FAKULTÄT
FÜR !NFORMATIK

Faculty of Informatics

Axiomatic Truth Theories and Reflection Principles

DIPLOMA THESIS

submitted in partial fulfillment of the requirements for the degree of

Diplom-Ingenieur

in

Logic and Computation

by

Arnaud De Coster, MSc
Registration Number 11704379

to the Faculty of Informatics

at the TU Wien

Advisor: Ao.Univ.Prof. Dipl.-Ing. Dr.techn. Christian Fermüller

Vienna, 25th January, 2021

Arnaud De Coster | Christian Fermüller

Erklärung zur Verfassung der Arbeit

Arnaud De Coster, MSc
Tongasse 4, 1030 Vienna

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Wien, 25. Jänner 2021

Arnaud De Coster

Acknowledgements

This thesis would never have seen the light of day if it were not for the many people who supported and guided me during my work on it. I can not thank my advisor Professor Chris Fermüller enough for the thorough and wide-ranging discussions we had. Your comments and (gentle) push-back on the drafts of this thesis have improved the presentation and content of the final work in every respect. I also want to thank Dr. Johannes Hafner for his eagle-eyed proofreading of the thesis, as well as the enriching discussions and pointers to literature he provided. Thanks are also due to Professor Georg Schiemer for suggesting the topic to me when I entered his office on a whim, looking for something that was on the boundary between philosophy and logic. Axiomatic truth theories turned out to be the perfect topic for delving into both!

I want to express my gratitude to the friends I've made in Vienna over the years who helped make my time here so pleasant. In particular, I want to thank every single member of the Vienna Forum for Analytic Philosophy for having provided such a warm and intellectually stimulating environment to this logician interloper. Additionally, I want to thank Michael and Tobias for proofreading and editing the German version of the abstract.

To my parents, my thanks for having always supported me to study whatever I found interesting, even when that interest was alien to them, and to my siblings for letting me talk about truth to them for unreasonable lengths of time without complaint. Finally, my thanks to you Brankica, for your ever-present support and love.

Kurzfassung

Der Begriff der Wahrheit nimmt in der Logik einen merkwürdigen Platz ein. Einerseits ist er zentral: Um die Vollständigkeit und Korrektheit eines deduktiven Systems zu zeigen muss bewiesen sein, dass alles Wahre mit allem Beweisbaren kongruent ist. Andererseits ist Wahrheit fast immer ein metatheoretischer Begriff, der in besagten deduktiven Systemen nicht vorhanden ist. Aus der Arbeit Tarskis wissen wir, dass es aufgrund des Lügner-Paradoxons unmöglich ist, das Konzept der Wahrheit direkt zu internalisieren. Axiomatische Wahrheitstheorien sind ein Versuch, dieses Paradoxon mittels ein primitives Wahrheitsprädikat, und geeigneter Axiome zu umgehen.

Das zentrale Ziel dieser Arbeit ist es, eine abgeschlossene Einführung in den Wahrheitsdeflationismus (die philosophische Position, dass der Begriff der Wahrheit nur eine untergeordnete linguistische Rolle zu erfüllen hat) und die ihm zugrunde liegenden axiomatischen Wahrheitstheorien zu geben. Insbesondere soll die Konservativitätsdebatte anhand der axiomatischen Wahrheitstheorien **TB** und **CT** vorgestellt und illustriert werden. Kritiker des Wahrheitsdeflationismus haben darauf hingewiesen, dass eine axiomatische Wahrheitstheorie nicht sowohl konservativ sein kann – also kein neues mathematisches Wissen liefert (wie **TB**) – als auch wahrheitstheoretisches Folgern validiert (wie **CT**).

In jüngerer Vergangenheit wurde eine implizite Akzeptanz bestimmter Reflexionsprinzipien, die (teilweise) die Korrektheit eines deduktiven Systems ausdrücken, als Antwort für den Wahrheitsdeflationismus vorgeschlagen. Demzufolge kann die Akzeptanz eines Reflexionsprinzips im Rahmen einer konservative Wahrheitstheorie wie **TB** zu einem stärkeren System wie **CT** führen. Die verschiedenen technischen Ergebnisse werden vorgestellt und die philosophische Motivation für diese implizite Akzeptanz bewertet. Es wird argumentiert, dass Reflexionsprinzipien akzeptabel und rechtfertigbar sind, ohne auf das starke Konzept der Wahrheit zurückzugreifen das sie für den Wahrheitsdeflationismus begründen sollen. Dies versetzt den Wahrheitsdeflationismus in die Lage, seiner Kritik zu widerstehen: Das Konzept der Wahrheit kann sowohl konservativ, in einem geeigneten Sinn, als auch linguistisch nützlich sein.

Abstract

The concept of truth occupies a curious place in logic. On the one hand it is central; showing completeness and correctness for a given deductive system is a matter of showing that what is true lines up with what is provable. On the other hand truth is almost always a meta-theoretical notion, which is not present within the deductive systems considered. Ever since the work of Tarski, we know that straightforwardly internalizing the concept of truth is impossible, due to the Liar paradox. Axiomatic theories of truth are an attempt at circumventing this paradox, by introducing a primitive truth predicate, and suitable axioms governing it.

The chief aim of this thesis is to give a self-contained introduction to truth deflationism, the philosophical position that the concept of truth has only a minor linguistic role to fulfill, and the axiomatic truth theories that underlie it. In particular, an account of the conservativity debate is given, as exemplified by the axiomatic truth theories **TB** and **CT**. Critics of truth deflationism have pointed out that an axiomatic truth theory can't be both conservative, roughly meaning that it does not provide genuinely new mathematical knowledge (like **TB**), and validate truth-theoretical reasoning (like **CT**).

Recently, a reply for the deflationist has been proposed that rests on an implicit commitment to certain reflection principles, which (partially) express the soundness of a deductive system. If one is committed to a reflection principle for a conservative truth theory, like **TB**, then one can obtain stronger systems like **CT**. The different technical results are presented, and the philosophical motivation for this implicit commitment is evaluated. It is argued that reflection principles are indeed acceptable, and crucially, can be justified without recourse to the strong concept of truth which they are supposed to recover for the truth deflationist. This leaves the truth deflationist in a position to withstand the criticism: the concept of truth can be conservative, in an appropriate sense, and linguistically useful too.

Contents

CHAPTER 1

Introduction

.

> Truth can never be told so as to be understood, and not be believ'd.
>
> William Blake, *Proverbs of Hell*

Truth has been entwined with western philosophy from its historical conception in the distant mists of the Presocratics onwards. In the beautiful philosophical poem of Parmenides, we find a conception of truth which has been hugely influential up to the present day. A young man has been carried away on a chariot by the daughters of the Sun, to meet a goddess who will guide him to *aletheia*, which is translatable as both truth and reality. This ambiguity is present at the heart of the poem itself. At his arrival, the goddess promises him that he will "Learn all things", not as a set of delivered statements, but by showing him how to think through them himself:

> But come now, I will tell you – and you, when you have
> heard the story, bring it safely away –
> which are the only routes of inquiry that are for thinking:
> the one, that is and that it is not possible for it not to be,
> is the path of Persuasion (for it attends upon Truth [aletheia]),
> the other, that it is not and that it is right that it not be,
> this indeed I declare to you to be a path entirely unable to
>   be investigated:
> For neither can you know what is not (for it is not to be
>   accomplished)
> nor can you declare it. [CM11, Fragment B2, p.57-58] [. . .] for thinking and
> being are the same. [CM11, Fragment B3, p.58]

The fragment itself is rife with hermeneutical challenges, with competing readings centred on the correct understanding of the verb ‘to be’ (esti) [Kim18, p.3]. It does however point to truth being a distinctively *ontologico-epistemic* concept [Sza18, p.16]. Truth on this understanding is not a property of sentences, but rather refers to true reality. It is an epistemic notion because of its link to that which can come to be known, and an ontological notion, because it identifies truth with that which is real. Note that reality is understood very differently from how it is now commonly understood; for Parmenides, and later Plato, that which is real is opposed to the phenomenal world of deceiving appearances in which we live. Reality (i.e. Platonic forms) is that which we apprehend through reason (*logos*), as opposed to the sensible objects of our daily life.

Skipping ahead a great deal, we find the most prominent modern incarnation of truth as a substantial notion in correspondence theory. Correspondence theory holds that truth is a property of truthbearers, in virtue of their relationship to truthmakers. Exactly what the truthbearers are depends on the flavour of correspondence theory considered. Paradigmatically, they are sentences or propositions. A truthmaker on the other hand is what makes some truthbearer true. These are also understood differently, but will often be states of affairs, facts, or objects. A classical formulation of correspondence theory is given by Russell: “Thus a belief is true when there is a corresponding fact, and is false when there is no corresponding fact” [Rus99, p. 129]. Some form of a correspondence theory of truth is today still the majority view of philosophers [BC14], surely in part due to its immediate obviousness. The main objection against it is its vagueness: how should one understand the correspondence relation, and what is the nature of facts? Attempting to dispel this vagueness often leads to unwieldy ontological commitments, like the existence of disjunctive facts [Hor11, p.13] or the Big Fact, which makes every true sentence true [Dav69].

