a justified belief in the corresponding arithmetical sentence $\varphi(x)$ on that basis coupled with the formal counterpart to the principle † above. So it seems like doubting the proposition asserting the validity of (UT2) would result in a peculiar epistemic non-equivalence between arithmetical sentences and $\mathcal{L}_T$-sentences, rather than rationally compel us to think we are unable to carry out c in a significant or competent way. So, as before, we will assume that this natural coherence is something we want. Then the proposition asserting the validity of (UT2) is also a presupposition of c.

Next we want to exhibit an attempt to justify p_{PA} which is external to c. We exhibit this justificatory attempt in a similar way as in section 4.1.1, for the type-free sequents (T1) and (T2), where we rely on *no* further extension validators. Here is the argument for the sequent (UT2) (the argument for (UT1) is similar). Suppose we want to justify the validity of a particular instance of the sequent (UT2), say $T(\ulcorner\varphi(\underline{x})\urcorner) \Rightarrow \varphi(\underline{x})$, for some $\mathcal{L}_{\mathsf{PA}}$-formula $\varphi(x)$. That is, suppose we want to justify $T(\ulcorner\varphi(\underline{x})\urcorner) \models \varphi(\underline{x})$. Fix a model $\mathcal{M}$ of PA such that $\mathcal{M} \models T(\ulcorner\varphi(\underline{x})\urcorner)$. Moving to the metalanguage, all this means is that $\mathcal{M} \models \varphi(x)$. And this is what we set out to conclude.

At this point, we have argued that we are warranted in extending PA by the uniform disquotational truth sequents (UT1) and (UT2) by induced entitlement of cognitive project. Since we have argued that induced entitlements of cognitive project can underwrite justified belief, we thereby arrive at a justified belief in the principles of the theory PA^U; i.e., we arrive at stage (3) of the stability-project c above. Next we argue that we are warranted in extending PA^U by the rule $(\mathrm{Ind}_{\mathcal{L}_{\mathsf{PA}^U}})$ by induced entitlement of the cognitive project c.

Let now p_{PA^U} be the extension validator for PA^U asserting the validity of the rule ($\mathrm{Ind}_{\mathcal{L}_{\mathsf{PA}^U}}$), and suppose we hold a justified belief in ZF. We need to show that p_{PA^U} is an induced entitlement of cognitive project c. To show that p_{PA^U} is an induced entitlement of cognitive project c, we need to show two things: (i) that p_{PA^U} is a presupposition of c, and (ii) that there exists an attempt to justify p_{PA^U} which is external to c.

First, we argue that p_{PA^U} is a presupposition of c. We will approach this in a different way than we have previously. We claim that the proposition asserting the validity of the reflection rule ($\mathrm{R}_{\mathsf{PA}^U}$) is a presupposition of c. The argument is similar to the argument we gave for the rule (R_{EA}) in section 4.1.2. At that point, we will have shown that the propositions asserting the validity of (UT1), (UT2), and the proposition asserting the validity of ($\mathrm{R}_{\mathsf{PA}^U}$), are presuppositions of c. We will then appeal to the following general principle: if the proposition asserting the validity of some (set of) principle(s) (P) is a presupposition of an arbitrary cognitive project, and (P) derives some other principle (R), then the proposition asserting the validity of (R) is also a presupposition of that cognitive project. Everything will then follow from our claims that the propositions asserting the validity of the uniform disquotational truth sequents (UT1) and (UT2), and the proposition asserting the validity of the rule ($\mathrm{R}_{\mathsf{PA}^U}$), are presuppositions of c. We show below how to recover fully extended induction (Ind_T) from $\mathsf{PA}^U + (\mathrm{R}_{\mathsf{PA}^U})$.

So first, let us argue that ($\mathrm{R}_{\mathsf{PA}^U}$) is a presupposition of c. As before, the idea is to distinguish between our justified belief in the *axioms* of PA^U, and our justified belief in the entire *theory* PA^U. We argue that the way in which our justified belief in the axioms of PA^U propagates to the entire theory

PA^U presupposes the validity of $(\mathrm{R}_{\mathsf{PA}^U})$. But this propagation of beliefs, from axioms to theory, must occur at some point during c. Thus, the success of c requires the validity of $(\mathrm{R}_{\mathsf{PA}^U})$.

Suppose we hold a justified belief in the axioms of PA^U. How would we come to hold a justified belief in an arbitrary theorem $\Gamma \Rightarrow \Delta$ of PA^U? (We suppose $\Gamma \Rightarrow \Delta$ is not itself an axiom of PA^U.) The plausible suggestion, as in section 4.1.2, is to exhibit a proof of $\Gamma \Rightarrow \Delta$ from the axioms of PA^U. But then our justified belief in $\Gamma \Rightarrow \Delta$ is grounded in *more* than just our justified belief in the *axioms* of PA^U. It is our reliance on the inference rules of classical logic when we write down proofs from the axioms of PA^U, coupled with our justified belief in the axioms of PA^U, by which we arrive at a justified belief in $\Gamma \Rightarrow \Delta$. So, consider a scenario in which we write down a proof of $\Gamma \Rightarrow \Delta$ from the axioms of PA^U. The point is that for our justified belief in the axioms of PA to transfer to $\Gamma \Rightarrow \Delta$, we require collateral warrant for the following principle: if $\Gamma \Rightarrow \Delta$ is provable from the axioms of PA, then $\Gamma \Rightarrow \Delta$ holds.

But the natural way of formalizing the informal claim for which we require collateral warrant *is* just the proposition asserting the validity of the following reflection rule $(\mathrm{R}_{\mathsf{PA}^U})$. Thus, if we infer justified belief in $\Gamma \Rightarrow \Delta$ on the basis of our justified belief in the axioms of PA^U and our reliance on the inference rules of classical logic when we construct proofs in PA^U, then we presuppose the validity of $(\mathrm{R}_{\mathsf{PA}^U})$. If we were to doubt the proposition asserting the validity of $(\mathrm{R}_{\mathsf{PA}^U})$ (where doubt might include even agnosticism about the validity of $(\mathrm{R}_{\mathsf{PA}^U})$), we would also in general doubt the significance or competence of the cognitive project c. For the success of c requires that we are able to arrive at a *justified belief* in any theorem of PA^U. And justified belief excludes doubt.

So, the proposition asserting the validity of the rule ($\mathrm{R}_{\mathsf{PA}^U}$) is a presupposition of c. At this point, we have argued that the propositions asserting the validity of the sequents (UT1) and (UT2), and the proposition asserting the validity of the rule ($\mathrm{R}_{\mathsf{PA}^U}$), are each presuppositions of c.

Next, we argue that if the propositions asserting the validity of some (set of) principle(s) (P) is a (are) presupposition(s) of an arbitrary cognitive project c, and (P) derives some other principle (R), then the proposition asserting the validity of (R) is a presupposition of that cognitive project. The argument is simple: suppose p_1 and p_2 are presuppositions of c asserting the validity of principles (P1) and (P2) respectively. Suppose that over some theory S, (P1) + (P2) derives (R). If one were to doubt the validity of (R), one would also (rationally) doubt the validity of either (P1) or (P2) (assuming doubting S is out of the question). Then in either case, we would thereby rationally doubt the significance or competence of c, because the propositions p_1 and p_2 asserting the validity of principles (P1) and (P2) respectively are presuppositions of c.

Finally, we show that over PA, the sequents (UT1) and (UT2) and the rule (R_{PA}) derive the rule (Ind_T):[10]

Lemma 8. $\mathrm{R}(\mathsf{PA}^U)$ derives (Ind_T).

Proof. Let $\varphi(x)$ be an arbitrary $\mathcal{L}_T$-formula with one free variable. Suppose that $\mathrm{R}(\mathsf{PA}^U)$ derives $\Gamma, \varphi(x) \Rightarrow \varphi(x+1), \Delta$. We want to show that $\mathrm{R}(\mathsf{PA}^U)$

[10] This is essentially Lemma 5 of (Fischer et al., 2017).

derives $\Gamma, \varphi(0) \Rightarrow \varphi(t), \Delta$ (here x is not free in $\varphi(0)$, and Δ, Γ, t are arbitrary). The following rule is admissible in PA^U,[11] for any $n \in \omega$:

$$\frac{\Gamma, \varphi(x) \Rightarrow \varphi(x+1), \Delta}{\Gamma, \varphi(0) \Rightarrow \varphi(\underline{n}), \Delta}$$

Thus, PA^U proves:

$$\Rightarrow \mathrm{Pr}^2_{\mathsf{PA}^U}(\ulcorner \Gamma, \varphi(x) \Rightarrow \varphi(x+1), \Delta \urcorner, \ulcorner \Gamma, \varphi(0) \Rightarrow \varphi(\dot{y}), \Delta \urcorner).$$

By our assumption and $(\mathrm{R}_{\mathsf{PA}^U})$ we obtain:

$$\Gamma, \varphi(0) \Rightarrow \varphi(y), \Delta.$$

□

Putting all of this together, the proposition p_{PA^U} asserting the validity of the rule $(\mathrm{Ind}_{\mathcal{L}_{\mathsf{PA}^U}})$ is a presupposition of our stability-project c.

Finally, we want to exhibit an attempt to justify p_{PA^U} which is external to c. So, suppose we want to justify:

$$\Gamma, \varphi(x) \Rightarrow \varphi(x+1), \Delta \models \Gamma, \varphi(0) \Rightarrow \varphi(t), \Delta,$$

where x is not free in the lower sequent, t is an arbitrary term, Γ, Δ are arbitrary sets of formulas, and $\varphi(v)$ is any formula in the language $\mathcal{L}_T$.

[11] In fact is admissible in weaker theories than PA^U.

Fix a standard model $\mathcal{M}$ such that:

$$\mathcal{M} \models \Gamma, \varphi(x) \Rightarrow \varphi(x+1), \Delta.$$

Moving to a metalanguage, this means that whenever $\mathcal{M} \models \Gamma \wedge \varphi(x)$, we have $\mathcal{M} \models \varphi(x+1) \vee \Delta$.

We want to conclude that:

$$\mathcal{M} \models \Gamma, \varphi(0) \Rightarrow \varphi(t), \Delta.$$

So, it would be enough to establish the claim that whenever $\mathcal{M} \models \Gamma \wedge \varphi(0)$, we have $\mathcal{M} \models \varphi(t) \vee \Delta$. Putting everything together, it would be enough to establish the following claim. If:

$$\big(\text{if } \mathcal{M} \models \Gamma \wedge \varphi(x) \text{ then } \mathcal{M} \models \varphi(x+1) \vee \Delta\big) \text{ and } \mathcal{M} \models \Gamma \wedge \varphi(0),$$

then:

$$\mathcal{M} \models \varphi(t) \vee \Delta.$$

In particular (by logic), it would be enough to establish the following claim. If:

$$\big(\text{if } \mathcal{M} \models \varphi(x) \text{ then } \mathcal{M} \models \varphi(x+1)\big) \text{ and } \mathcal{M} \models \varphi(0),$$

then:

$$\mathcal{M} \models \varphi(t).$$

We suggest the following is the natural way of establishing this claim: we *prove by induction* on complexity of formulas that $\mathcal{M} \models \varphi(t)$.

But let us think about what has happened, when we prove $\mathcal{M} \models \varphi(t)$ by induction on complexity of formulas. Since we want to conclude $\mathcal{M} \models \varphi(t)$, the formula we want to run our induction on is the following formula of one free variable: $\mathcal{M} \models \varphi(v)$. And so we cannot possibly hope to prove that $\mathcal{M} \models \varphi(t)$ by induction on the complexity of formulas in the theory PA^U itself. First, PA^U cannot express the notion of a model of itself. Second, the induction schema of PA^U applies only to arithmetical formulas, but the formula $\varphi(v)$ is an $\mathcal{L}_T$ formula. In particular, a similar thing has happened as in the case of reflection: any theory which would be able to derive $\mathcal{M} \models \varphi(t)$ as the conclusion of an instance of induction has to be stronger than PA^U itself.