Around the same time as Russell and Moore set out the classical formulations of correspondence theory, we find the seeds of a completely different view. Frege was likely the first to point out that when truth is taken to be primarily an element of *language*, it is much less of an impressive concept:

> It is worth noticing that the sentence “I smell the scent of violets”, surely has the same meaning as the sentence “It is true that I smell the scent of violets”. So it seems that by my ascription of the property of truth to it, nothing is thereby added to the thought itself.[1][FP03, p.34]

This statement is an example of what later came to be known as a *deflationary* approach to truth. In contrast to the substantial concept of truth of correspondence theory, truth can be understood as an insubstantial concept, which adds very little, if anything, to our discourse. It is this approach which underlies the subject of this thesis. Truth

[1] Beachtenswert ist es auch, daß der Satz “Ich rieche Veilchenduft” doch wohl denselben Inhalt hat wie der Satz “es ist wahr, daß ich Veilchenduft rieche”. So scheint denn dem Gedanken dadurch nichts hinzugefügt zu werden, daß ich ihm die Eigenschaft der Wahrheit beilege. (Own translation)

deflationism as a doctrine of truth is hard to pin down completely. However, virtually all explications of truth deflationism in one way or other do justice to Frege's remark by including the so-called T-schema :

T-schema: For all sentences (propositions) φ : "φ'' is true iff φ.

One reason truth deflationism can not be easily summarized is that for most deflationists, the T-schema does not give a definition of truth, but only partially characterizes its use. The disagreement on *what else* truth is for is the reason that several positions can be found under the banner of deflationism. F.P. Ramsey was the first to argue that the truth predicate, through the T-schema, can be seen to be redundant [Ram27]. A.J. Ayer, full of the positivist spirit of the day, went on to claim that truth is not even a genuine concept:

> [On the T-schema] And this shows that the words 'true' and 'false' are not used to stand for anything, but function in the sentence merely as assertion and negation signs. That is to say, truth and falsehood are not genuine concepts. Consequently there can be no logical problem concerning the nature of truth. [Aye35]

A watershed moment in the analysis of the concept of truth was Tarski's seminal 1933 paper [Tar33]. Tarski realized that, lacking a formal *theory* of the concept of truth, some issues can simply not be understood clearly. In the paper, we find the undefinability theorem which roughly states that, within a language, no truth predicate can be defined which fulfills the T-schema. He showed this by considering the liar sentence which asserts its own falsehood, and roughly takes the form of "This sentence is not true." Tarski took this limiting result as a sign that constructing a universally valid definition of truth was out of the question, and instead went on to study it in the context of formal theories, where a strict separation between the meta-language (containing the concept of truth), and the object language (the domain we are interested in) can be maintained. Although it would be wrong to claim that Tarski himself saw his analysis as a deflationary one, his approach of considering *formal theories of truth* would become a mainstay in the different explications of truth deflationism. Similar to Tarski's separation of the object language and meta-language (a distinction going back to at least Hilbert) for analyzing the concept of truth, truth deflationists nowadays often study axiomatic theories of truth in order to get a grip on the precise consequences of one's deflationist tenets. Specifically, the *base theory* of arithmetic is usually taken as the domain of interest on the level of language, with different formulations of the truth-theory over arithmetic on the level of the meta-language. Arithmetic has several nice properties which makes it a useful sand-box to put one's deflationary truth theory to the test, not the least of which is that it has a ready arsenal of 'obvious' truths.

Jumping ahead once again, to the end of the 1990's, a flurry of new work on formal truth theories brought the question of the uses of truth in sharp relief. Horwich defended a minimalist theory of truth, where the T-schema is taken to fully explicate the deflationary concept of truth [Hor98]. But he explicitly did not take truth to be a redundant concept. Following Quine, he argued for its usefulness in expressing *generalizations.* We recognize that in classical logic each sentence is either true or false. Without the concept of truth, this can be expressed for each sentence individually. I can assert that "Snow is white or snow is not white" without explicit recourse to truth. Generalizing this pattern to each possible assertion *does* require the truth predicate, by putting it as "Every sentence S of the form 'p or not p' is true." As the decade came to its close, it became clear that these instrumental uses of truth were less innocent than presumed. The axiomatic theory encapsulating the T-schema, from now on referred to as **TB**, turned out to be too weak to actually derive the generalizations touted as important to truth deflationism. A stronger truth-theory, which we will call **CT**, does derive these generalizations and so is prima facie in a better position to be understood as the correct formalization. However, this deductive strength comes at a price. As pointed out by both Shapiro and Ketland, **CT** is not *conservative* over the base theory of arithmetic, meaning that it proves some arithmetical statements that are not provable within arithmetic themselves [Sha98][Ket99]. This result was put forward as an explicit challenge to the deflationist. It seems hard to reconcile the notion that truth is a light concept, with at most some linguistic uses, if one's truth theory is able to prove new results over the domain it is formulated over.

It is this *impasse* which is central to the thesis. We require the deductive strength of a theory like **CT** in order to defend the – essentially inferential – uses we put truth to in our language. Simultaneously, we need a reply to the challenge that such a truth theory fails to be deflationist since it allows us to derive new statements, previously out of reach. Here the second strand of the thesis comes in. We know by the Gödel incompleteness theorems that for a sufficiently expressive system, like arithmetic, there are certain true statements that are unprovable within arithmetic. In particular, the system is unable to prove its own 'consistency', where the concept of consistency is translated into the language of arithmetic. As Turing already observed in his PhD thesis, this suggests a method of completing a given formal system: Simply add the consistency statement as an axiom [Tur36]. Of course, the resulting formal system is incomplete too, and so can be completed further by adding the corresponding consistency statement, and so on. Traditionally, a reflection principle is understood as statement which expresses the soundness of a given formal system, that is, formalizing that if a system proves φ, it holds that φ is true. It is straightforward to show that the addition of a reflection principle implies the consistency of a system, so that the addition of reflection principles are a natural way of generating ever more complete systems.

Historically, reflection principles have been used to unfold the full implications of a given formal theory. The idea is that if one trusts a given formal system, that is, believes the axioms and inference rules to be true, adding a reflection principle as an axiom amounts to making this implicit trust explicit. The new system thus obtained will then be a better approximation of the full implications of the initial formal theory. Two well-known foundational projects that have been approached in this way are Kreisel's analysis of finitism [Kre58], and Feferman's analysis of predicativism [Fef64]. Recently, a similar approach has been suggested to face the conservativity challenge to truth deflationism [HL17][Cie10][FNH17]. By starting with an acceptable, but deductively deficient conservative truth theory, one can reach stronger, non-conservative truth theories through the addition of a reflection principle. The lack of conservativity of the truth theories we put forward is then not a phenomenon intrinsic to truth, but a result of the gradual unfolding of our conservative concept of truth. This unfolding through reflection itself is not to be motivated by the way of a prior concept of truth, on pain of avoiding circularity. If this can be done, and this thesis argues that it can, this provides the truth deflationist with a coherent response to the challenge that one's formal truth theory ought to be deductively fertile, and yet conservative too.

This thesis will not further address the different positions sheltered underneath the umbrella of truth deflationism, since we hone in rapidly on the conservativity challenge. At the expense of breadth, this makes it possible to be more explicit about the assumptions underlying this work. First, we will consider only axiomatic truth theories, where the base theory is some arithmetical theory. The purpose of these truth theories is to give an *implicit* definition of a truth predicate T. Truth is on this understanding a property of sentences only, identified with the extension of the T predicate. A sentence is true if and only if the truth theory **Th** proves $T(\ulcorner\varphi\urcorner)$, where $\ulcorner\varphi\urcorner$ is the 'name' of the sentence. The purpose of truth is a merely linguistic one, it makes certain inferences easier, and allows us to express generalizations that are impossible to express otherwise. It is, after Horsten, a *logico-linguistic* notion [Hor11, p.65-66], rather than the substantial notion found in correspondence theories of truth. Because other concepts of truth will also be encountered in the thesis, we summarize the different meanings of 'true' we will encounter. This disambiguation will be especially important when we come to the analysis of the defense of reflection principles. There are (at least) four different concepts of truth present in this thesis:

1. **Truth with capital 'T'**: A non-deflationary understanding of truth, ranging from the visions of the mystic to truth as a property of truthbearers by virtue of a direct relation to the world, as in correspondence theory.

2. **Truth in a truth-theory**: A logico-linguistic notion. We understand truth as the T-predicate in a formalized theory, such as **CT** or **TB**.

3. **Arithmetical truth**: Truth as a property of sentences in virtue of the sentences holding for the natural numbers. Well-known to be unformalizable in first-order arithmetic.

4. **Truth as validity**: Validity of a sentence in a theory means that the sentence is satisfied in all models of the theory.

The axiomatic truth theories we will study, chief among them **TB** and **CT**, are theories which uphold a strict separation between the level of the base-theory and the meta-level of the truth theory. Such a truth theory will only be applicable in specific domains, rather than for natural language as a whole, since natural language is universal, that is, every concept ought to be translatable in natural language. Hence, no meta-theory can be available for natural language. Truth theories which do not make this separation, so-called type-free theories, exist and are well-studied, for example the theories **FS** and **KF** discussed in [Hal01a, Part 3]. Although the conservativity challenge is just as relevant for theories like **KF** and **FS**, our concern is not with completeness, but with evaluating the proposed use of reflection principles to face it. For this reason we will not look at any type-free theories in this thesis.

We now address some worries the reader might have about the T-schema, and truth deflationism in general. At first glance, the T-schema seems to simply kick the can down the road. It is very well to hold that “snow is white” is true if and only if snow is white, but what does the right-hand side in this bi-conditional stand for? For the deflationist, it certainly can’t be the fact that snow is white, since this is just a version of correspondence theory. The deflationary point of view is that the truth predicate has only a linguistic role to play, and that the right setting to understand truth is a language. So, we have to understand the T-schema as saying that one *asserts* that “snow is white” is true if and only if one (could come to) asserts that snow is white. In this sense, truth is parasitical on what one could come to assert. That is all well and good, but how can we come to assert that snow is white without first knowing it to be true? Presumably, I will come to assert that snow is white only if I came to know it (and am not the lying sort). How I come to know this is an epistemological question, and can be answered without recourse to truth. Looking outside of my window on a wintry morning and trusting in the reliability of my senses might be sufficient. What truth deflationism attempts to offer are the laws of truth in *general*. Coming to know *particular* truths in a particular domain will require theory and knowledge appropriate to that domain.

The aim of this thesis is to give a unified exposition of the technical, historical, and philosophical background (much of which is scattered in the literature) necessary to understand this recent interplay between axiomatic truth theories and reflection principles. First, in Chapter 2, we go over the incompleteness theorems, and some classic results on reflection principles. Since the truth predicate is a syntactical notion, some intuition to what can and cannot be expressed is useful. Hence, we devote more time than is usual to the details of coding syntactical concepts in an arithmetical theory. Next, Chapter 3 gives a broad historical overview of the foundationalist projects in mathematics of Hilbert, Kreisel, and Feferman, and the role of reflection principles therein. Only the analysis of finitism in Section 3.3 will be necessary later on, so the reader is free to skim the chapter if they like. In Chapter 4 we go over Tarski’s undefinability theorem, which is the paradoxical soil from which all truth theories spring, and introduce the two truth theories central to this thesis: **TB** and **CT**. The differences in deductive strength of **TB** and **CT** are considered, and compared to the desirable uses of the truth predicate according to truth deflationists. We will see that **TB** is deficient in this regard, while **CT** is capable of formalizing the desirable truth-theoretic reasoning. The conservativity challenge is discussed in detail in Chapter 5. We show that **TB** is syntactically conservative over arithmetic, and **CT** is not. This gives rise to the *impasse* that is central to this thesis: truth theories cannot be conservative over the base theory, *and* validate desirable truth-theoretic reasoning. Chapter 6 presents the different technical results on adjoining reflection principles over truth-theories, as a response to the conservativity challenge. In Chapter 7 we take a look at the literature surrounding the *implicit commitment thesis* (ICT), which roughly states that we are committed to additional, independent reflection principles by accepting a theory. It will be argued that in accepting a conservative truth theory, we are justified in accepting additional reflection principles over the theory, and so justified in accepting a non-conservative truth theory as deflationists.

As mentioned before, many of the results are scattered in the literature, since the application of reflection principles to axiomatic truth theories is still very recent. For convenience, Table 1.1 summarizes these results and convenient sources.

Topic	Source	Description
Truth theory books	[Hor11]	An easy-going introduction to the topic which is low on formality.
	[Hal01a]	More expansive and formal introduction to the field. Contains most complete proof of non-conservativity of **CT**.
	[Cie17]	Specialized monograph with recent results on conservativity and reflection principles.
Conservativity challenge	[Sha98]	Classic paper that put forward the conservativity challenge for truth deflationists.
	[Ket99]	Contemporaneous paper which posed the conservativity challenge.
Reflection principles	[Tur36]	Turing's PhD thesis develops ordinal logics, which are a precursor to theories obtained through iterated reflection.
	[Kre58]	Kreisel uses reflection over **PRA** to unfold the concept of finitist proof.
	[Fef64]	Feferman unfolds the concept of predicative mathematics using a kind of generalized uniform reflection.
	[KL68]	The original source of the results relating induction to reflection.
Reflection principles and truth	[Cie10]	The first explicit defense of using reflection principles to face the conservativity challenge. Considers uniform reflection over the theory $\mathbf{CT}^{-}$.
	[HL17]	Considers uniform reflection over the truth theories **TB** and **PTB**.
	[FNH17]	Studies reflection over a type-free truth theory formulated in a partial logic.
Implicit commitment thesis (ICT)	[Dea14]	Dean's paper gives a nice overview of the historical background of reflection principles, as well as criticizes ICT by the way of certain epistemically stable theories.
	[NP19]	The authors give a defense of ICT, where ICT is understood as being minimally committed to a conservative truth-theory over the accepted theory.
	[Hor]	A phenomenological analysis of the process of reflection over an accepted theory.

Table 1.1: Summary and description of the main sources for the topics covered.

A Computer Scientist's Apology

This thesis has been written as part of studies in computer science. Some readers might demur at seeing formal truth theories being studied in this context. Should all this truth business not be relegated to the philosophers wholesale? To a certain extent I concur. As far as I know, there are no immediate applications of the work on formal truth theories to computer science. Taken a slightly wider view however, formal truth theories connect many themes within applied logic and computer science. There is of course the centrality of truth as a semantic concept in logic. As Frege wrote:

> [Logic] stands in a similar relation to truth as physics to weight or heat. Discovering truths is the task of all sciences: it falls to logic to identify the laws of truth.[2] [FP03, p.30]

From this point of view, studying the basic laws of truth in the form of formal truth theories is foundational to anything else we might use logic for, even if the insights so gained are not immediately useful.

Formal truth theories are also a great case study of what I consider to be a recurrent theme in applying formal logic. The difficulty faced in applied logic is often not obtaining the technical results (although also often far from simple), but determining whether the formalization of a given situation itself cuts any ice. Modelling the functioning of the truth predicate in language is in that sense no different from modelling epistemic agents, defeasible reasoning, or bounded rationality. Formal truth theories can only enrich our understanding of the trade-offs one inevitably has to make when modelling things as slippery as natural language or rationality.

Finally, formal truth theories illustrate another common phenomenon in applied logic: the balancing-act between expressivity and tractability. The principal reason that higher-order logic is avoided in applications is that its consequence relation is not effective, that is, the theorems of the theory cannot be determined computationally. This phenomenon can be found time and time again in different guises in computer science. Different modal logics are used throughout the field, for example in knowledge representation, because they are more expressive than propositional logic, but tractable in a way that first-order logic is not. What we sacrifice in expressivity compared to first-order logic is gained on the side of the (polynomial) complexity of the algorithms used in automated deduction. Another example of this trade-off is Gentzen's famous cut-elimination theorem for sequent calculi, acccording to which proofs with cuts can be transformed into proofs without cuts. Informally, a cut corresponds to the usage of a lemma in a proof. Cut-elimination is very useful in automated deduction, since the correct choice of lemma to use is difficult to automate. The price one pays is that a proof without cuts can, in the worst case, be

[2] [Logik] verhält sich zur Wahrheit etwa so wie die Physik zur Schwere oder Wärme. Wahrheiten entdecken, ist Aufgabe aller Wissenschaften: der Logik kommt es zu, die Gesetze des Wahrseins zu erkennen. (Own translation)

non-elementarily longer than the original proof. A similar result can be found for truth theories. Fischer has shown that a *conservative* truth theory over arithmetic, essentially a variant of **CT**, is able to reduce proof-length in a non-elementary way over proofs in arithmetic [Fis14]. This gives mathematical content to the oft repeated claim that the truth predicate has an expressive function: it makes our proofs shorter, while not doing work that could not be done without it.

It is my hope that the reader is of the sort that is seduced foremost by the prospect of getting a better grip on the notion of truth, as I was. Failing that, I think they will at least find that the topics explored in this thesis will connect with the more traditional topics of computer science in a way which is fruitful, even if not directly applicable.

CHAPTER 2

Limits of Our Language

The title of this chapter, while somewhat arch, refers to the many different incompleteness results known of sufficiently expressive systems. These results stem from a common cause, leading several authors to speak of the incompleteness phenomenon instead. One of the settings in which the phenomenon occurs is when one attempts to speak about truth within the language of the system. In contrast to the more well-known Gödel incompleteness theorems, which are concerned with *syntactical* notions, truth is a prime example of a *semantical* notion. The purpose of this chapter is to build up the apparatus necessary to prove the incompleteness theorems, as well as introduce the reflection principles, and show how these relate.

2.1 Peano Arithmetic

We are interested in first-order Peano arithmetic since it offers enough expressivity to code sentences, thus allowing us to 'speak of' the language within the language. This is crucial to implement the T-schema, where 'is true' is understood as newly introduced predicate. There are in fact weaker systems of arithmetic, in the sense that they prove less theorems, for which this coding can still be implemented. Nevertheless, we stick with PA, since it is the most well-known and used system of arithmetic, and the results are usually easy to 'translate' to different systems. The exposition that follows is based on [Kay91] and [HP98]. First we define our first-order language.

As is standard, relations, functions, and variables will be represented by non-logical symbols, where variables will be denoted by $x, y, z, u, v, w, \dots$ On occasion we will use $\bar{x}$ to denote a tuple of free variables $\langle x_1, ..., x_n \rangle$. Our first-order language will also contain the following logical symbols:

- Boolean connectives: $\wedge, \vee, \neg$;

- Equality: $=$;
- Quantifiers: $\forall$, $\exists$;
- Brackets: ( ).

Aside from the the usual brackets, we will often use the following brackets where it improves legibility: $[\,], \{\,\}$. The terms of our first-order language are elements of the minimal set inductively defined by the following rules:

1. All variables and constant symbols are terms.
2. If $t_1, ..., t_n$ are terms, and f is a function symbol of arity n, then $f(t_1, ..., t_n)$ is a term.

The well-formed formulas, from now on simply referred to as formulas, of the language are then elements of the minimal set inductively defined by:

1. If $t_1, ..., t_n$ are terms and P is a predicate symbol of arity n, then $P(t_1, ..., t_n)$ is a formula;
2. If t_1 and t_2 are terms, then $t_1 = t_2$ is a formula;
3. If φ is a formula, then $\neg\varphi$ is a formula;
4. If φ and ψ are both formulas, then $(\varphi \wedge \psi)$ is a formula, and $(\varphi \vee \psi)$ is a formula;
5. If φ and ψ are both formulas, then $(\varphi = \psi)$ is a formula;
6. If $\varphi(x)$ is a formula and x a variable, then $\forall x\varphi(x)$ and $\exists x\varphi(x)$ are both formulas.