So, we need a theory which can: (1) talk about standard models of PA^U, (2) can define the truth predicate T for formulas in the language of PA, and (3) lies outside the scope of c. And as usual, ZF is such a theory. So let $\psi(v)$ be the formula with one free variable in the language of set theory which says that $\mathcal{M} \models \varphi(v)$. Then the formal counterpart to the metatheoretic claim, whose validity we are relying on in our justificatory attempt, is the following instance of induction:

$$\Big(\forall x(\psi(x) \to \psi(x+1) \land \psi(0)\Big) \to \psi(t),$$

which is derivable in ZF. So when we establish $\mathcal{M} \models \varphi(t)$ by induction in this way, we rely on the following claim: that the rule $(\mathrm{Ind}_{\mathcal{L}_{\mathsf{ZF}}})$ is valid. That is, we have relied on a version of the very principle whose validity we set out to justify in the first place. But the proposition asserting the validity of $(\mathrm{Ind}_{\mathcal{L}_{\mathsf{ZF}}})$ is itself an extension validator for ZF. Thus, in our attempt to justify the validity of $(\mathrm{Ind}_{\mathcal{L}_{\mathsf{PA}^U}})$, we relied on an extension validator for $\mathsf{ZF} \supseteq \mathsf{PA}^U$. Since we hold a justified belief in ZF, and ZF is not embedded in the scope of c, we have shown that our attempt to justify the validity of $(\mathrm{Ind}_{\mathcal{L}_{\mathsf{PA}^U}})$ is external to c.

This brings us to our conclusion. We have argued that we are warranted in extending PA^U by the rule (Ind_T) by induced entitlement of cognitive project. Since we have argued that induced entitlements of cognitive project can underwrite (mathematically) justified belief, we thereby arrive at a justified belief in the principles of the theory $(\mathsf{PA}^U)_\omega$; i.e., we arrive at stage (5) of the stability-project c above. This completes our argument for the theory $(\mathsf{PA}^U_{\mathrm{Ax}_{\mathsf{PA}}})_\omega$.

6.1.2 Justified belief in $\mathsf{PA}^T_{\mathrm{Ax}_{\mathsf{PA}}}$

Next, consider the following stability-project c for PA, a reconstruction of the idea that the first-orderist, from a justified belief in PA, may coherently also claim to hold a justified belief in the principles of the theory $\mathsf{PA}^T_{\mathrm{Ax}_{\mathsf{PA}}}$:

(1) We start out with a justified belief in the principles of the theory PA.

($\vdots$) $\vdots$

(n) We are thereby justified in believing the principles of $\mathsf{PA}^T_{\mathrm{Ax}_{\mathsf{PA}}}$.

We want to show that we can fill in the $\vdots$ in this process in the following way:

(1) We start out with a justified belief in the principles of the theory PA.

(2) We are warranted in extending PA by the sequents ($T\ =_1$), ($T\ =_2$), ($T\wedge_1$), ($T\wedge_2$), ($T\neg_1$), ($T\neg_2$), ($T\forall_1$), and ($T\forall_2$), by induced entitlement of this cognitive project.

(3) We are thereby justified in believing the principles of PA^T.

(4) We are thereby justified in believing the principles of $\mathsf{PA}^T_{\mathrm{Ax}_{\mathsf{PA}}}$.

Analogously to the stability-project defined in section 6.1.1, here notice that the move from step (3) to (4) is legitimized by Leigh's Theorem 1: the theory PA^T can interpret the rule (D). Thus, if we hold a justified belief in the axioms of PA^T, then relying only on the validity of the corresponding reflection rule ($\mathrm{R}_{\mathsf{PA}^T}$), the rule (D) inherits our justified belief in an ordinary way.

Let p_{PA} be an extension validator asserting the validity of any of the compositional truth sequents. Suppose we hold a justified belief in ZF. We want to show that p_{PA} is an induced entitlement of cognitive project c. To show that p_{PA} is an induced entitlement of cognitive project c, we need to show two things: (i) that p_{PA} is a presupposition of c, and (ii) that there exists an attempt to justify p_{PA} which is external to c.

First we argue that p_{PA} is a presupposition of c. We will do this in the same way that we argued the proposition asserting the validity of the rule

$(\mathrm{Ind}_{\mathcal{L}_{\mathsf{PA}^U}})$ was a presupposition of c. We have already argued that the proposition asserting the validity of the sequents (UT1), (UT2), and the proposition asserting the validity of the reflection rule $(\mathrm{R}_{\mathsf{PA}^U})$, are presuppositions of c. We again appeal to the following general principle: if the proposition asserting the validity of some (set of) principle(s) (P) is a presupposition of an arbitrary cognitive project, and (P) derives some other principle (R), then the proposition asserting the validity of (R) is also a presupposition of that cognitive project. Everything will then follow from our claim that the proposition asserting the validity of the sequents (UT1), (UT2), and the proposition asserting the validity of the reflection rule $(\mathrm{r}_{\mathsf{PA}^U})$, are presuppositions of c. We show next how to recover the fully compositional truth sequents from $\mathsf{PA}^U + (\mathrm{R}_{\mathsf{PA}^U}) = \mathrm{R}(\mathsf{PA}^U)$.[12]

Lemma 9. In $\mathrm{r}(\mathsf{PA}^U)$ we can derive $(T =_{1\text{–}2})$, $(T\wedge_{1\text{–}2})$, $(T\neg_{1\text{–}2})$, and $(T\forall_{1\text{–}2})$.[13]

Proof. We note that PA^U derives each instance of the following schematic versions of the compositional clauses:

1. $\mathsf{PA}^U \vdash T(s \underset{.}{=} t) \Leftrightarrow \mathrm{val}(s) = \mathrm{val}(t)$ for all closed terms s, t;

2. $\mathsf{PA}^U \vdash T(\varphi) \wedge T(\psi) \Leftrightarrow T(\varphi \underset{.}{\wedge} \psi)$ for all sentences φ, ψ;

3. $\mathsf{PA}^U \vdash \neg T(\varphi) \Leftrightarrow T(\underset{.}{\neg}\varphi)$ for all sentences φ;

4. $\mathsf{PA}^U \vdash \forall x T(\varphi(\underline{x})) \Leftrightarrow T(\underset{.}{\forall} x \varphi(x))$ for all sentences φ;

Here $\Leftrightarrow$ is used to indicate that both "directions" of the sequents are derivable;

[12] This strategy is similar to that in Lemma 4 of Fischer et al., 2017.

[13] Hence we can do the same in $\mathrm{R}(\mathsf{PA}^U)$.

i.e. for 1 above that both the sequents

$$T(s \mathrel{\dot{=}} t) \Rightarrow \mathrm{val}(s) = \mathrm{val}(t)$$

and

$$\mathrm{val}(s) = \mathrm{val}(t) \Rightarrow T(s \mathrel{\dot{=}} t)$$

are derivable for all closed terms s, t.

Formalizing these facts in PA^U, we obtain (for example)

$$\Rightarrow \mathrm{Pr}_{\mathsf{PA}^U}(\ulcorner \mathrm{Sent}_{\mathcal{L}_T}(x \mathrel{\dot{=}} y), \mathrm{val}(x) = \mathrm{val}(y) \Rightarrow T(x \mathrel{\dot{=}} y)\urcorner).$$

Thus, in $\mathrm{r}(\mathsf{PA}^U)$, we can move to the full quantifiable statement:

$$\mathrm{Sent}_{\mathcal{L}_T}(x \mathrel{\dot{=}} y), \mathrm{val}(x) = \mathrm{val}(y) \Rightarrow T(x \mathrel{\dot{=}} y),$$

which is compositional clause $(T =_1)$. The other cases are entirely similar. □

Thus, any proposition p_{PA} asserting the validity of one of the fully compositional truth sequents is a presupposition of our stability-project c.

Finally we have to exhibit an attempt to justify the proposition p_{PA} asserting the validity of all the fully compositional truth sequents, which is external to c. To give the general idea, we will exhibit an attempt to justify the proposition asserting the validity of the fully compositional truth sequent $(T\forall_1)$. We argue that we can do this by relying on no other extension validators. So suppose we want to justify:

$$\mathrm{Sent}_{\mathcal{L}_T}(\underset{\cdot}{\forall} yx), \forall y T(x[y/v]) \models T(\underset{\cdot}{\forall} yx).$$

Fix a model $\mathcal{M}$ of PA such that:

$$\mathcal{M} \models \mathrm{Sent}_{\mathcal{L}_T}(\underset{\cdot}{\forall} yx) \wedge \forall y T(x[y/v]).$$

Moving to the metalanguage, this means that $\mathcal{M} \models \mathrm{Sent}_{\mathcal{L}_T}(\underset{\cdot}{\forall} yx)$ and $\mathcal{M} \models \forall y T(x[y/v])$. In particular, $\mathcal{M} \models \forall y T(x[y/v])$. This means that for all y in the domain of $\mathcal{M}$, we have $\mathcal{M} \models T(x[y/v])$. And all the latter means is that $\mathcal{M} \models x[y/v]$. So, unpacking our assumptions, we have concluded that for all y in the domain of $\mathcal{M}$ we have $\mathcal{M} \models x[y/v]$. Equivalently, this means that $\mathcal{M} \models \forall y x[y/v]$. But all *this* means is just that $\mathcal{M} \models T(\forall y x[y/v])$. We have essentially invoked the idea from chapter 4, that $\mathcal{M} \models \varphi$ and $\mathcal{M} \models T(\varphi)$ *mean* the same thing. This, coupled with the meaning of the logical symbol $\forall$, is what drives this line of reasoning. The cases for the other compositional truth sequents are similar.

This brings us to our conclusion. We have argued that we are warranted in extending PA by fully compositional truth sequents by induced entitlement of cognitive project. Since we have argued that induced entitlements of cognitive project can underwrite belief, we thereby arrive at a justified belief in the principles of the theory PA^T; i.e., we arrive at stage (3) of the stability-project c above. This completes our argument for the theory $\mathsf{PA}^T_{\mathsf{Ax}_{\mathsf{PA}}}$.

Let us summarize. We have argued that the following two cognitive projects capture the idea from chapter 2, that on the basis of their justified belief in PA, the first-orderist may claim to thereby hold a justified belief in the theories $(\mathsf{PA}^U_{\mathrm{Ax}_{\mathsf{PA}}})_\omega$ and $\mathsf{PA}^T_{\mathrm{Ax}_{\mathsf{PA}}}$ considered individually.

Stability-project 1 for PA:

(1) We start out with a justified belief in the principles of the theory PA.

(2) We are warranted in extending PA by the sequents (UT1) and (UT2) by induced entitlement of this cognitive project.

(3) We are thereby justified in believing the principles of PA^U.

(4) We are warranted in extending PA^U by the rule (Ind_T) by induced entitlement of this cognitive project.

(5) We are thereby justified in believing the principles of $(\mathsf{PA}^U)_\omega$.

(6) We are warranted in extending $(\mathsf{PA}^U)_\omega$ by the rule (D) by justification.

(7) We are thereby justified in believing the principles of $(\mathsf{PA}^U_{\mathrm{Ax}_{\mathsf{PA}}})_\omega$.

Stability-project 2 for PA:

(1) We start out with a justified belief in the principles of the theory PA.

(2) We are warranted in extending PA by the compositional truth sequents by induced entitlement of this cognitive project.

(3) We are thereby justified in believing the principles of PA^T.

(4) We are warranted in extending PA^T by the rule (D) by justification.

(5) We are thereby justified in believing the principles of $\mathsf{PA}^T_{\mathrm{Ax}_{\mathsf{PA}}}$.