We also make use of the implication, equivalence, and 'there exists a unique' symbols with the following shorthands:

- $(\varphi \rightarrow \psi)$ for $(\neg\varphi \vee \psi)$;
- $(\varphi \leftrightarrow \psi)$ for $[(\varphi \rightarrow \psi) \wedge (\psi \rightarrow \varphi)]$;
- $\exists!x\, \psi(x)$ for $\exists x[\varphi(x) \wedge \forall y(\varphi(y) \rightarrow \varphi(x))]$.

We will use the following shorthand notation to denote quantification over sequences of variables:

- $\forall\bar{x}\varphi(\bar{x})$ for $\forall x_1, \ldots, \forall x_n\varphi(x_1, \ldots, x_n)$, where $\bar{x} = \langle x_1, \ldots, x_n\rangle$;
- $\exists\bar{x}\varphi(\bar{x})$ for $\exists x_1, \ldots, \exists x_n\varphi(x_1, \ldots, x_n)$, where $\bar{x} = \langle x_1, \ldots, x_n\rangle$.

From now on, we will follow the convention that the outermost brackets in a formula are not shown. Finally, the language of Peano arithmetic, denoted by L_{PA} , is extended from the first-order language with the following non-logical symbols $+, \cdot, 0, S$, with the intended (yet to be axiomatized) meaning of addition, multiplication, zero, and the successor function respectively. For the functions of addition and multiplication we will use the more familiar infix notation rather than the usual prefix notation.

We now introduce the *theory* **PA**. There are two ways to think of a (first-order) theory. The most general way is to declare a theory as simply being a set of sentences. This definition has the advantage that it includes theories which are not recursively axiomatizable, meaning that there is no algorithm which decides for a given sentence whether it is a member of the theory or not. The classic example of this is the theory of true arithmetic, the set of first-order sentences which are true of the natural numbers. This result follows from the incompleteness theorems which are the subject of this chapter. The second way to think of a logical theory is as a deductive system. A deductive system consists of recursive set of axioms, and inference rules which allow one to derive new sentences belonging to the theory from the axioms. So, one can think of a theory as either being a set of sentences, or a core of axioms and inference rules which derive precisely this set of sentences. Usually we will take a theory to be a deductive system, but it is good to keep in mind that this does not cover all cases (true arithmetic in particular).

In the case of Peano arithmetic, aside from the axioms of first-order logic with identity, and the inference rules of first-order logic, the theory of Peano arithmetic (**PA**) is axiomatized by:

PA-1 $\neg\exists x\, S(x) = 0$

PA-2 $\forall x, y\, S(x) = S(y) \to x = y$

PA-3 $\forall x\, x + 0 = x$

PA-4 $\forall x, y\, x + S(y) = S(x + y)$

PA-5 $\forall x\, x \cdot 0 = 0$

PA-6 $\forall x, y\, x \cdot S(y) = (x \cdot y) + x$

PA-7 For all $\varphi(x) \in L_{PA}$: $\varphi(0) \wedge \forall x\, [\varphi(x) \to \varphi(S(x))] \to \forall x \varphi(x)$

The last element of the list is in fact an axiom schema, capturing the principle of mathematical induction.

We will often find ourselves discussing extensions of Peano arithmetic, which include in addition to the axioms of **PA** more sentences, perhaps in a richer language. These arithmetical theories will be denoted by the generic **Th**, with L_{Th} denoting the corresponding language. In addition, we will require these extended theories to be recursively axiomatizable. In particular, **PA** is recursively axiomatizable. When two theories $\mathbf{Th_1}$

and $\mathbf{Th_2}$ are deductively equivalent, meaning that they derive the same set of sentences, we will write this as $\mathbf{Th_1} \equiv \mathbf{Th_2}$.

Aside from the symbol '0', no numerals are present explicitly in **PA**. Other numerals are usually defined in the following way:

Definition 2.1.1. A numeral is a term of L_{PA} of the form '$S(\dots S(0)\dots)$'. If the number of successor symbols prefixed to '0' is n, the numeral will be denoted by $\underline{n}$.

We will also need the 'smaller than' relation in what follows.

Definition 2.1.2. The $<$ relation is defined in **PA** (using infix notation) as $x < y := \exists z(x + z = y \land x \neq y)$.

The semantics for first-order logic are assumed to be known but bear repeating. Given a language L_{Th}, an L_{Th}-structure $\mathcal{I}$ is a triple $\langle U, I, \alpha\rangle$, where U is a non-empty set, called the domain, I is the interpretation function, and α is a variable assignment. The interpretation function I assigns for each each n-ary predicate symbol P of L_{Th} an n-ary relation $I(P) \subseteq U^n$, each n-ary function symbol f an n-ary function: $I(f) : U^n \to U$, and for each constant symbol c an element $I(c) \in U$. The equality symbol '=' will always be interpreted by the identity relation $\{\langle a, a\rangle \mid a \in U\}$. Finally, the variable assignment is a function which assigns to each variable x in L_{Th} an element $\alpha(x) \in U$. We will be loose with notation and write $a \in \mathcal{I}$, when meaning $a \in U$.

Definition 2.1.3. Given a structure $\mathcal{I} = \langle U, I, \alpha\rangle$, the evaluation of a term t is denoted by $t^{\mathcal{I}}$ and defined inductively as:

- $t^{\mathcal{I}} = I(c)$ if t is a constant c;
- $t^{\mathcal{I}} = \alpha(x)$ if t is a variable x;
- $t^{\mathcal{I}} = f({t_1}^{\mathcal{I}}, \dots, {t_n}^{\mathcal{I}})$ if t is given by $I(f)(t_1, \dots, t_n)$.

Definition 2.1.4. The following inductive conditions relate a structure $\mathcal{I} = \langle U, I, \alpha\rangle$ and a formula φ, where $\mathcal{I} \models \varphi$ is understood as $\mathcal{I}$ *modelling* φ:

- If φ is an atomic formula, that is, of the form $P(t_1, \dots, t_n)$, then $\mathcal{I} \models \varphi$ iff $\langle {t_1}^{\mathcal{I}}, \dots, {t_n}^{\mathcal{I}}\rangle \in I(P)$.
- $\mathcal{I} \models \neg\varphi$ iff $\mathcal{I} \not\models \varphi$.
- $\mathcal{I} \models (\varphi \land \psi)$ iff $\mathcal{I} \models \varphi$ and $\mathcal{I} \models \psi$.
- $\mathcal{I} \models (\varphi \lor \psi)$ iff $\mathcal{I} \models \varphi$ or $\mathcal{I} \models \psi$.
- $\mathcal{I} \models \forall x\varphi(x)$ iff $\mathcal{I}' \models \varphi$ for each $\mathcal{I}' = \langle U, I, \alpha \cup \{x \leftarrow u\}\rangle$, with $u \in U$. Here, $\alpha' := \alpha \cup \{x \leftarrow u\}$ is the mapping for which holds that $\alpha'(y) = \alpha(y)$ for each variable $y \neq x$ in L_{Th}, and $\alpha'(x) = u$.

A structure $\mathcal{I}$ which models each axiom in a theory **Th** will be named a model of **Th**, and denoted by M. A formula φ is *true* in M iff $M \models \varphi$. A formula φ is valid iff it is true in every model M of **Th**. In the case of Peano arithmetic, we will single out the structure of the natural numbers $\mathbb{N}$ as the standard model. When the claim is made that φ is true with no model M specified, it will be tacitly understood to be true in the standard model, which is also called arithmetical truth. Note that the standard model is given by the *structure* $\mathbb{N}$, rather than the set of natural numbers $\mathcal{N}$. On occasion, we will abuse notation by writing $M \models \varphi(a)$, where a is an object in the domain of M, rather than $M \models \varphi(x)$ under the variable assignment $\alpha(x) = a$. This abuse of notation will extend to any formula in which objects of the domain occur, so that the formula occurring in $M \models \forall x(x < c \to x + 1 \leq c)$, where $c \in M$, can be seen as *partially interpreted* in M.

Since we will be studying axiomatic truth theories, which contain a primitive truth predicate T, we will often consider *expanded models*.

Definition 2.1.5. A model M is an expansion of a model K iff the only difference between M and K is that M contains new relations, functions, and/or constant elements that do not occur in M.

In this thesis, the most common expansion is a model $\langle U, +_M, \cdot_M, S_M, 0_M, T, \alpha\rangle$, abbreviated as (M, T), expanded from $M := \langle U, +_M, \cdot_M, S_M, 0_M, \alpha\rangle$, where $M \models$ **PA**.

As the name ‘standard model’ suggests, **PA** also has non-standard models that are not isomorphic to $\mathbb{N}$. Given the existence of non-standard models for **PA**, some apprehension in identifying the numeral $\underline{n}$ with $n \in \mathbb{N}$ is natural. Perhaps the numeral $\underline{n}$ could be interpreted by non-standard numbers in non-standard models? That this is not the case, and the identification of the numeral $\underline{n}$ with $n \in \mathbb{N}$ is unproblematic, is the upshot of the next result. First we need to define the model-theoretic notion of initial segments and end-extensions:

Definition 2.1.6. If N and M are L_{PA}-structures, with N a substructure of M, then N is an *initial segment* of M, or equivalently, M is an *end-extension* of N (in symbols: $N \subseteq_e M$) if and only if:

$$\text{For all } x \in N,\ \text{for all } y \in M : M \models y < x \implies y \in N.$$

The following result relates $\mathbb{N}$ and the non-standard models, and allows us to identify $\mathbb{N}$ with $\{\underline{n}^M | n \in \mathbb{N}\}$:

Theorem 2.1.1. If $M \models$ **PA**, then the map $\mathbb{N} \to M$ given by $n \to \underline{n}^M$ is an embedding of $\mathbb{N}$ onto the smallest initial segment of M.

For this reason we will be sloppy and only use the numeral notation where confusion could arise.

We will also need the Overspill lemma which, loosely speaking, states that any property which holds on an initial segment 'spills over' beyond the initial segment (see [Cie17, p. 11]). First we define the notational convenience of bounded quantification:

Definition 2.1.7. Assuming t is a term of L_{PA} not containing the variables in $\bar{x}$, we abbreviate

$$\begin{aligned}&\forall\bar{x}\ (\bar{x} < t \rightarrow \dots)\text{ as }\forall\bar{x}{<}t\ (\dots)\\&\text{and}\\&\exists\bar{x}\ (\bar{x} < t \wedge \dots)\text{ as }\exists\bar{x}{<}t\ (\dots),\end{aligned}$$

where $\bar{x} < t$ means $\bigwedge_i x_i < t$ for each $x_i \in \bar{x}$. The quantifiers are said to occur in *bounded* form.

Lemma 2.1.2 (Overspill Lemma). Let **Th** be an extension of **PA**, with induction extended to every formula of **Th**. Let M be a model of **Th** whose arithmetical reduct, meaning the model one obtains after removing all interpretations of non-arithmetical symbols, is non-standard. Let I be a proper initial segment of M closed under the successor operation, and let $\varphi(x, \bar{a})$ be a formula of the language L_{Th}, with $\bar{a}$ a tuple of parameters, i.e. objects, from M. If

$$\text{for all } b \in I : M \models \varphi(b, \bar{a}),$$

then there is an element $c \in M$ such that $c > I$ and

$$M \models \forall x \leq c\varphi(x, \bar{a}).$$

Proof. Assume that the conditions stated in the lemma hold. Assume for a contradiction that for no $c > I$ the condition $M \models \forall x \leq c\varphi(x, \bar{a})$ is satisfied. Then the formula:

$$\psi(x, \bar{a}) := \forall y < x\varphi(y, \bar{a}),$$

defines I in M. Now, since $0 \in I$, and since I is closed under the successor operation, $M \models \psi(0, \bar{a}) \wedge [\psi(x, \bar{a}) \rightarrow \psi(S(x), \bar{a})]$. Hence, by induction it holds that $M \models \forall x\psi(x, \bar{a})$, and so $I = M$, contradicting the assumption that I is a proper initial segment. □

2.2 Incompleteness Results

As often in mathematics, the various truth theories spring from an antecedent failure — the celebrated incompleteness theorems. As we will see, the concept of truth is inexpressible within Peano arithmetic, or any other theory more expressive. Given these negative results, it is easy to forget just how much can be expressed within Peano arithmetic. In fact, the limiting incompleteness theorems are valid precisely due to this expressive power. The purpose of this section is to trace the border of expressiveness, and to recapitulate the consequent incompleteness theorems. As Jean-Yves Girard has pointed out, the details can obscure what is at the heart of these theorems:

> This result, like the late paintings of Claude Monet, is easy to perceive, but from a certain distance. A close look reveals only fastidious details that one perhaps does not want to know. [Gir11, p.15]

Nevertheless, we will give a sketch of how to get the trickeries of coding right, if only because it will be good practice for when we come to truth theories, where truth predicates have to apply to coded sentences. The first step towards developing the syntactical apparatus necessary is defining the arithmetical hierarchy.

Definition 2.2.1 (**Arithmetical Hierarchy**)**.** We give an inductive definition of the arithmetical hierarchy. A formula in which all its quantifiers occur in bounded form is classified as Δ_0:

$$Q_1\bar{x_1}{<}r\ Q_2\bar{x_2}{<}s\ \dots\ Q_n\bar{x_n}{<}t\varphi(\bar{x_1},\bar{x_2},\dots,\bar{x_n},\bar{z}),$$

where $Q_i \in \{\exists, \forall\}$, $Q_i\bar{x_i}$ is possibly a sequence of quantified variables, and φ is quantifier-free. By definition, a Δ_0-formula is also classified as Σ_0 and Π_0. A Σ_{n+1}-formula is of the form:

$$\exists\bar{x}\varphi(\bar{x},\bar{z}),$$

where φ is a Π_n-formula. A Π_{n+1}-formula is of the form:

$$\forall\bar{x}\varphi(\bar{x},\bar{z}),$$

where φ is a Σ_n-formula. Finally, the Δ_n-formulas are those that are equivalent to both Σ_n- and Π_n-formulas.

Note that the definition of Δ_n depends on the underlying theory over which the equivalence is taken. Strictly speaking, not all formulas are classified in the arithmetical hierarchy. However, it is well known that every first-order formula is equivalent to some formula in prenex normal form, meaning that all quantifiers occur as a prefix to a formula which is quantifier-free. Modulo equivalence then, every formula is classified. We will also occasionally write $\varphi \in \Delta_n$ instead of saying that φ is a Δ_n-formula, and similarly for the other classes.

At the heart of the incompleteness theorems lies self-referentiality: the ability of a theory to 'speak of itself'. What is meant is the following: syntactical properties of formulas, such as being the negation of a different formula, being a numeral or a term, and even being provable within a theory, are properties we would like to have as (defined) predicates within the theory. However, formulas are not first-class citizens of our theory — predicates only apply to terms. It is for this reason that we need a *Gödel numbering* of formulas, encoding each formula (and proofs) of the language as natural numbers. A Gödel numbering is an injective function $\sigma \to \ulcorner\sigma\urcorner$ taking sequences of symbols of the language to natural numbers. For the incompleteness theorems to be proved, the coding of formulas and syntactical properties needs to be done in a 'natural' way. In particular, the coding must be computable, as well as its inverse where defined, and similar for the syntactical predicates.

As a reminder, ‘computable’ means that there is an effective method, i.e. an algorithm, to solve the problem under consideration. This means that for each input the algorithm will give the (correct) output. By the Church-Turing thesis, we can refrain from referring to an underlying model of computation when using the intuitive notion of computability. As algorithms can be seen as functions from natural numbers to natural numbers, the computable functions are exactly those functions for which an (extensionally) equivalent algorithm exists. These functions are also called recursive, and we will do so in the remainder of the thesis. Similarly, a subset S of the natural numbers will be called recursive iff there is a recursive function f of which holds that for each $x \in \mathcal{N}$: $f(x) = 1$ iff $x \in S$ and $f(x) = 0$ iff $x \notin S$. A proper introduction to these notions can for example be found in [Rog87].

While formulas and proofs can be seen as the objects of discussion in our theory by virtue of a Gödel numbering, we still need to precisify what it means to ‘express’ a syntactical property in the theory.

Definition 2.2.2. A total function $f : \mathbb{N}^k \to \mathbb{N}$ is represented by a formula $\varphi(x_1, \dots, x_k, y)$ in an arithmetical theory **Th** iff for all $\langle n_1, \dots, n_k \rangle \in \mathbb{N}^k$:

$$\begin{aligned}&\mathbf{Th} \vdash \exists! y \varphi(\langle \underline{n}_1, \dots, \underline{n}_k \rangle, y)\\ &\text{and}\\ &\text{if } l = f(\langle n_1, \dots, n_k \rangle) \text{ then } \mathbf{Th} \vdash \varphi(\langle \underline{n}_1, \dots, \underline{n}_k \rangle, \underline{l}).\end{aligned}$$

Likewise, a set $S \subseteq \mathbb{N}^k$ is represented by a formula $\varphi(x_1, \dots, x_k)$ in a theory **Th** iff for all $\bar{n} \in \mathbb{N}^k$:

$$\begin{aligned}&\text{if } \langle n_1, \dots, n_k \rangle \in S \text{ then } \mathbf{Th} \vdash \varphi(\langle \underline{n}_1, \dots, \underline{n}_k \rangle)\\ &\text{and}\\ &\text{if } \langle n_1, \dots, n_k \rangle \notin S \text{ then } \mathbf{Th} \vdash \neg\varphi(\langle \underline{n}_1, \dots, \underline{n}_k \rangle).\end{aligned}$$

The following theorem and corollary capture just how much can in fact be represented in an arithmetical theory:

Theorem 2.2.1. Let $f : \mathbb{N}^k \to \mathbb{N}$ be a recursive function. Then there is a Σ_1-formula representing f in **PA**.

Corollary 2.2.1.1. All recursive sets $A \subseteq \mathbb{N}$ are represented by a Σ_1-formula in **PA**.

The previous theorem and its corollary are the reason we will call functions f that are representable by a Σ_1-formula provably recursive.

As we will later show through the incompleteness theorems, not all true sentences, where true means ‘holds in the standard model’, are provable in **PA**. Nevertheless a restricted version does hold:

Theorem 2.2.2. Σ_1-Completeness
Let Σ_1-$\mathbb{N}$ denote $\{\varphi \mid \varphi \in \Sigma_1,\ \varphi \text{ closed},\ \mathbb{N} \models \varphi\}$. Then $\mathbf{PA} \vdash \Sigma_1$-$\mathbb{N}$.