We also set out to explain how the first-orderist is not thereby justified in believing the principles of the unified theory $(\mathsf{PA}^T_{\mathrm{Ax}_{\mathsf{PA}}})_\omega$. This boils down to explaining how the first-orderist is not thereby warranted in extending PA by *both* the fully compositional truth sequents for PA, *and* the fully extended induction rule. We claim to have achieved this. For if c is any stability-project for PA, then $(\mathsf{PA}^T)_\omega$ cannot witness the success of c: $(\mathsf{PA}^T)_\omega$ is not embedded in the scope of c. So, in the presence of the other, one of the (set of) fully compositional truth sequents for PA, or the fully extended induction rule, cannot be warranted by induced entitlement of c. By defining the right kind of cognitive project, and by keeping the goal of the cognitive project in sharp focus, we have a story about how it is possible that on the basis of their justified belief in PA, the first-orderist may claim to thereby hold a justified belief in the theories $(\mathsf{PA}^U_{\mathrm{Ax}_{\mathsf{PA}}})_\omega$ and $\mathsf{PA}^T_{\mathrm{Ax}_{\mathsf{PA}}}$ considered individually, but claim to not thereby hold a justified belief in both of these theories at the same time. Everything occurs in the context of a specific cognitive goal.

We note also that the goal of the first-orderist's stability projects captures the essence of reconciling the notion of $\mathcal{L}_{\mathsf{PA}}$-epistemic stability, and the weak version of the ICT, from chapter 2. Recall that PA is $\mathcal{L}_{\mathsf{PA}}$-epistemically stable if there exists a coherent rationale for accepting PA that does not entail or otherwise rationally oblige a theorist to accept statements in the language of PA which cannot be derived from the axioms of PA. Recall also the corresponding weak version of the ICT: anyone who accepts the axioms of PA is thereby also implicitly committed to accepting various additional statements Γ which are formally independent of the axioms of PA.[14] At this point, to say that the notion of $\mathcal{L}_{\mathsf{PA}}$-epistemic stability is reconcilable with the corresponding version of the weak ICT is just to say that the first-orderist can successfully carry out a stability-project for PA.

So, at this stage, we have answered the following question from chapter 3: what could the warrant $\mapsto$ possibly consist in, such that we can make sense of the following scenario?

$$\begin{aligned}\text{justified belief in } \mathsf{PA} &\mapsto \text{ justified belief in } \mathsf{PA}^{T}_{\mathrm{Ax}_{\mathsf{PA}}}\\ \text{justified belief in } \mathsf{PA} &\mapsto \text{ justified belief in } (\mathsf{PA}^{U}_{\mathrm{Ax}_{\mathsf{PA}}})_{\omega}\\ \text{justified belief in } \mathsf{PA} &\not\mapsto \text{ justified belief in } (\mathsf{PA}^{T}_{\mathrm{Ax}_{\mathsf{PA}}})_{\omega}\end{aligned}$$

The warrant $\mapsto$ is essentially induced entitlement of a stability-project c.

Before we return to our second fundamental question of interest: what is the epistemological force behind implicit commitments? let us define one other kind of cognitive project in our typed setting.

6.2 JAB-projects

At this point, we can also easily formulate a typed, one-stage version of Hypatia's iterated JMB project which would make sense of a scenario in which we *are* warranted in extending PA by both the fully compositional truth sequents for PA, and the fully extended induction rule. In such a scenario, we do thereby arrive at a justified belief in the reflection principle (R_{PA}).

[14] Recall that "acceptance" here is the all-encompassing version of the term.

Let us call a cognitive project c is a *JAB-project for* S (for the "justification of new arithmetical beliefs") just in case it has the following structure:

(1) We start out with a justified belief in the principles of the theory S.

($\vdots$) $\vdots$

(n) We are thereby justified in believing the principles of $\mathsf{S} + \mathsf{E}$, where $\mathsf{E} \subseteq \mathsf{SC} \cup \mathsf{IC}$ and $\mathsf{S} + \mathsf{E}$ is not conservative over S.

Let S be a suitable theory and c be a JAB-project for S. Recall that the collection:

$$\mathcal{E}_c = \{\mathsf{S}^* \supseteq \mathsf{S} : \mathsf{S}^* = \mathsf{S} + \mathsf{E} \text{ for some subset } \mathsf{E} \subseteq \mathsf{SC} \cup \mathsf{IC}\}$$

consists of all the possible extensions of S by our semantic and schematic components.

Definition 28. Let S be a suitable theory and c be a JAB-project for S. We call the collection:

$$\mathcal{S}_c = \{\mathsf{S}^* \supseteq \mathsf{S} : \mathsf{S}^* \in \mathcal{E}_c \text{ and } \mathsf{S}^* \text{ is a non-conservative extension of } \mathsf{S}\}$$

the *scope* of c.

Definition 29. Let T be a theory and c be a JAB-project for S. We say that T *is embedded in the scope of* c just in case there exists a theory S^* in the scope of c such that $\mathsf{S}^* \vdash \mathsf{T}$.

We say that a theory T *witnesses the success of* c just in case T is embedded in $\mathcal{S}_c$, $\mathsf{T} \supsetneq \mathsf{S}$, and one holds a justified belief in T. It is not necessarily the case that if T witnesses the success of a JAB-project c, then T is not conservative over S. Rather, the theory we acquire a justified belief in by the *end* of a JAB-project witnesses the success of that project, and is not conservative over S. The list of extension validators p_{T} is the same as before, as is the definition of an attempt to justify p_{T}, and the associated set of bases of justificatory attempts. Our definitions of an attempt to justify p_{T} being external/internal to apply to stability-projects straightforwardly generalize to JAB-projects. Similarly, our definition of a principle being warranted by induced entitlement of a stability-project straightforwardly generalizes to JAB-projects.

The following is an example of a JAB-project for PA:

(1) We start out with a justified belief in the principles of the theory PA.

(2) We are warranted in extending PA by the sequents $(T=_1)$, $(T=_2)$, $(T\wedge_1)$, $(T\wedge_2)$, $(T\neg_1)$, $(T\neg_2)$, $(T\forall_1)$, and $(T\forall_2)$, by induced entitlement of this cognitive project.

(3) We are thereby justified in believing the principles of PA^T.

(4) We are warranted in extending PA^T by the rule $(\mathrm{Ind}_{\mathcal{L}_{\mathsf{PA}^T}}) = (\mathrm{Ind}_T)$ by induced entitlement of this cognitive project.

(5) We are thereby justified in believing the principles of $(\mathsf{PA}^T)_\omega$.

In particular, the theory $(\mathsf{PA}^T)_\omega$, in which we hold a justified belief by successfully undertaking this cognitive project, includes the reflection principle

(R_{PA}).

With two types of cognitive project defined in our typed setting, let us finally return to our second fundamental question of interest: what is the *epistemological force* behind implicit commitments?

6.3 The force of *should*

Let us briefly retrace our steps. We began by asking: mathematically, what are implicit commitments? Our investigation was motivated by the idea of epistemic stability and the implicit commitment thesis. We proposed a framework for analyzing implicit commitments which clarifies how the notion of epistemic stability is reconcilable with non-trivial versions of the implicit commitment thesis. This framework cashed out implicit commitments as various principles extending a base theory of arithmetic S:

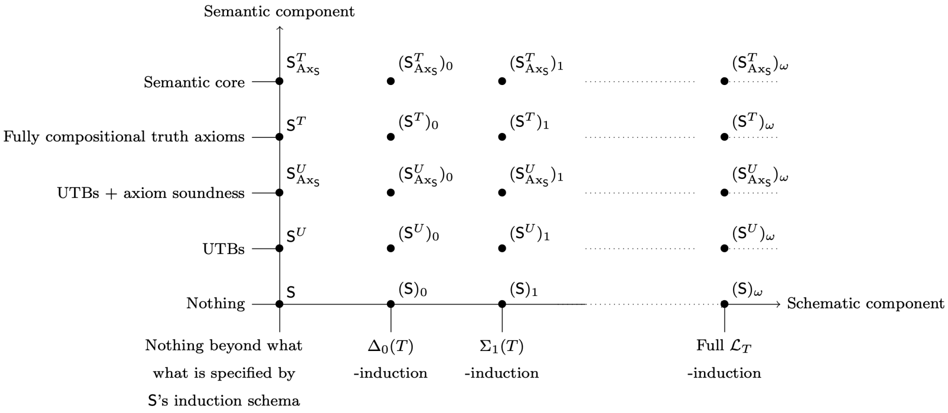
Semantic component
Semantic core
Fully compositional truth axioms
UTBs + axiom soundness
UTBs
Nothing
$\mathsf{S}^T_{\mathrm{Ax_S}}$
$(\mathsf{S}^T_{\mathrm{Ax_S}})_0$
$(\mathsf{S}^T_{\mathrm{Ax_S}})_1$
$(\mathsf{S}^T_{\mathrm{Ax_S}})_\omega$
S^T
$(\mathsf{S}^T)_0$
$(\mathsf{S}^T)_1$
$(\mathsf{S}^T)_\omega$
$\mathsf{S}^U_{\mathrm{Ax_S}}$
$(\mathsf{S}^U_{\mathrm{Ax_S}})_0$
$(\mathsf{S}^U_{\mathrm{Ax_S}})_1$
$(\mathsf{S}^U_{\mathrm{Ax_S}})_\omega$
S^U
$(\mathsf{S}^U)_0$
$(\mathsf{S}^U)_1$
$(\mathsf{S}^U)_\omega$
S
$(\mathsf{S})_0$
$(\mathsf{S})_1$
$(\mathsf{S})_\omega$
Schematic component
Nothing beyond what what is specified by S's induction schema
$\Delta_0(T)$ -induction
$\Sigma_1(T)$ -induction
Full $\mathcal{L}_T$ -induction

Figure 6.1: The semantic and schematic components of implicit commitment

We may think of one's implicit commitments $\mathsf{I}(\mathsf{S})$ in justifiably believing S as any of the theories of Figure 6.1. That is, we may think of one's implicit commitments in justifiably believing S as axiomatized by the axioms of S, together with suitable combinations of the following:

(1) Uniform disquotational truth principles for S.

(2) Fully compositional truth principles for S.

(3) Fully extended S-induction to the language $\mathcal{L}_T$.

(4) The axiom soundness principle for S.

The final question we want to answer is this: if we hold a belief in the principles of S, why *should* we also believe the principles of $\mathsf{I}(\mathsf{S})$?

The essential idea, which we formulated over chapters 5 and 6, is that the theories of Figure 6.1 all lie at the end of one of the kinds of cognitive projects we have defined in chapter 6. In other words, any of the theories of Figure 6.1 which are conservative over S are such that, if we held a justified belief in that theory, then it would witness the success of a stability-project for S. Any of the theories of Figure 6.1 which are not conservative over S are such that, if we held a justified belief in that theory, then it would witness the success of a JAB-project for S. The overall point is that our implicit commitments $\mathsf{I}(\mathsf{S})$ in holding a justified belief in S consist of principles such that, if we held a justified belief in $\mathsf{I}(\mathsf{S})$, then $\mathsf{I}(\mathsf{S})$ witnesses the success of a particular kind of cognitive project.

So, we can say why we *should* believe the principles of $\mathsf{I}(\mathsf{S})$, by saying what underlies the force of the warrant we have for these principles, in the contexts of these kinds of cognitive projects. And what we have seen is that essentially, everything goes via the following three (sets of) principles:

(1) Uniform disquotational truth principles for S.

(2) Fully compositional truth principles for S.

(3) Fully extended S-induction to the language $\mathcal{L}_T$.

So, let us fix an arbitrary subset $\mathcal{S}$ of (1)–(3) for the moment, and ask: why *should* we also believe the principles of $\mathcal{S}$?

Drawing together all of our observations, we argue that there are (at least) four underlying kinds of force to *should*. Suppose we have set out to undertake a cognitive project c, either a stability-project for S, or a JAB-project for S, as appropriate. We have argued that the principles of $\mathcal{S}$ are warranted by induced entitlement of the corresponding cognitive project c. This means two things: (i) the propositions asserting the validity of the principles of $\mathcal{S}$ are induced entitlements of c, and (ii) if we held a justified belief in $\mathsf{S}+\mathcal{S}$, then $\mathsf{S}+\mathcal{S}$ would witness the success of c. Suppose for concreteness also that we hold a justified belief in ZF, which forms the basis of the justificatory attempts involved, and which also derives the principles of $\mathcal{S}$ outright.

The first two reasons derive from condition (i) above: the propositions asserting the validity of the principles of $\mathcal{S}$ are induced entitlements of c. These two reasons correspond to the two clauses of a proposition's being an induced entitlement of cognitive project c. In particular, the first reason why we should believe the principles of $\mathcal{S}$ is: the propositions asserting the validity of the principles of $\mathcal{S}$ are presuppositions of c. Thus, these propositions are such that if we were to doubt them, we could not rationally maintain that our project was still significant or competent. We *have* to believe that the principles are valid, in order to undertake our cognitive project.

The second reason why we should believe the principles of $\mathcal{S}$ is: there exist attempts to justify the validity of the principles of $\mathcal{S}$ which are external to c. Thus, we have every reason to think that inferences made on the basis of the principles of $\mathcal{S}$ are valid inferences. We note that this seems to be more compelling than in the case of mere entitlements of cognitive project, where we are required only to *lack* sufficient countervailing reasons. Induced entitlements do better than this: we *have* sufficient reasons to believe that the principles of $\mathcal{S}$ are valid. Furthermore, justificatory attempts of mere entitlements of cognitive project are infinitely regressive, but we have argued that we can exhibit justificatory attempts of induced entitlements in a non-regressive manner relative to c. As we have seen, in articulating these reasons (i.e., in making the justificatory attempts), we rely on *some* further propositions, which we have argued are in turn of no more secure a prior understanding than the propositions we set out to justify in the first place. But the success of c does not *rationally* require us to hold a justified belief in these propositions. So justificatory attempts which are external to c seem to be more secure than infinitely regressive justificatory attempts in Wright's sense.

So, the first two reasons why we should believe the principles of $\mathcal{S}$ concern the validity of the principles of $\mathcal{S}$. If we are hoping to achieve a justified belief in the principles of the resulting theory $\mathsf{S} + \mathcal{S}$, we want to be sure that the principles via which we hope to arrive at this justified belief, and whose validity we *have to* rely on, do not result in belief in falsities. Because the propositions asserting the validity of the principles of $\mathcal{S}$ are induced entitlements of our cognitive project, we have every reason to think that this is indeed the case.

The third and fourth reasons why we should believe the principles of $\mathcal{S}$ derive from condition (ii) above: if we held a justified belief in $\mathsf{S} + \mathcal{S}$, then $\mathsf{S} + \mathcal{S}$ would witness the success of c. In particular, the third reason why we should believe the principles of $\mathcal{S}$ is: because believing the principles of $\mathcal{S}$ is the first epistemic step towards achieving our cognitive goal. All we have left to do is justify that belief. The fourth reason why we should believe the principles of $\mathcal{S}$ is: there exists a source of ordinary mathematical justification for the principles of $\mathcal{S}$ themselves. In particular, in every example we have seen, the basis $\mathsf{B} = \mathsf{ZF}$ of our attempts to justify the validity of the principles of $\mathcal{S}$ is also such that B derives, outright, the principles of $\mathcal{S}$. We leveraged this idea to argue that induced entitlements can underwrite (mathematically) justified belief. Thus, if we think that:

> if we hold a belief in the principles of a suitable theory T and T derives φ, then we should also believe φ,

then whenever

$$\text{justified belief in } \mathsf{S} \ \mapsto\ \text{justified belief in } \mathcal{S}$$

and $\mapsto$ is induced entitlement of some cognitive project c, then we should also believe the principles of $\mathcal{S}$ because we think the principles of $\mathcal{S}$ are justified in the ordinary mathematical sense of derivability.

So, if we hold a justified belief in a theory S, then we should believe the principles of $\mathcal{S}$ above because we have every reason to think that the principles of $\mathcal{S}$ do not result in belief in falsities; believing the principles of $\mathcal{S}$ sets us up to achieve our cognitive goal; and because the warrant we have for the principles of $\mathcal{S}$ witnesses a source of mathematical justification for the principles of $\mathcal{S}$.

Now let us consider a general theory of implicit commitments $\mathsf{I}(\mathsf{S})$, any of the theories of Figure 6.1 above. Suppose we have set out to undertake a suitable cognitive project, either a stability-project for S, or a JAB-project for

S, and $\mathsf{I}(\mathsf{S})$ is the theory in which we hold a justified belief at the end of our undertaking. If we hold a belief in the principles of S, why *should* we also believe arbitrary $\varphi \in \mathsf{I}(\mathsf{S})$?

We have already considered the case where φ is one of (1)–(3) above. To illustrate the idea, we will consider the cases where φ is the axiom soundness principle for S, or the reflection principle $(\mathrm{R}_{\mathsf{S}})$ for S.

First let φ be the axiom soundness principle for S. On one hand, in the two stability-projects we outlined earlier, our warrant for extending S^U by axiom soundness for S was simply justification itself, understood in the sense of derivability. This is legitimized because S^U interprets axiom soundness for S. On the other hand, suppose we set out to undertake a JAB-project for S, and we arrive at a justified belief in the principles of $(\mathsf{S}^T)_\omega$. Then the axiom soundness principle for S is derivable in $(\mathsf{S}^T)_\omega$. The point is that whenever we are engaged in either kind of cognitive project, and we hold a justified belief in S, we *should* believe the axiom soundness principle for S because:

(i) we should believe the axioms of S^U (respectively $(\mathsf{S}^T)_\omega$), and

(ii) we thereby have an ordinary source of mathematical justification for the axiom soundness principle for S.

So while our warrant for axiom soundness is not induced entitlement itself of our cognitive project, our warrant for axiom soundness is underwritten by induced entitlement. If we have set out to undertake a stability-project for S, then the force of our induced entitlement to the disquotational truth principles, coupled with the force of ordinary mathematical justification (understood as derivability), is what underlies the force of the should, when we say that we should believe the axiom soundness principle for S. Similarly, if we have set out to undertake a JAB-project for S, then the force of our induced entitlement to the fully compositional truth principles and fully extended induction, coupled with the force of ordinary mathematical justification (understood as derivability), is what underlies the force of the should, when we say that we should believe the axiom soundness principle for S.

Now let φ be the reflection principle (R_{S}) for S. (R_{S}) forms part of our implicit commitments whenever we are engaged in a JAB-project, in which the fully compositional truth principles and fully extended induction schema are warranted by induced entitlement. For over S^U, these principles are equivalent to (R_{S}). The point is that whenever we are engaged in a JAB-project, and we hold a justified belief in S, we *should* believe (R_{S}) because:

(i) we should believe the axioms of $(\mathsf{S}^T)_\omega$, and

(ii) we thereby have an ordinary source of mathematical justification for (R_{S}).

Again, while our warrant for (R_{S}) is not induced entitlement itself of our cognitive project, our warrant for (R_{S}) is underwritten by induced entitlements. The force of our induced entitlement to the compositional truth principles and fully extended induction, coupled with the force of ordinary mathematical justification (understood as derivability), is what underlies the force of the should, when we say that we should believe (R_{S}).

We note also that the context of the cognitive goal is what makes it possible to reconcile the idea that reflection principles *can* form part of our implicit commitments on the basis of a justified belief in an arithmetical theory S, with the idea of epistemic stability, which excludes reflection principles from our implicit commitments. Reflection principles form part of our implicit commitments when our cognitive goal is to justify new *arithmetical* beliefs, on the basis of a justified belief in S. But epistemic stability dovetails with the idea that reflection principles do not form part of our implicit commitments when our cognitive goal is to see what other mathematical principles we might believe, which are not themselves new *arithmetical* principles, on the basis of our justified belief in S.

We could tell a similar story about any arbitrary principle of $\mathsf{I}(\mathsf{S})$. Putting everything together, this is our story about why, if we hold a belief in the principles of S, we *should* also believe the principles of $\mathsf{I}(\mathsf{S})$. This is what underlies the *commitment* of implicit commitments. The force of the *should* consists in the force behind the induced entitlements of our cognitive project, coupled with the force of ordinary mathematical justification, understood in the sense of derivability. This is our answer to our second central question of interest: what is the epistemological force behind implicit commitments?[15]

[15] We note also that on the story we have told, the sense in which implicit commitments are *implicit* is this: they form part of one's cognitive goal, if one sets out to see what other mathematical principles one might come to believe, on the basis of a justified belief in S.

Chapter 7

Conclusion

Our goal in this dissertation was to answer the following question: what are implicit commitments of theories of arithmetic? Drawing on the account in (Nicolai & Piazza, 2019), we proposed a mathematical conception of implicit commitments by cashing out sets of implicit commitments of an arithmetical theory S as various theories $\mathsf{I}(\mathsf{S})$ extending S. In general, $\mathsf{I}(\mathsf{S})$ is axiomatized by a combination of semantic and schematic principles. However, unlike the account in (Nicolai & Piazza, 2019), we argued that there is no reason to think that any semantic (or schematic) principles are *fixed* among our implicit commitments.

Part of our broader motivation for this framework stemmed from the following three ideas: epistemic stability, the implicit commitment thesis, and the idea that one's implicit commitments in accepting an arithmetical theory S are typically understood to include reflection principles for S. We refined the notions of epistemic stability and the implicit commitment thesis. By doing so, we offered a framework in which these three ideas can coexist. Unlike Dean (2015), we think that it is possible for epistemic stability to be compatible with the implicit commitment thesis. In particular, it is possible for (1) there to exist a coherent rationale for accepting a given arithmetical theory S that does not entail or otherwise rationally oblige a theorist to accept statements in the language of S, which cannot be derived from the axioms of S, and (2) anyone who accepts the axioms of S to be implicitly committed to accepting various additional statements which are formally independent of S. Cashing out theories of implicit commitments as theories of truth extending S makes this possible. Reflection principles occur among one's implicit commitments just in case one's implicit commitments include both the fully compositional truth principles for S-sentences, and the fully extended S-induction schema to the expanded language of truth. Unlike Łełyk and Nicolai (2022) and Horsten (2021), we think that epistemic stability is compatible with the idea that reflection principles occur among one's implicit commitments. In this way, we hope to have vindicated the idea of epistemic stability.