Thanks to the Church-Turing thesis it is easy to convince oneself that many syntactic properties are indeed recursive, and can thus be used effectively as defined predicates within **PA**. It suffices to come up with an algorithm which checks whether or not the property in question holds of the code of the sentence, formula, or term in question. In particular, we will make use of the following defined predicates and functions:

- $x = Neg(y)$, iff x is the code of the expression coded by y, and prefixed with the negation symbol;
- $x = Con(y, z)$, iff x is the code of the expression which is the conjunction of the expressions coded by y and z respectively;
- $Var(x), Tm(x), Tm^c(x), Fm(x)$, iff x is the code of a variable, term, closed term, or formula respectively;
- $\underline{n} = Numeral(x)$, iff x is the code of the numeral $\underline{n}$;
- $x = Sub(z, v, t)$, iff x is the code of the formula coded by z, with v the code of the variable occurring in that expression substituted with the term coded by t.

As an example of what an explicit coding requires we will construct the formula corresponding to $Tm(x)$. First we need to be able to code sequences. Which coding device (e.g. by usage of the Chinese remainder theorem) is implemented is not important. That this coding exists, and is provably recursive within **PA**, is the subject of the following lemma:

Lemma 2.2.3. There is a Δ_0 formula $\theta(x, y, z)$, abbreviated by $(x)_y = z$, such that:

$$\mathbf{PA} \vdash \forall x, y\ \exists! z\ (x)_y = z$$

and

$$\mathbf{PA} \vdash \forall x, y, z\ \exists w\ (\forall i{<}y (w)_i = (x)_i \land (w)_y = z)$$

In effect, $(x)_y = z$ indicates that x codes the sequence $(x)_0, \dots, (x)_y, \dots$ where $(x)_y$ is given by z. It is also the case that $Len(x)$, the function from sequences to their length, is Δ_0. With sequences neatly coded, we can be bit more concrete about our Gödel numbering. Given a function ν which takes each symbol of the alphabet to a unique natural number, each string of symbols $\sigma_1\sigma_2 \dots \sigma_n$ will be coded as the unique x for which $(x)_1 = \nu(\sigma_1), (x)_2 = \nu(\sigma_2) \dots (x)_n = \nu(\sigma_n)$. We also introduce the concatenation operation $x \cap y$, which returns the code of the sequence z representing the concatenation of the sequences coded by x and y respectively. That is, if x codes the sequence $(x)_0, \dots, (x)_n$, and y codes $(y)_0, \dots, (y)_m$, then $x \cap y$ codes the sequence $(x)_0, \dots, (x)_n, (y)_0, \dots, (y)_m$.

Using the concatenation operation, new (provably recursive) functions can be introduced. For example, $\ulcorner(\urcorner \cap \ulcorner x\urcorner \cap \ulcorner+\urcorner \cap \ulcorner y\urcorner \cap \ulcorner)\urcorner$ is the function which returns the code of the expression which consists of the left bracket symbol, the expression coded by x, the addition symbol, the expression coded by y, and the right bracket symbol, in that order. It represents a function of arity 2, with arguments x, and y. We will abbreviate it as $\ulcorner(x+y)\urcorner$, and similarly for other functions so introduced. All this being in place, the formula for $Tm(x)$ is given by:

$$Tm(x) = \exists s\; Termseq(s \cap (x)), \text{ where}$$

$$\begin{aligned} Termseq(s) = \forall i < Len(s) :\; & \\ & (s)_i = \ulcorner 0\urcorner \lor \\ & \exists j{\leq}i\ (s)_i = \ulcorner v_j\urcorner \lor \\ & \exists j{<}i\ (s)_i = \ulcorner S((s)_j)\urcorner \lor \\ & \exists j,k{<}i\ (s)_i = \ulcorner((s)_j + (s)_k)\urcorner \lor \\ & \exists j,k{<}i\ (s)_i = \ulcorner((s)_j \cdot (s)_k)\urcorner. \end{aligned}$$

The ‘trick’ behind the formula is to encode the compositionality of our language by having subsequent elements of the sequence be the composition of previous elements. An expression is a term if it ‘fits’ at the end of such a sequence.

The following lemma, which comes in both syntactic and semantic variants, is key to the proofs of the incompleteness theorems, but will also play an important role in our discussion of truth theories. Proofs of both can be found in [Smi13, Chapter 24].

Lemma 2.2.4. Syntactic diagonal lemma For each formula $\varphi(x)$ of a theory **Th**, extending **PA** and possibly introducing new non-logical vocabulary, there is a sentence ψ of such that $\mathbf{Th} \vdash \psi \leftrightarrow \varphi(\ulcorner\psi\urcorner)$.

Lemma 2.2.5. Semantic diagonal lemma For each formula $\varphi(x)$ in a language L_{Th}, which includes L_{PA}, there is a sentence ψ of L_{Th} such that $\psi \leftrightarrow \varphi(\ulcorner\psi\urcorner)$ is true (in the standard model).

For the incompleteness theorems, but also for defining the reflection principles, we will need a provability predicate, representing the property of a sentence being provable. The idea is to mimic the external notion of provability for a theory by coding proofs, which are nothing but strings of formulas that follow by inference rules from axioms. In the case of **PA** for example, if we mimic a Hilbert-style proof system, where the only inference rules are given by modus ponens and generalization, the notion of a proof in Peano arithmetic, $Pr_{PA}(x,y)$, standing for ‘x codes a proof of y’ in **PA**, is represented (without proof) by:

$$\begin{aligned} Pr_{PA}(x,y) = Seq(y) \land End(y) = x \land & \\ \forall i < len(y)[& LogAx((y)_i) \lor EqAx((y)_i) \lor Ax_{PA}((y)_i) \\ & \lor \exists j,k{<}i\ MP((y)_j,(y)_k,(y)_i) \\ & \lor \exists j{<}i\ Gen((y)_j,(y)_i)]. \end{aligned}$$

In this definition we have made use of the following -definable in **PA**- predicates and functions:

- $Seq(y)$, representing that y codes a sequence;
- $End(y)$, the function returning the last element of y;
- $LogAx(y)$, the predicate representing that y codes a logical axiom of the theory;
- $EqAx(y)$, the predicate representing that y codes an equality axiom;
- $Ax_{PA}(y)$, the predicate representing that y codes one of the axioms proper to Peano arithmetic;
- $MP(x, y, z)$: the predicate representing that z is the code of the formula resulting by modus ponens from the formulas coded by x and y;
- $Gen(x, y)$: the predicate representing that y is the code of a formula following from the formula coded by x through generalization.

The one-place provability predicate $Pr_{PA}(y)$, representing that y is provable in **PA**, is then defined as $\exists x Pr_{PA}(x, y)$. It is possible to construct $Pr_{PA}(x, y)$ in such a way that it is Δ_0, so that $Pr_{PA}(y)$ is in Σ_1.

The next theorem captures that $Pr_{PA}(x)$ is a faithful translation of the external notion of provability. We first define ω-consistency.

Definition 2.2.3. A recursively axiomatizable theory **Th** extending **PA** is ω-inconsistent if $\mathbf{Th} \vdash \varphi(\underline{n})$ for each $\underline{n}$, but $\mathbf{Th} \vdash \exists x \neg \varphi(x)$. A recursively axiomatizable theory **Th** is ω-consistent if it is not ω-inconsistent.

Theorem 2.2.6. For any sentence φ:

1. If $\mathbf{PA} \vdash \varphi$, then $\mathbf{PA} \vdash Pr_{PA}(\ulcorner \varphi \urcorner)$.
2. Assume that **PA** is ω-consistent. If $\mathbf{PA} \vdash Pr_{PA}(\ulcorner \varphi \urcorner)$, then $\mathbf{PA} \vdash \varphi$.

Proof. For the first claim, assume that $\mathbf{PA} \vdash \varphi$. Then there is a proof d with code $\ulcorner d \urcorner$, such that d is a proof of φ with code $\ulcorner \varphi \urcorner$. Since the relation 'x is proof of y in **PA**' is represented by $Pr_{PA}(x, y)$, it follows that $\mathbf{PA} \vdash Pr_{PA}(\ulcorner d \urcorner, \ulcorner \varphi \urcorner)$. Clearly, $\mathbb{N} \models Pr_{PA}(\ulcorner d \urcorner, \ulcorner \varphi \urcorner)$, hence $\mathbb{N} \models \exists x Pr_{PA}(x, \ulcorner \varphi \urcorner)$, and so by definition, $\mathbb{N} \models Pr_{PA}(\ulcorner \varphi \urcorner)$. Since Pr_{PA} is Σ_1, it follows by Theorem 2.2.2 that $\mathbf{PA} \vdash Pr_{PA}(\ulcorner \varphi \urcorner)$.

For the second claim, assume that $\mathbf{PA} \vdash Pr_{PA}(\ulcorner \varphi \urcorner)$, and that **PA** is ω-consistent. For a contradiction, assume that $\mathbf{PA} \nvdash \varphi$. Then, for all n, it is not the case that n codes a proof of φ. Since the relation 'x is a proof of y' is represented by $Pr_{PA}(x, y)$, it follows that $\mathbf{PA} \vdash \neg Pr_{PA}(n, \ulcorner \varphi \urcorner)$, for each n. Since by assumption $\mathbf{PA} \vdash \exists x Pr_{PA}(x, \ulcorner \varphi \urcorner)$, PA is ω-inconsistent after all. □

In fact, the theorem can be weakened. All that is needed is that **PA** be Σ_1-sound, meaning that if **PA** $\vdash \varphi$, where φ is a Σ_1-sentence, then $\mathbb{N} \models \varphi$.