We then approached the question: what is the epistemological force underlying implicit commitments? That is, what is the force underlying the *commitment* of implicit commitments? Our approach was motivated by our case study of first-orderism. In particular, we asked: what could the warrant

$\mapsto$ possibly consist in, such that we can make sense of the following scenario?

$$\begin{aligned} \text{accept } \mathsf{PA} &\mapsto \text{accept } \mathsf{PA}^{T}_{\mathrm{Ax}_{\mathsf{PA}}} \\ \text{accept } \mathsf{PA} &\mapsto \text{accept } (\mathsf{PA}^{U}_{\mathrm{Ax}_{\mathsf{PA}}})_{\omega} \\ \text{accept } \mathsf{PA} &\not\mapsto \text{accept } (\mathsf{PA}^{T}_{\mathrm{Ax}_{\mathsf{PA}}})_{\omega} \end{aligned}$$

We argued that the warrant $\mapsto$ cannot consist in ordinary empirical justification. Whether or not empirical justification is really the kind of warrant at play in mathematical contexts, empirical justification is closed under conjunction, and so cannot stand in for the warrant $\mapsto$. We also argued that the warrant $\mapsto$ cannot consist in either of two typical understandings of mathematical justification. First, we examined the idea of mathematical justification as the kind of warrant supporting our belief in distinguished families of axioms. This kind of warrant is independent in general of our beliefs about other families of axioms. Thus, independent axiom justification cannot stand in for the warrant $\mapsto$. For the warrant $\mapsto$ is a dependent kind of warrant. In particular, it depends on our initial belief in the axioms of PA. At that point we fixed an understanding of the broad notion of acceptance featuring in chapter 2. What we mean, when we say that the first-orderist accepts PA in the scenario above, is that the first-orderist has an independent mathematical justification for believing the axioms of PA. Thus, we refined our question of interest: what could the warrant $\mapsto$ possibly consist in, such that we can make sense of the following scenario?

$$
\begin{aligned}
\text{justified belief in } \mathsf{PA} &\mapsto \text{ justified belief in } \mathsf{PA}^{T}_{\mathrm{Ax}_{\mathsf{PA}}}\\
\text{justified belief in } \mathsf{PA} &\mapsto \text{ justified belief in } (\mathsf{PA}^{U}_{\mathrm{Ax}_{\mathsf{PA}}})_{\omega}\\
\text{justified belief in } \mathsf{PA} &\not\mapsto \text{ justified belief in } (\mathsf{PA}^{T}_{\mathrm{Ax}_{\mathsf{PA}}})_{\omega}
\end{aligned}
$$

We then turned to the idea of a dependent kind of mathematical justification, understood in the sense of derivability. That is, if one holds an independent mathematical justified belief in the axioms of a theory T and T derives φ, then one is warranted in justifiably believing φ. But mathematical justification in this sense cannot stand in for the warrant $\mapsto$ either. For in general, the principles of the first-orderist's implicit commitments are not derivable from the axioms of PA. Thus, traditional notions of justification cannot be what we are after.

Next we introduced Crispin Wright's notion of entitlements of cognitive project, and investigated the scenario in (Fischer et al., 2021), who apply entitlements to contexts of rational theory acceptance. We argued that entitlement of cognitive project also cannot stand in for the warrant $\mapsto$. The fundamental problem with entitlements of cognitive project in our context, is that entitlements cannot underwrite justified belief. This does not depend on whether we are talking about justified belief in an empirical sense, or justified belief in either of our mathematical senses. Whatever it is that distinguishes the kind of belief we hold in families of axioms justified by independent traditional methods, from the kind of belief we hold in general towards *any* mathematical principles, the analog of entitlements in contexts of mathematical theory acceptance cannot underwrite this distinguished kind of belief. During this discussion, we also fixed an understanding of *trust*: trust is what we place in propositions warranted by entitlement of cognitive project, on the basis of which we can form *some* kind of belief. But what we cannot form on the basis of entitlements alone is justified belief. Justification does not come for free.

However, we gleaned several useful structural features of entitlements of cognitive project during our discussion. In particular we were introduced to the ideas of cognitive projects, and cognitive goals. Using these ideas, we proposed a kind of warrant which we argued can stand in for $\mapsto$ above: induced entitlements of cognitive project. The theories of implicit commitments on the right hand sides in the scenario above correspond to the goals of certain kind of cognitive project. We argued that justified belief in the principles of those theories is warranted by (or is at least underwritten by) induced entitlement of the corresponding cognitive project. Induced entitlements are a dependent kind of warrant that differ from mathematical justification, understood in the sense of derivability. For the source of induced entitlements of cognitive project is not the source of mathematical justification, understood in the sense of derivability. However, we designed induced entitlements of cognitive project to witness such a justificatory source. Thus, we argued that the kinds of beliefs we form on the basis of induced entitlements are fundamentally justified, in a mathematical sense.

So, our implicit commitments occur in the context of a cognitive goal, and ultimately, the warrant we have for them is induced entitlement of the corresponding cognitive project. Thus, the *commitment* of implicit commitments consists in the commitment underlying induced entitlements. We suggested four reasons why one is committed to a principle warranted by induced entitlement of cognitive project. Three reasons center the cognitive project one is engaged in. If one seriously hopes to achieve the cognitive goal of one's project, then one must *rationally* presuppose the validity of certain principles warranted by induced entitlement. But one has every reason to think that these principles *are* valid: we can justify this in perfectly ordinary ways. For example, one may justify the validity of disquotational truth sequents by way of the meaning of the truth predicate. One may justify the validity of reflection principles by way of proof, and relying on stronger reflection principles. Neither of these justificatory strategies are necessarily regressive from the point of view of one's cognitive project. Furthermore, believing the principles warranted by induced entitlement puts one in a position to achieve one's cognitive goal. The final reason fits mathematical justification, understood in the sense of derivability, into the picture. The principles warranted by induced entitlement are also such that we may independently come to believe them. All things considered, we think these reasons are the most natural understanding of the *commitment* underlying implicit commitments.

Recall (from chapter 1) that part of the motivation for our project was that it is unclear how to understand the various epistemic notions featuring in existing accounts of the warrant we have for implicit commitments. Let us comment on these accounts in light of what we have said.

Fischer (2021) argue that our trust in reflection principles is as warranted as our trust in a theory S. On our understanding of trust, the account we have proposed seems to differ structurally from the account in (Fischer, 2021). For we have been interested in cases where we hold a justified belief in S, which is different than merely trusting S. Nonetheless, modulo this difference, if we understand "trust in S" as "beliefs in the axioms of S," then we disagree that our trust in reflection principles is as warranted as our trust in S. For we have argued that the warrant we have for reflection principles is induced entitlement of some cognitive project. This is not the same kind of warrant which informs our beliefs in the axioms of S in the first place.

If instead "trust in S" is understood instead as "beliefs in the principles of the theory of S," then perhaps we are in agreement with Fischer (2021). On our story, for one's beliefs in the axioms of S to propagate to the general theory of S, one must hold that a corresponding reflection principle for S is *valid*. If the latter is what is meant by "trust in a reflection principle," then perhaps we are in agreement. However, we distinguish between holding that a corresponding reflection principle for S is *valid*, and acquiring a justified belief in the reflection principle itself. We have argued holding that a reflection principle is valid does not automatically warrant extending S by the principle itself, for the latter depends on the goal of one's cognitive project. Thus, if "trust in a reflection principle" is understood as believing the reflection principle itself, we disagree with Fischer (2021). One can hold a belief in the principles of S without also holding a belief in the corresponding reflection principle itself as a result.

Fischer et al. (2021) and Horsten and Leigh (2016) argue that accepting a theory S in conjunction with a disquotational conception of truth provides sufficient warrant to accept reflection principles for S. Again, on our understanding of acceptance,[1] the account we have proposed seems to differ structurally from these accounts. Acceptance is the output of an entitlement, a kind of epistemic attitude which has no evidence (mathematical or otherwise) as input. This differs from justified belief. But again, modulo this difference, we disagree that accepting a theory S in conjunction with a disquotational conception of truth provides sufficient warrant to accept reflection principles for S. For in general, the kind of theory of implicit commitments corresponding to the goal of cognitive project one is engaged in, whereby one is warranted by induced entitlement in extending S by a disquotational conception of truth, need not include reflection principles for S. For a similar reason, this is where we disagree with the accounts in (Ketland, 2005, 2010; Shapiro, 1998), who argue that accepting S in conjunction with a fully compositional conception of truth provides sufficient warrant to accept reflection principles for S.

Finally, some accounts claim that the warrant for reflection principles comes from a "process of reflection" on the accepted theory S (Cieśliński, 2010; Horsten, 2021; Tennant, 2002). For example, during the process of reflection, one notes that one is ready to accept any sentence φ for which one can produce a proof in S. This gives one a reason to accept any sentence φ for which one can produce a proof in S (Cieśliński, 2010). The process of reflection seems a little unclear to us. In the spirit of our account, let us read acceptance as justified belief. In what sense is one *ready* to justifiably believe

[1] The understanding we fixed in chapter 4, rather than the umbrella term we used in chapter 2.

φ? Why does being ready to justifiably believe φ count as a reason to justifiably believe φ? One's justification seems to have appeared from nowhere, and we have explicitly avoided this idea. But perhaps *general* aspects of some of these accounts do align with our own. We have argued that the validity of a corresponding reflection principle is necessary for propagating our justified belief in the axioms of a theory to the general theory itself. So perhaps when we note that we are ready to justifiably believe any sentence φ for which we can produce a proof in S, this amounts to our realization that we cannot doubt the validity of the corresponding reflection principle. However, on the account we have put forward, this realization does not itself warrant our justified belief in the corresponding reflection principle. For the warrant we have for reflection principles occurs in the context of a cognitive project. In particular, our warrant for reflection principles occurs when reflection principles are compatible with our cognitive *goal*.

Overall, we hope to have provided a more thorough understanding of the implicit commitments of theories of arithmetic than existing accounts, whereby our epistemological understanding of implicit commitments coheres with our mathematical understanding of implicit commitments. At the very least, by fixing an understanding of the epistemic terminology ubiquitous in existing literature on implicit commitments, we hope to have offered a platform for uniform appraisal of existing commentaries.

7.1 Future directions

Let us draw things to a close by commenting on some future directions of this work. One issue we raised in chapter 3 concerned the difference, if any, between justification in mathematical contexts and justification in epistemic contexts. We proceeded on the general assumption that justification differs between these two contexts. In particular, we supposed that the kind of evidence underlying justification in epistemic contexts differs from the kind of evidence underlying justification in mathematical contexts. While we have had nothing to say about it, we also supposed that there is *some* notion of evidence at work in mathematical contexts: this is what distinguishes the kind of belief we hold in families of axioms justified by independent traditional methods, from the kind of belief we hold in any general mathematical statement. The force underlying mathematical evidence, if there is such a thing at all, is a topic we would like to investigate in future work. On one hand, if there is any such force, this would further inform our understanding of implicit commitments: we have argued that *whatever it is* that underlies mathematical justification, induced entitlements carry such a thing. On the other hand, perhaps it will turn out that there is no such force. We note that we do not think this significantly affects the current project: we still think *something* must distinguish the kind of belief we hold in families of axioms justified by independent traditional methods, from the kind of belief we hold in general towards any general mathematical principles. Even if this distinguishing factor is not underwritten by any evidential force, we may still hope to be able to say how mathematically justified beliefs are preserved in the kinds of cognitive project we have been interested in.

We also hope to be able to isolate canonical consistency sentences in our mathematical framework. For example, on the story we have told, if one's goal is to arrive at newly justified arithmetical beliefs by undertaking a JAB-project for PA, then one is implicitly committed to $\mathrm{Con}(\mathsf{PA})$. For one is implicitly committed to the reflection principle for PA, which derives $\mathrm{Con}(\mathsf{PA})$. In particular, if one holds a justified belief in PA, one should believe $\mathrm{Con}(\mathsf{PA})$ whenever one is engaged in such a cognitive project. But the kind of theory corresponding to the goal of a JAB-project is more than sufficient to deliver justified belief in $\mathrm{Con}(\mathsf{PA})$. For such a theory includes *fully* compositional truth principles, and *fully* extended induction. We would like to be able to isolate just how much compositional truth, and just how much extended induction, is both necessary and sufficient for deriving $\mathrm{Con}(\mathsf{PA})$. By doing so, we hope to be able to articulate precisely what is meant by saying: if one holds a justified belief in PA, then one should believe $\mathrm{Con}(\mathsf{PA})$. We would hope to make the case that the force of this should is underwritten by induced entitlements of a particular kind of cognitive project, whose cognitive goal corresponds to a theory which delivers $\mathrm{Con}(\mathsf{PA})$, and nothing stronger.