The fact that **PA** $\vdash \varphi$ implies **PA** $\vdash Pr_{PA}(\ulcorner\varphi\urcorner)$ is sufficient to prove the (strengthened) first incompleteness theorem, known as the Gödel-Rosser incompleteness theorem. For the second incompleteness theorem, we will need the provability predicate to fulfill two additional conditions. Together, these conditions are known as the Hilbert-Bernays-Löb (HBL) derivability conditions.

Definition 2.2.4. Let Th be an axiomatizable theory extending PA. the HBL derivability conditions for a provability predicate $Pr_{Th}(x)$, standing for 'x is provable in Th', are given by:

HBL-1 For arbitrary $\varphi \in L_{Th}$: If $Th \vdash \varphi$ then $PA \vdash Pr_{Th}(\ulcorner\varphi\urcorner)$,

HBL-2 For arbitrary $\varphi, \psi \in L_{Th}$: $PA \vdash Pr_{Th}(\ulcorner\varphi\urcorner) \wedge Pr_{Th}(\ulcorner\varphi \to \psi\urcorner)$
$\to Pr_{Th}(\ulcorner\psi\urcorner)$,

HBL-3 For arbitrary $\varphi \in L_{Th}$: $PA \vdash Pr_{Th}(\ulcorner\varphi\urcorner) \to Pr_{Th}(\ulcorner Pr_{Th}(\varphi)\urcorner)$.

The first HBL condition was proved for **PA** in Theorem 2.2.6, and one can prove that the second and third HBL conditions hold (although this is more difficult, see [Smi13, Chapter 26] for the proofs). We will later also need the formalized counter-part of Lemma 2.2.2:

Theorem 2.2.7. Let Th be an axiomatizable extension of PA.

1. For any Σ_1-sentence φ of L_{Th}, $PA \vdash \varphi \to Pr_{Th}(\ulcorner\varphi\urcorner)$.
2. For any Σ_1-formula $\varphi(\bar{x})$ of L_{Th}, $PA \vdash \varphi(\bar{x}) \to Pr_{Th}(\ulcorner\varphi(\dot{\bar{x}})\urcorner)$, with $\bar{x}$ the only free variables in φ.

Before we state and prove the Gödel-Rosser incompleteness theorem, we define the Rosser provability predicate. It has the peculiar property that it is *extensionally* correct (i.e. satisfies the first HBL-condition and its converse) but is *intensionally* incorrect, that is, does not express our external notion of provability (see [Fef60] for a discussion of this issue).

Definition 2.2.5. Given a provability predicate $Pr_{Th}(x, y)$ which fulfills HBL-1, we define the two-place Rosser provability predicate $Pr^R_{Th}(x, y)$ as:

$$Pr^R_{Th}(x, y) := Pr_{Th}(x, y) \wedge \forall z{<}x \neg Pr_{Th}(z, Neg(y)).$$

The one-place Rosser provability predicate $Pr^R_{Th}(y)$ is then given by $\exists x Pr^R_{Th}(x, y)$.

Note that this provability predicate is Σ_1, given the construction of $Pr_{Th}(x, y)$ as Δ_0, since both negation and bounded quantification do not increase the complexity. We can understand this provability predicate as saying that there is a proof witnessing φ such that there is no smaller witness to the contrary.

We are now in a position to prove the Gödel-Rosser incompleteness theorem, also known as the first incompleteness theorem.

Theorem 2.2.8 (**Gödel-Rosser incompleteness theorem**)**.** Let **Th** be a consistent, recursively axiomatizable theory extending **PA**. Then there exists a Σ_1-sentence φ (called the Rosser sentence) such that neither $\mathbf{Th} \vdash \varphi$ nor $\mathbf{Th} \vdash \neg\varphi$.

Proof. By the diagonal lemma we instantiate the fixed point φ:

$$\mathbf{PA} \vdash \varphi \leftrightarrow Pr^R_{Th}(\ulcorner \neg\varphi \urcorner).$$

There are two cases to be considered.

For the first case assume $\mathbf{Th} \vdash \varphi$. Then also $\mathbf{Th} \vdash Pr^R_{Th}(\underline{\mathrm{m}}, \ulcorner \varphi \urcorner)$ for some numeral $\underline{\mathrm{m}}$, where m codes a derivation of φ, by the first HBL-condition and Σ_1-completeness. By the definition of φ we also have that:

$$\mathbf{PA} \vdash \exists x[Pr^R_{Th}(x, \ulcorner \neg\varphi \urcorner) \land \forall z{<}x \neg Pr^R_{Th}(z, \ulcorner \varphi \urcorner)].$$

Reasoning within **PA** we have that for such an x witnessing $\exists x Pr^R_{Th}(x, \ulcorner \neg\varphi \urcorner)$ either $x \leq \underline{\mathrm{m}}$ or $\underline{\mathrm{m}} \leq x$. Since $\forall z{<}x \neg Pr^R_{Th}(z, \ulcorner \varphi \urcorner)$ holds, it must be the case that $x \leq \bar{m}$, which implies that:

$$\mathbf{Th} \vdash \exists x{\leq}\bar{m} Pr^R_{Th}(x, \ulcorner \neg\varphi \urcorner)$$

On the other hand, by assumption **Th** is consistent, $\mathbf{Th} \nvdash \neg\varphi$ and so also

$$\forall x{\leq}\bar{m} \neg Pr^R_{Th}(x, \ulcorner \neg\varphi \urcorner)$$

holds. But then it is provable in **PA** and we obtain a contradiction in **Th**.

Assume for the second case that $\mathbf{Th} \vdash \neg\varphi$. By definition of φ we have:

$$\mathbf{Th} \vdash \forall x[Pr^R_{Th}(x, \ulcorner \neg\varphi \urcorner) \to \exists z{<}x Pr^R_{Th}(z, \ulcorner \varphi \urcorner)]$$

Similarly as before, there is a code $\underline{\mathrm{m}}$ of a derivation of $\neg\varphi$ such that $\mathbf{Th} \vdash Pr^R_{Th}(\underline{\mathrm{m}}, \ulcorner \neg\varphi \urcorner)$. But then it follows that $\mathbf{Th} \vdash \exists z{<}\bar{m} Pr^R_{Th}(z, \ulcorner \varphi \urcorner)$. Since by assumption **Th** is consistent, $\mathbf{Th} \nvdash \varphi$ and so we have that

$$\forall z{\leq}\underline{\mathrm{m}} \neg Pr^R_{Th}(z, \ulcorner \varphi \urcorner)$$

holds, from which a contradiction follows in **Th**. □

Note that the Gödel-Rosser incompleteness theorem also gives us the means to see that the Rosser sentence φ, which is of Σ_1 complexity, is in fact false (in the standard model). We know by Σ_1-completeness that if the Rosser sentence were true, then $\mathbf{PA} \vdash \varphi$. So φ is false. This means that $\mathbb{N} \models \neg\varphi$. But since $\neg\varphi$ is of Π_1-complexity, this cannot be proved in **PA**.

The second incompleteness theorem concerns the consistency of a theory, and only holds if this provability predicate is a standard one, meaning that it fulfills all three HBL derivability conditions.

Definition 2.2.6. Let **Th** be an axiomatizable theory extending **PA** and let Pr_{Th} be a (not necessarily standard) provability predicate. The consistency statement provided by Pr_{Th} is the formula $\neg Pr_{Th}(\ulcorner 0 = 1 \urcorner)$, which we denote as $Con(Pr_{Th})$.

The choice of $0 = 1$ as contradiction is arbitrary if the provability predicate is standard, as the following lemma shows:

Lemma 2.2.9. Let **Th** be an axiomatizable theory extending **PA** and Pr_{Th} be a standard provability predicate. Then, for any sentence φ we have that: $\mathbf{Th} \vdash Con(Pr_{Th}) \leftrightarrow \neg Pr_{Th}(\ulcorner \varphi \urcorner) \vee \neg Pr_{Th}(\ulcorner \neg\varphi \urcorner)$.

Proof. For the direction from right to left: Since $\mathbf{Th} \vdash \neg 0 = 1$, by HBL-1 we also have $\mathbf{Th} \vdash Pr_{Th}(\ulcorner \neg 0 = 1 \urcorner)$ and:

$$\mathbf{Th} \vdash \neg Con(Pr_{Th}) \to Pr_{Th}(\ulcorner 0 = 1 \urcorner) \wedge Pr_{Th}(\ulcorner \neg 0 = 1 \urcorner).$$

By *ex falso quodlibet* and by HBL-1:

$$\mathbf{Th} \vdash Pr_{Th}(\ulcorner 0 = 1 \to (\neg 0 = 1 \to \varphi)\urcorner)$$

By two applications of HBL-2 we derive:

$$\mathbf{Th} \vdash \neg Con(Pr_{Th}) \to Pr_{Th}(\ulcorner \varphi \urcorner),$$

and similarly,

$$\mathbf{Th} \vdash \neg Con(Pr_{Th}) \to Pr_{Th}(\ulcorner \neg\varphi \urcorner).$$

Combining these two statements and contraposition obtains the result. As for the converse, we obtain by application of HBL-1:

$$\mathbf{Th} \vdash Pr_{Th}(\ulcorner \varphi \to (\neg\varphi \to 0 = 1)\urcorner).$$

By two applications of HBL-3:

$$\mathbf{Th} \vdash (Pr_{Th}(\ulcorner \varphi \urcorner) \wedge Pr_{Th}(\ulcorner \neg\varphi \urcorner)) \to Con(Pr_{Th}).$$

Contraposition gives the result. □

We are now in a position to prove the second incompleteness theorem:

Theorem 2.2.10 (**Second incompleteness theorem**)**.** Let **Th** by an axiomatizable, consistent extension of **PA**. Then $\mathbf{Th} \nvdash Con(Pr_{Th})$, where Pr_{Th} is a standard provability predicate.