Finally, our framework gives rise to mathematical sentences φ such that in the context of a suitable cognitive project, one should believe φ, but one has no reason to believe $\neg\varphi$. $\mathrm{Con}(\mathsf{PA})$ is an example of such a sentence. On one hand, we have seen an example of a cognitive project whose successful undertaking results in our justifiably believing $\mathrm{Con}(\mathsf{PA})$ (any JAB-project for PA). On the other hand, if $\neg\mathrm{Con}(\mathsf{PA})$ were warranted by induced entitlement of some cognitive project, then in particular there would exist a theory in which we held an independently justified belief which derives $\neg\mathrm{Con}(\mathsf{PA})$. We take it that there are no such theories. Thus, we suggest that there are no cognitive projects whose successful undertaking results in our justifiably believing $\neg\mathrm{Con}(\mathsf{PA})$. If one holds a justified belief in PA, then by leveraging the idea of cognitive projects, we have articulated a sense in which one should believe $\mathrm{Con}(\mathsf{PA})$, but should not believe $\neg\mathrm{Con}(\mathsf{PA})$. In future work we hope to investigate other natural principles which can be characterized in this manner, and principles which cannot.

Bibliography

Addison, J. (1958). Separation principles in the hierarchies of classical and effective descriptive set theory. *Fundamenta Mathematicae*, *46*(2), 123–135. doi:https://doi.org/10.4064/fm-46-2-123-135

Addison, J., & Moschovakis, Y. N. (1968). Some consequences of the axiom of definable determinateness. *Proceedings of the National Academy of Sciences*, *59*(3), 708–712. doi:https://doi.org/10.1073/pnas.59.3.708

Baire, R., Borel, E., Hadamard, J., & Lebesgue, H. (1982). Five letters on set theory (G. H. Moore, Trans.). In, *Zermelo's axiom of choice: Its origins, development, & influence* (pp. 311–320). Springer-Verlag. (Original work published 1905).

Barton, N., Ternullo, C., & Venturi, G. (2020). On forms of justification in set theory. *The Australasian Journal of Logic*, *17*(4), 158–200. doi:https://doi.org/10.26686/ajl.v17i4.6579

Bernays, P. (1946). [Review of the article *The philosophy of Bertrand Russell*, by K. Gödel]. *The Journal of Symbolic Logic*, *11*(3), 75–79. doi:https://doi.org/10.2307/2266736

Bettazzi, R. (1892). Sui punti di discontinuità delle funzioni di variabile reale. *Circolo Matematico di Palermo, Rendiconti*, *6*, 173–195. Retrieved from https://link.springer.com/content/pdf/10.1007/BF03012379.pdf

Bettazzi, R. (1896). Gruppi finiti ed infiniti di enti. *Accademia delle Scienze di Torino, Classe di Scienze Fisiche, Matematiche, e Naturale*, *31*, 506–512. Retrieved from https://www.biodiversitylibrary.org/item/44259#page/564/mode/1up

Blackwell, D. (1967). Infinite games and analytic sets. *Proceedings of the National Academy of Sciences*, *58*(5), 1836–1837. doi:https://doi.org/10.1073/pnas.58.5.1836

Boolos, G. (1983). The iterative conception of set. In P. Benacerraf & H. Putnam (Eds.), *Philosophy of mathematics (2nd ed.)* (pp. 486–502). Cambridge University Press. (Original work published 1971).

Brouwer, L. E. J. (1975). Over de Grondslagen der Wiskunde (A. Heyting, Trans.). In, *L. E. J. Brouwer: Collected works, vol i* (pp. 11–101). North-Holland. (Original work published 1907).

Burge, T. (1993). Content preservation. *The Philosophical Review*, *102*(4), 457–488. doi:https://doi.org/10.2307/1523046

Burge, T. (1998). Computer proof, apriori knowledge, and other minds: The sixth philosophical perspectives lecture. *Nous*, *32*(S12), 1–37. doi:https://doi.org/10.1111/0029-4624.32.s12.1

Burge, T. (2003). Perceptual entitlement. *Philosophy and Phenomenological Research*, *67*(3), 503–548. doi:https://doi.org/10.1111/j.1933-1592.2003.tb00307.x

Burge, T. (2013). Self and self-understanding. In T. Burge (Ed.), *Cognition through understanding: Self-knowledge, interlocution, reasoning, reflection. philosophical essays, volume 3* (pp. 187–226). Oxford University Press.

Buss, S. R. (1986). *Bounded arithmetic.* Bibliopolis.

Cantor, G. (1883). Über unendliche, lineare Punktmannigfaltigkeiten. V. *Mathematische Annalen*, *21*, 545–591.

Cantor, G. (1967). Letter to Dedekind. In J. V. Heijenoort (Ed.), *From Frege to Gödel: A source book in mathematical logic, 1879–1931* (pp. 113–117). Harvard University Press. (Original work published 1899).

Casullo, A. (2007). What is entitlement? *Acta Analytica*, *22*(4), 267–279. doi:https://doi.org/10.1007/s12136-007-0012-y

Cieśliński, C. (2010). Truth, conservativeness, and provability. *Mind*, *119*(474), 409–422. doi:https://doi.org/10.1093/mind/fzq034

Cieśliński, C. (2017). *The epistemic lightness of truth: Deflationism and its logic.* Cambridge University Press.

Cohen, P. (1971). Comments on the foundations of set theory. In D. Scott (Ed.), *Proceedings of symposia in pure mathematics, vol. 13, part i* (pp. 9–15). American Mathematical Society.

Coliva, A. (Ed.). (2012). *Mind, meaning and knowledge: Themes from the philosophy of Crispin Wright.* Oxford University Press.

Coliva, A. (2015). The extended rationality view extended. In *Extended rationality* (pp. 153–180). doi:https://doi.org/10.1057/9781137501899_6

Coliva, A. (2020). Against (neo-Wittgensteinian) entitlements. In P. J. Graham & N. J. L. L. Pedersen (Eds.), *Epistemic entitlement* (pp. 327–343). doi:https://doi.org/10.1093/oso/9780198713524.003.0012

Cook, R. T. (2016). Conservativeness, cardinality, and bad company. In P. A. Ebert & M. Rossberg (Eds.), *Abstractionism: Essays in philosophy of mathematics* (pp. 223–246). Oxford University Press.

Davis, M. (1964). Infinite games of perfect information. In M. Dresher, L. S. Shapley, & A. W. Tucker (Eds.), *Advances in game theory (AM-52), volume 52* (pp. 85–101). Princeton University Press.

Dean, W. (2015). Arithmetical reflection and the provability of soundness. *Philosophia Mathematica*, *23*(1), 31–64. doi:https://doi.org/10.1093/philmat/nku026

Dedekind, R. (1965). Was sind und was sollen die Zahlen? In *Was sind und was sollen die Zahlen? Stetigkeit und Irrationale Zahlen* (pp. 1–47). Vieweg+Teubner Verlag. (Original work published 1888).

Drake, F. (1974). *Set theory*. North-Holland.

Dummett, M. (1963). The philosophical significance of Gödel's theorem. In M. Dummett (Ed.), *Truth and other enigmas* (pp. 186–201). Harvard University Press.

Ebert, P. A., & Rossberg, M. (Eds.). (2016). *Abstractionism: Essays in philosophy of mathematics*. Oxford University Press.

Feferman, S. (1962). Transfinite recursive progressions of axiomatic theories. *Journal of Symbolic Logic*, *27*(3), 259–316. doi:https://doi.org/10.2307/2964649

Feferman, S. (1964). Systems of predicative analysis. *Journal of Symbolic Logic*, *29*(1), 1–30. doi:https://doi.org/10.2307/2269764

Feferman, S. (1991). Reflecting on incompleteness. *Journal of Symbolic Logic*, *56*(1), 1–49. doi:https://doi.org/10.2307/2274902

Feferman, S. (2000). Why the programs for new axioms need to be questioned. *Bulletin of Symbolic Logic*, *6*(4), 401–413.

Fenstad, J. E. (1971). The axiom of determinateness. In J. E. Fenstad (Ed.), *Proceedings of the second Scandinavian logic symposium* (pp. 41–61). North-Holland.

Field, H. (1986). The deflationary conception of truth. In G. MacDonald & C. Wright (Eds.), *Fact, science and morality: Essays on A. J. Ayer's language, truth and logic* (pp. 55–117). Blackwell.

Field, H. (1999). Deflating the conservativeness argument. *Journal of Philosophy*, *96*(10), 533–540. doi:https://doi.org/10.2307/2564613

Fischer, M. (2021). Another look at reflection. *Erkenntnis*. doi:https://doi.org/10.1007/s10670-020-00363-9

Fischer, M., Horsten, L., & Nicolai, C. (2021). Hypatia's silence: Truth, justification, and entitlement. *Noûs*, *55*(1), 62–85. doi:https://doi.org/10.1111/nous.12292

Fischer, M., Nicolai, C., & Horsten, L. (2017). Iterated reflection over full disquotational truth. *Journal of Logic and Computation*, *27*(8), 2631–2651. doi:https://doi.org/10.1093/logcom/exx023

Fraenkel, A. A. (1927). *Zehn Vorlesungen über die Grundlegung der Mengenlehre*. Teubner.

Fraenkel, A. A. (1928). *Einleitung in die Mengenlehre (3rd further expanded edn.)* Springer.

Fraenkel, A. A., Bar-Hillel, Y., & Lévy, A. (1973). *Foundations of set theory.* North-Holland.

Franzén, T. (2004). *Inexhaustibility. A non-exhaustive treatment. Lecture Notes in Logic. Association for Symbolic Logic (Vol. 16).* A. K. Peters.

Gaifman, H. (1974). Elementary embeddings of models of set-theory and certain subtheories. In T. Jech (Ed.), *Proceedings of symposia in pure mathematics, vol. 13, part ii* (pp. 33–102). American Mathematical Society.

Gödel, K. (1983a). Russell's mathematical logic. In P. Benacerraf & H. Putnam (Eds.), *Philosophy of mathematics, 2nd ed.* (pp. 444–469). Cambridge University Press. (Original work published 1944).

Gödel, K. (1983b). What is Cantor's continuum problem? In P. Benacerraf & H. Putnam (Eds.), *Philosophy of mathematics, 2nd ed.* (pp. 470–485). Cambridge University Press. (Original work published 1947).

Gödel, K. (1986). On formally undecidable propositions of Principia Mathematica and related systems i. In S. Feferman, J. W. Dawson Jr., S. C. Kleene, G. H. Moore, R. M. Solovay, & J. Van Heijenoort (Eds.), *Collected works, volume i, publications 1929–1936* (pp. 145–195). Oxford University Press. (Original work published 1931).

Graham, P. J., & Pedersen, N. J. L. L. (2020). Recent work on epistemic entitlement. *American Philosophical Quarterly*, *57*(2), 193–214. doi:https://doi.org/10.2307/48570848

Halbach, V. (2011). *Axiomatic theories of truth.* Cambridge University Press.

Halbach, V., & Horsten, L. (2006). Axiomatizing Kripke's theory of truth. *Journal of Symbolic Logic*, *71*(2), 677–712. doi:https://doi.org/10.2178/jsl/1146620166

Hallett, M. (1984). *Cantorian set theory and limitation of size*. Oxford University Press.

Hardy, G. H. (1906). The continuum and the second number class. *Proceedings of the London Mathematical Society*, *s2-4*(1), 10–17. doi:https://doi.org/10.1112/plms/s2-4.1.10

Harrington, L. (1978). Analytic determinacy and $0^{\#}$. *Journal of Symbolic Logic*, *43*(4), 685–693. doi:https://doi.org/10.2307/2273508

Hausdorff, F. (1907). Untersuchungen über Ordnungstypen. *Königlich Sächsischen Gesellschaft der Wissenschaften zu Leipzig, Math.-Phys. Klasse, Sitzungsberichte*, *59*, 84–159.

Hessenberg, G. (1906). *Grundbegriffe der Mengenlehre (vol. i)*. Vandenhoeck & Ruprecht.

Hilbert, D., & Bernays, P. (1968). *Grundlagen der Mathematik* (2nd ed.). Springer.

Horsten, L. (2021). On reflection. *The Philosophical Quarterly*, *71*(4). doi:https://doi.org/10.1093/pq/pqaa083

Horsten, L., & Leigh, G. E. (2016). Truth is simple. *Mind*, *126*(501), 195–232. doi:https://doi.org/10.1093/mind/fzv184

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

Isaacson, D. (1996). Arithmetical truth and hidden higher-order concepts. In W. D. Hart (Ed.), *The philosophy of mathematics* (pp. 203–224). Oxford University Press.

Jensen, R. (2023). *Manuscript on fine structure, inner model theory, and the core model below one woodin cardinal.* [Manuscript in preparation.] Retrieved from https://www.mathematik.hu-berlin.de/~raesch/org/jensen/pdf/book%5C_skript%5C_feb%5C_11%5C_2020.pdf

Jourdain, P. E. B. (1904). On the transfinite cardinal numbers of well-ordered aggregates. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *7*(37), 61–75. doi:https://doi.org/10.1080/14786440409463088

Jourdain, P. E. B. (1905). On transfinite cardinal numbers of the exponential form. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *9*(49), 42–56. doi:https://doi.org/10.1080/14786440509463254

Kanamori, A. (1995). The emergence of descriptive set theory. In J. Hintikka (Ed.), *From Dedekind to Gödel: Essays on the development of the foundations of mathematics* (pp. 241–262). Springer.

Kanamori, A., & Magidor, M. (1978). The evoution of large cardinal axioms in set theory. In G. H. Müller & D. S. Scott (Eds.), *Lecture notes in mathematics, vol. 669* (pp. 99–275). Springer-Verlag.

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

Ketland, J. (2005). Deflationism and the Gödel phenomena: Reply to Tennant. *Mind*, *114*(453), 75–88. doi:https://doi.org/10.1093/mind/fzi075

Ketland, J. (2010). Truth, conservativeness, and provability: Reply to Cieśliński. *Mind*, *119*(474), 423–436. doi:https://doi.org/10.1093/mind/fzq039

Keyser, C. (1905). Some outstanding problems for philosophy. *The Journal of Philosophy, Psychology and Scientific Methods*, *2*(8), 207–213.

Kleene, S. C. (1950). A symmetric form of Gödel's theorem. *Koninklijke Nederlandse Akademie van Wetenschappen: Proceedings of the Section of Sciences*, *53*, 800–802. Retrieved from https://dwc.knaw.nl/DL/publications/PU00014670.pdf

Koellner, P. (2006). On the question of absolute undecidability. *Philosophia Mathematica*, *14*(2), 153–188. doi:https://doi.org/10.1093/philmat/nkj009

Koellner, P. (2009). On reflection principles. *Annals of Pure and Applied Logic*, *157*(2–3), 206–219. doi:https://doi.org/10.1016/j.apal.2008.09.007

Koellner, P. (2014). Large cardinals and determinacy. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Spring 2014). Metaphysics Research Lab, Stanford University.

Kondo, M. (1939). Sur l'uniformisation des complemèntaires analytiques et les ensembles projectifs de la seconde classe. *Japanese Journal of Mathematics*, *15*, 197–230. doi:https://doi.org/10.4099/jjm1924.15.0_197

Kotlarski, H. (1968). Bounded induction and satisfaction classes. *Zeitschrift für Mathematische Logik und Grundlagen der Mathematik*, *32*(31–34), 531–544. doi:https://doi.org/10.1002/malq.19860323107

Kotlarski, H., Krajewski, A., & Lachlan, A. H. (1981). Construction of satisfaction classes for nonstandard models. *Canadian Mathematical Bulletin*, *24*(3), 283–293. doi:https://doi.org/10.4153/cmb-1981-045-3

Kreisel, G. (1960). La prédicativité. *Bulletin de la Societété Mathématique de France*, *88*, 371–391.

Kreisel, G., & Lévy, A. (1968). Reflection principles and their use for establishing the complexity of axiomatic systems. *Zeitschrift für Mathematische Logik und Grundlagen der Mathematik*, *14*(7–12), 97–142. doi:https://doi.org/10.1002/malq.19680140702

Kunen, K. (1970). Some applications of iterated ultrapowers in set theory. *Annals of Mathematical Logic*, *1*(2), 179–227. doi:https://doi.org/10.1016/0003-4843(70)90013-6

Kuratowski, C. (1936). Sur les théorèmes de séparation dans la théorie des ensembles. *Fundamenta Mathematicae*, *26*, 183–191. doi:https://doi.org/10.4064/fm-26-1-183-191

Leigh, G. (2015). Conservativity for theories of compositional truth via cut-elimination. *The Journal of Symbolic Logic*, *80*(3), 845–865. doi:https://doi.org/10.1017/jsl.2015.27

Łełyk, M., & Nicolai, C. (2022). A theory of implicit commitment. *Synthese*. doi:https://doi.org/10.1007/s11229-022-03601-5

Łełyk, M., & Wcisło, B. (2017). Models of weak theories of truth. *Archive for Mathematical Logic*, *56*(5–6), 453–474. doi:https://doi.org/10.1007/s00153-017-0531-1

Levi, B. (1902). Intorno alla teoria degli aggregati. *Istituto Lombardo di Scienze e Lettere, Rendiconti*, *35*(2), 863–868.

Lévy, A. (1979). *Basic set theory*. Springer.

Lindström, P. (1997). *Aspects of incompleteness*. Cambridge University Press.

Lusin, N. (1927). Sur les ensembles analytiques. *Fundamenta Mathematicae*, *10*(1), 1–95. doi:https://doi.org/10.4064/fm-10-1-1-95

Lusin, N. (1930). Sur le problème de M. J. Hadamard d'uniformisation des ensembles. *Comptes Rendus Acad. Sci. Paris*, *190*(1), 349–351. Retrieved from https://gallica.bnf.fr/ark:/12148/bpt6k3143v/f351.item#

Lusin, N., & Novikoff, P. (1935). Choix effective d'un point dans un complémentaire analytique arbitraire donné par un crible. *Fundamenta Mathematicae*, *25*, 559–560. doi:https://doi.org/10.4064/fm-25-1-559-560

Maddy, P. (1988a). Believing the axioms. i. *Journal of Symbolic Logic*, *53*(2), 481–511. doi:https://doi.org/10.1017/s0022481200028425

Maddy, P. (1988b). Believing the axioms. ii. *Journal of Symbolic Logic*, *53*(3), 736–764. doi:https://doi.org/10.2307/2274569

Maddy, P. (1990). *Realism in mathematics*. Clarendon Press.

Maddy, P. (1997). *Naturalism in mathematics*. Oxford University Press.

Maddy, P. (2011). *Defending the axioms*. Oxford University Press.

Maddy, P. (2016). *What do philosophers do?: Skepticism and the practice of philosophy*. Oxford University Press.

Madison, B., & Waxman, D. (2021). Stable and unstable theories of truth and syntax. *Mind*, *130*(518), 439–473. doi:https://doi.org/10.1093/mind/fzaa034

Majors, B. (2015). What entitlement is. *Acta Analytica*, *30*(4), 363–387. doi:https://doi.org/10.1007/s12136-015-0252-1

Mansfield, R. (1970). Perfect subsets of definable sets of real numbers. *Pacific Journal of Mathematics*, *35*(2), 451–457. doi:https://doi.org/10.2140/pjm.1970.35.451

Marshall, M. V. (1989). Higher order reflection principles. *Journal of Symbolic Logic*, *54*(2), 474–489. doi:https://doi.org/10.2307/2274862

Martin, D. A. (1968). The axiom of determinateness and reduction principles in the analytical hierarchy. *Bulletin of the American Mathematical Society*, *74*(4), 687–689. doi:https://doi/org/10.1090/s0002-9904-1968-11995-0

Martin, D. A. (1970). Measurable cardinals and analytic games. *Fundamenta Mathematicae*, *66*(3), 287–291. doi:https://doi.org/10.4064/fm-66-3-287-291

Martin, D. A. (1976). Hilbert's first problem: The continuum hypothesis. In F. E. Browder (Ed.), *Proceedings of symposia in pure mathematics, vol. 28* (pp. 81–92). American Mathematical Society.

Martin, D. A. (1977). Descriptive set theory. In J. Barwise (Ed.), *Handbook of mathematical logic* (pp. 783–815). North-Holland.

Martin, D. A. (2012). Projective sets and cardinal numbers: Some questions related to the continuum problem. In A. S. Kechris, B. Löwe, & J. R. Steel (Eds.), *Wadge degrees and projective ordinals: The Cabal seminar, volume ii* (pp. 484–508). Cambridge University Press.

Martin, D. A., & Steel, J. R. (1983). The extent of scales in $l(\searrow)$. In A. S. Kechris, B. Löwe, & J. R. Steel (Eds.), *Games, scales, and Suslin cardi-*

nals: The Cabal seminar, volume i (pp. 110–120). Cambridge University Press.

Martin, D. A., & Steel, J. R. (1989). A proof of projective determinacy. *Journal of the American Mathematical Society*, *2*(1), 71–125. doi:https://doi.org/10.1090/s0894-0347-1989-0955605-x

Mauldin, R. D. (Ed.). (1981). *The Scottish book*. Birkhauser.

McCallum, R. (2021). Intrinsic justifications for large-cardinal axioms. *Philosophia Mathematica*, *29*(2), 195–213. doi:https://doi.org/10.1093/philmat/nkaa038

McGee, V. (1997). How we learn mathematical language. *The Philosophical Review*, *106*(1), 35–68. doi:https://doi.org/10.2307/2998341

Meadows, T. (2021). Did Descartes make a diagonal argument? *Journal of Philosophical Logic*, *51*(2), 219–247. doi:https://doi.org/10.1007/s10992-021-09620-w

Mirimanoff, D. (1917). Les antinomies de Russell et de Burali-Forti et le problème fondamental de la théorie des ensembles. *L'Enseignement Mathématique*, *19*, 37–52.

Moore, G. H. (1982). *Zermelo's axioms of choice*. Springer-Verlag.

Moretti, L. (2021). Problems for wright's entitlement theory. In L. Moretti & N. J. L. L. Pederson (Eds.), *Non-evidentialist epistemology* (pp. 121–138). doi:https://doi.org/10.1163/9789004465534_007

Moschovakis, Y. N. (2009). *Descriptive set theory (2nd. ed.)* American Mathematical Society.

Mostowski, A. (1952). *Sentences undecidable in formalized arithmetic: An exposition of the theory of Kurt Gödel*. North-Holland.

Müller, S., Schindler, R., & Woodin, W. H. (2020). Mice with finitely many woodin cardinals from optimal determinacy hypotheses. *Journal of Mathematical Logic*, *20*(Supp01), 1950013. doi:https://doi.org/10.1142/s0219061319500132

Mycielski, J., & Świerczkowski, S. (1964). On the Lebesgue measurability and the axiom of determinateness. *Fundamenta Mathematicae*, *54*(1), 67–71. doi:https://doi.org/10.4064/fm-54-1-67-71

Neeman, I. (2002). Inner models in the region of a Woodin limit of Woodin cardinals. *Annals of Pure and Applied Logic*, *116*(1–3), 67–155. doi:https://doi.org/10.1016/s0168-0072(01)00103-8

Nelson, E. (1986). *Predicative arithmetic*. Princeton University Press.

Neta, R. (2009). Mature human knowledge as a standing in the space of reasons. *Philosophical Topics*, *37*(1), 115–132. doi:https://doi.org/10.5840/philtopics200937119

Nicolai, C., & Piazza, M. (2019). The implicit commitment of arithmetical theories and its semantic core. *Erkenntnis*, *84*(4), 913–937. doi:https://doi.org/10.1007/s10670-018-9987-6

Novikoff, P. (1931). Sur les fonctions implicites mesurables B. *Fundamenta Mathematicae*, *17*, 8–25. doi:https://doi.org/10.4064/fm-17-1-8-25

Novikoff, P. (1935). Sur la séparabilité des ensembles projectifs de seconde classe. *Fundamenta Mathematicae*, *25*, 459–466. doi:https://doi.org/10.4064/fm-25-1-459-466

Oxtoby, J. C. (1957). The Banach-Mazur game and Banach category theorem. In M. Dresher, A. W. Tucker, & P. Wolfe (Eds.), *Contributions to the theory of games (AM-39), volume iii* (pp. 159–163). Princeton University Press.

Parsons, C. (1983). Sets and modality. In C. Parsons (Ed.), (2019) *Mathematics in philosophy: Selected essays* (pp. 298–342). Cornell University Press.

Peano, G. (1890). Démonstration de l'intégrabilité des équations différentielles ordinaires. *Mathematische Annalen*, *37*, 182–228. Retrieved from https://eudml.org/doc/157517

Peano, G. (1902). Confronto col formulario. *Rivista di matematica*, *8*, 7–11. Retrieved from https://babel.hathitrust.org/cgi/pt?id=mdp.39015033442792%5C&view=1up%5C&seq=9

Pedersen, N. J. L. L., & Rossberg, M. (2010). Open-endedness, schemas and ontological commitment. *Noûs*, *44*(2), 329–339. doi:https://doi.org/10.1111/j.1468-0068.2010.00742.x

Pederson, N. J. L. L. (2009). Entitlement, value and rationality. *Synthese*, *171*(3), 443–457. doi:https://doi.org/10.1007/s11229-008-9330-x

Poincaré, H. (1906). Les mathématiques et la logique. *Revue de métaphysique et de morale*, *14*(3), 294–317.

Potter, M. (2004). *Set theory and its philosophy: A critical introduction*. Oxford University Press.

Quine, W. V. O. (1969). *Set theory and its logic*. Harvard University Press.

Reinhardt, W. N. (1974). Remarks on reflection principles, large cardinals and elementary embeddings. In T. Jech (Ed.), *Proceedings of symposia in pure mathematics, vol. 13, part ii* (pp. 189–205). American Mathematical Society.

Roberts, S. (2017). A strong reflection principle. *The Review of Symbolic Logic*, *10*(4), 651–662. doi:https://doi.org/10.1017/s1755020317000223

Rowbottom, F. (1971). Some strong axioms of infinity incompatible with the axiom of constructibility. *Annals of Mathematical Logic*, *3*(1), 1–43. doi:https://doi.org/10.1016/0003-4843(71)90009-x

Russell, B. (1907). On some difficulties in the theory of transfinite numbers and order types. *Proceedings of the London Mathematical Society*, *s2-4*(1), 29–53. doi:https://doi.org/10.1112/plms/s2-4.1.29

Schindler, R., & Steel, J. R. (2014). *The core model induction.* Retrieved from https://ivv5hpp.uni-muenster.de/u/rds/core_model_induction.pdf

Schütte, K. (1965a). Eine Grenze für die Beweisbarkeit der Transfiniten Induktion in der verzweigten Typenlogik. *Archiv für Mathematische Logik und Grundlagen-forschung*, *7*(1–2), 45–60. doi:https://doi.org/10.1007/bf01972460

Schütte, K. (1965b). Predicative well-orderings. In J. N. Crossley & M. Dummett (Eds.), *Formal systems and recursive functions* (pp. 280–303). North-Holland.

Scott, D. (1961). Measurable cardinals and constructible sets. *Bulletin de l'Académie Polonaise des Sciences. Série des Sciences Mathématiques, Astronomiques et Physiques*, *7*, 145–149.

Scott, D. (1974). Axiomatizing set theory. In T. Jech (Ed.), *Proceedings of symposia in pure mathematics, vol. 13, part ii* (pp. 207–214). American Mathematical Society.

Scott, D. (1977). Foreword. In J. L. Bell (Ed.), *Boolean-valued models and independence proofs in set theory*. Clarendon Press.

Shapiro, S. (1998). Proof and truth: Through thick and thin. *The Journal of Philosophy*, *95*(10), 493–521. doi:https://doi.org/10.2307/2564719

Shapiro, S., & Uzquiano, G. (2016). Ineffability within the limits of abstraction alone. In P. A. Ebert & M. Rossberg (Eds.), *Abstractionism: Essays in philosophy of mathematics* (pp. 283–307). Oxford University Press.

Shoenfield, J. R. (1967). *Mathematical logic*. Addison-Wesley.

Shoenfield, J. R. (1977). Axioms of set theory. In J. Barwise (Ed.), *Handbook of mathematical logic* (pp. 321–344). North-Holland.

Sierpinski, W. (1930). Sur l'uniformisation des ensembles measurables (B). *fuma*, *16*, 136–139. doi:https://doi.org/10.4064/fm-16-1-136-139

Silins, N. (2012). Explaining perceptual entitlement. *Erkenntnis*, *76*, 243–261. doi:https://doi.org/10.1007/s10670-011-9304-0

Silver, J. (1966). *Some applications of model theory in set theory* (Doctoral dissertation, University of California, Berkeley).

Simpson, S. G. (2009). *Subsystems of second order arithmetic*. Cambridge University Press.

Skolem, T. (1929). Über die Grundlagendiskussionen in der Mathematik. In J. E. Fenstad (Ed.), (1970) *Thoralf Skolem: Selected works in logic* (pp. 207–225). Universitetsforlaget.

Skolem, T. (1967). The foundations of elementary arithmetic established by means of the recursive mode of thought, without the use of apparent variables ranging over infinite domains. In J. V. Heijenoort (Ed.), *From Frege to Gödel: A source book in mathematical logic, 1879–1931* (pp. 303–333). Harvard University Press. (Original work published 1923).

Solovay, R. M. (1967). A nonconstructible Δ^1_3 set of integers. *Transactions of the American Mathematical Society*, *127*(1), 50–75. doi:https://doi.org/10.1090/s0002-9947-1967-0211873-5

Solovay, R. M. (1969). On the cardinality of Σ^1_2 sets of reals. In J. J. Bulloff, T. C. Holyoke, & S. W. Hahn (Eds.), *Foundations of mathematics: Symposium papers commemorating the sixtieth birthday of Kurt Gödel* (pp. 58–73). doi:https://doi.org/10.1007/978-3-642-86745-3_7

Solovay, R. M. (1970). A model of set-theory in which every set of reals is Lebesgue measurable. *Annals of Mathematics*, *92*(1), 1–56. doi:https://doi.org/10.2307/1970696

Solovay, R. M., Reinhardt, W. N., & Kanamori, A. (1978). Strong axioms of infinity and elementary embeddings. *Annals of Mathematical Logic*, *13*(1), 73–116. doi:https://doi.org/10.1016/0003-4843(78)90031-1

Steel, J. R. (2000). Mathematics needs new axioms. *Bulletin of Symbolic Logic*, *6*(4), 422–433.

Steel, J. R. (2005). PFA implies ADL(R). *Journal of Symbolic Logic*, *70*(4), 1255–1296. doi:https://doi.org/10.2178/jsl/1129642125

Tait, W. W. (1981). Finitism. *The Journal of Philosophy*, *78*(9), 524–546. doi:https://doi.org/10.2307/2026089

Tait, W. W. (2001). Gödel's unpublished papers on the foundations of mathematics. *Philosophia Mathematica*, *9*(1), 87–126. doi:https://doi.org/10.1093/philmat/9.1.87

Tait, W. W. (2005). Constructing cardinals from below. In W. W. Tait (Ed.), *The provenance of pure reason: Essays in the philosophy of mathematics and its history*. Oxford University Press.

Tennant, N. (2002). Deflationism and the Gödel phenomena. *Mind*, *111*(443), 551–582. doi:https://doi.org/10.1093/mind/111.443.551

Tennant, N. (2005). Deflationism and the Gödel phenomena: Reply to Ketland. *Mind*, *114*(453), 89–96. doi:https://doi.org/10.1093/mind/fzi089

Tiles, M. (2004). *The philosophy of set theory: An historical introduction to Cantor's paradise*. Courier Corporation.

Turing, A. (1939). Systems of logic based on ordinals. *Proceedings of the London Mathematical Society*, *S2-45*, 161–228. doi:https://doi.org/10.1112/plms/s2-45.1.161

Wang, H. (1963). *A survey of mathematical logic*. North Holland.

Wang, H. (1983). The concept of set. In P. Benacerraf & H. Putnam (Eds.), *Philosophy of mathematics, 2nd ed.* (pp. 530–570). Cambridge University Press. (Original work published 1974).

Wcisło, B., & Łełyk, M. (2017). Notes on bounded induction for the compositional truth predicate. *The Review of Symbolic Logic*, *10*(3), 455–480. doi:https://doi.org/10.1017/s1755020316000368

Weir, A. (2003). Neo-Fregeanism: An embarrassment of riches. *Notre Dame Journal of Formal Logic*, *44*(1). doi:https://doi.org/10.1305/ndjfl/1082637613

Welch, P. D. (2015). Large cardinals, inner models, and determinacy: An introductory overview. *Notre Dame Journal of Formal Logic*, *56*(1). doi:https://doi.org/10.1215/00294527-2835083

Welch, P. D. (2017). Obtaining Woodin's cardinals. In A. Caicedo, J. Cummings, P. Koellner, & P. Larson (Eds.), *Logic in Harvard: Conference celebrating the birthday of Hugh Woodin* (pp. 161–176). AMS Series, Contemporary Mathematics, vol. 690.

Weyl, H. (1949). *Philosophy of mathematics and natural science*. Princeton University Press.

Woodin, W. H. (1988). Supercompact cardinals, sets of reals, and weakly homogeneous trees. *Proceedings of the National Academy of Sciences*, *85*(18), 6587–6591. doi:https://doi.org/10.1073/pnas.85.18.6587

Woodin, W. H. (2005). The continuum hypothesis. In R. Cori, A. Razborov, S. Todorčević, & C. Wood (Eds.), *Logic Colloquium 2000, volume 19 of Lecture Notes in Logic* (pp. 143–197). Association of Symbolic Logic.

Woodin, W. H. (2017). In search of ultimate-L: The 19th midrasha mathematicae lectures. *The Bulletin of Symbolic Logic*, *23*(1), 1–109. doi:https://doi.org/10.1017/bsl.2016.34

Wright, C. (1965). Facts and certainty. *Proceedings of the British Academy*, *71*, 429–472.

Wright, C. (1994). About “the philosophical significance of Gödel’s theorem”: Some issues. In B. McGuinness & G. Oliveri (Eds.), *The philosophy of Michael Dummett* (pp. 167–202). Kluwer.

Wright, C. (2002). (Anti-)sceptics simple and subtle: G.E. Moore and John McDowell. *Philosophy and Phenomenological Research*, *65*(2), 330–348. doi:https://doi.org/10.1111/j.1933-1592.2002.tb00205.x

Wright, C. (2004). Warrant for nothing (and foundations for free)? *Aristotelian Society Supplementary Volume*, *78*(1), 167–212. doi:https://doi.org/10.1111/j.0309-7013.2004.00121.x

Wright, C. (2005). Intuition, entitlement and the epistemology of logical laws. *Dialectica*, *58*(1), 155–175. doi:https://doi.org/10.1111/j.1746-8361.2004.tb00295.x

Wright, C. (2012). Replies part iv: Warrant transmission and entitlement. In A. Coliva (Ed.), *Mind, meaning and knowledge: Themes from the philosophy of Crispin Wright* (pp. 451–486). Oxford University Press.

Zermelo, E. (1967a). A new proof of the possibility of a well-ordering. In J. V. Heijenoort (Ed.), *From Frege to Gödel: A source book in mathematical logic, 1879–1931* (pp. 183–198). Harvard University Press. (Original work published 1908).

Zermelo, E. (1967b). Proof that every set can be well-ordered. In J. V. Heijenoort (Ed.), *From Frege to Gödel: A source book in mathematical logic, 1879–1931* (pp. 139–141). Harvard University Press. (Original work published 1904).

Index