Proof. We instantiate the diagonal lemma with the Gödel sentence:

$$\varphi \leftrightarrow \neg Pr_{Th}(\ulcorner \varphi \urcorner)$$

We show that $\mathbf{Th} \vdash \varphi \leftrightarrow Con(Pr_{Th})$, from which the result follows by the first incompleteness theorem. For the direction from left to right, notice that

$$\mathbf{Th} \vdash 0 = 1 \rightarrow \varphi,$$

and so by the HBL-1 and HBL-2 also:

$$\begin{aligned} &\mathbf{Th} \vdash Pr_{Th}(\ulcorner 0 = 1 \rightarrow \varphi \urcorner) \\ &\mathbf{Th} \vdash Pr_{Th}(\ulcorner 0 = 1 \urcorner) \rightarrow Pr_{Th}(\ulcorner \varphi \urcorner) \end{aligned}$$

Since we have by definition of φ that $\varphi \rightarrow \neg Pr_{Th}(\ulcorner \varphi \urcorner)$, it follows that $\mathbf{Th} \vdash \varphi \rightarrow \neg Pr_{Th}(\ulcorner 0 = 1 \urcorner)$. Conversely, we can reason as follows in **Th**:

$$\begin{aligned}
&\mathbf{Th} \vdash Pr_{Th}(\ulcorner \varphi \urcorner) \rightarrow \neg\varphi \qquad \text{Definition of } \varphi \\
&\mathbf{Th} \vdash Pr_{Th}(\ulcorner Pr_{Th}(\varphi) \rightarrow \neg\varphi \urcorner) \qquad \text{By HBL-1} \\
&\mathbf{Th} \vdash Pr_{Th}(\ulcorner Pr_{Th}(\varphi) \urcorner) \rightarrow Pr_{Th}(\ulcorner \neg\varphi \urcorner) \qquad \text{By HBL-2} \\
&\mathbf{Th} \vdash Pr_{Th}(\ulcorner \varphi \urcorner) \rightarrow Pr_{Th}(\ulcorner \neg\varphi \urcorner) \qquad \text{By HBL-3} \\
&\mathbf{Th} \vdash Pr_{Th}(\ulcorner \varphi \urcorner) \rightarrow [Pr_{Th}(\ulcorner \neg\varphi \urcorner) \wedge Pr_{Th}(\ulcorner \varphi \urcorner)] \qquad \text{By propositional reasoning} \\
&\mathbf{Th} \vdash [\neg Pr_{Th}(\ulcorner \neg\varphi \urcorner) \vee \neg Pr_{Th}(\ulcorner \varphi \urcorner)] \rightarrow \neg Pr_{Th}(\ulcorner \varphi \urcorner) \qquad \text{By contraposition} \\
&\mathbf{Th} \vdash Con(Pr_{Th}) \rightarrow \varphi \qquad \text{By lemma 2.2.9 and definition of } \varphi
\end{aligned}$$

□

Note that in the case of a non-standard provability predicate, the proof of Lemma 2.2.9 does not go through, and the second incompleteness theorem does not apply. For example, the non-standard Feferman provability predicate Δ_f [Fef60] is extensionally correct, that is,

$$\mathbf{PA} \vdash \varphi \text{ iff } \mathbf{PA} \vdash \Delta_f(\ulcorner \varphi \urcorner).$$

Since it does not fulfill all of the HBL-conditions it is not a contradiction to the second incompleteness theorem that $\mathbf{PA} \vdash Con(\Delta_f)$.

2.3 Reflection Principles

In the proofs of the incompleteness theorems, the Gödel sentence, which asserts its own unprovability, played a crucial role. In 1952 Henkin raised the question of the status of the fixed points of $Pr_{Th}(x)$, that is, the sentences S for which $\mathbf{Th} \vdash S \leftrightarrow Pr_{Th}(S)$ [SKH52]. It is not a priori obvious what the answer should be, or that the answer should be the same for each of these fixed points. The solution is given by Löb's theorem, which can be proved elegantly from the second incompleteness theorem (and vice versa), as pointed out by Kripke and Kreisel years after the original proof was given [Boo94]. The proof, and the discussion of reflection principles in the remainder of this section essentially follow Smorynski [Smo77] and Beklemishev [Bek05].

Theorem 2.3.1 (Löb's Theorem). Let φ be a sentence of an axiomatizable extension **Th** of **PA**. Then

$$\mathbf{Th} \vdash Pr_{Th}(\ulcorner\varphi\urcorner) \to \varphi \text{ iff } \mathbf{Th} \vdash \varphi,$$

where Pr_{Th} is a standard provability predicate.

Proof. The direction from right to left follows immediately from HBL-1. For the direction from left to right, assume that $\mathbf{Th} \nvdash \varphi$. Then $\mathbf{Th} + \varphi$, standing for $\mathbf{Th} \cup \{\varphi\}$, is consistent. By the second incompleteness theorem, $\mathbf{Th} + \neg\varphi \nvdash Con(Pr_{Th+\neg\varphi})$, or equivalently $\mathbf{Th} + \neg\varphi \nvdash \neg Pr_{Th+\neg\varphi}(\ulcorner 0 = 1\urcorner)$. The latter statement is equivalent to:

$$\mathbf{Th} + \neg\varphi \nvdash \neg Pr_{Th}(\ulcorner\neg\varphi \to 0 = 1\urcorner).$$

But this is equivalent by propositional reasoning to:

$$\mathbf{Th} + \neg\varphi \nvdash \neg Pr_{Th}(\ulcorner\varphi\urcorner).$$

Applying the deduction theorem and contraposition yields the result:

$$\mathbf{Th} \nvdash Pr_{Th}(\ulcorner\varphi\urcorner) \to \varphi.$$

By contraposition on the assumption that $\mathbf{Th} \vdash \varphi$ we obtain the direction from left to right. □

Löb's theorem confirms that the fixed points of $Pr_{Th}(x)$ are all provable. The implication $Pr_{Th}(\ulcorner\varphi\urcorner) \rightarrow \varphi$ can be seen as expressing 'Th is sound'. What Löb's theorem tells us is that **Th** is very humble in this claim, only asserting its soundness for statements it can actually derive. One is naturally led to consider certain *reflection* principles, which schematically assert more comprehensive forms of the soundness of **Th** (with respect to a standard provability predicate Pr_{Th}):

Local Reflection Principle

$Rfn(Th) : Pr_{Th}(\ulcorner\varphi\urcorner) \rightarrow \varphi,\qquad \varphi$ is a sentence.

Uniform Reflection Principle

$RFN(Th) : \forall\bar{x}[Pr_{Th}(\varphi(\ulcorner\dot{\bar{x}}\urcorner)) \rightarrow \varphi(\bar{x})],\qquad \varphi$ only has $\bar{x}$ as free variables.

In the statement of the uniform reflection principle we have used Feferman's dot notation. Since $\ulcorner\varphi(x)\urcorner$ is a numeral, expressions such as $\forall x Pr_{Th}(\ulcorner\varphi(x)\urcorner)$ are not well-formed. Instead, Feferman's dot notation is used to express that x is a variable to be quantified over, and can be unpacked as follows:

$$\forall x P(\ulcorner\varphi(\dot{x})\urcorner) := \forall x \exists n, z[n = Numeral(x) \wedge z = Sub(\ulcorner\varphi(x)\urcorner, x, n) \wedge P(z)]$$

Since sequences can be coded into a single number by Lemma 2.2.3, the schema

$$\forall x(Pr_{Th}(\ulcorner\varphi(\dot{x})\urcorner) \rightarrow \varphi(x)),$$

where φ only has the single free variable x, is equivalent to $RFN(\mathbf{Th})$, if **Th** is sufficiently expressive to perform this coding. As before, we assume that **Th** is an axiomatizable extension of **PA**, so that this equivalence always holds. We will exploit this fact in some of the proofs that follow. First we show that the addition of either one of the previous reflection principles implies the consistency of the underlying theory.

It is immediate that $\mathbf{Th} + RFN(\mathbf{Th}) \vdash Rfn(\mathbf{Th})$. Also, $\mathbf{Th} + Rfn(\mathbf{Th}) \vdash Con(Pr_{Th})$ since $\mathbf{Th} + Rfn(\mathbf{Th}) \vdash Pr_{Th}(\ulcorner 0 = 1\urcorner) \rightarrow 0 = 1$. Since already $\mathbf{PA} \vdash 0 \neq 1$, by contraposition of the previous statement we have that $\mathbf{Th} + Rfn(\mathbf{Th}) \vdash \neg Pr_{Th}(\ulcorner 0 = 1\urcorner)$.

The previous reflection principles already overshoot the target with respect to consistency: as it turns out, consistency is equivalent to a restricted reflection principle. We denote the restriction of the local reflection principle to a class of formulas Γ as $Rfn_{\Gamma}(\mathbf{Th})$, and similarly for the uniform reflection principle.

Theorem 2.3.2. Let **Th** be an axiomatizable extension of **PA**. Over **PA**, the following are equivalent:

1. $Con(Pr_{Th})$;
2. $Rfn_{\Pi_1}(\mathbf{Th})$;

3. $RFN_{\Pi_1}(\mathbf{Th})$

Proof. The implication from 2. to 1. and 3. to 2. follow from the preceding discussion. For the implication from 1. to 3., let $\varphi(x) \in \Pi_1$, with only x free. Then $\neg\varphi(x) \in \Sigma_1$ and by Theorem 2.2.7:

$$\mathbf{PA} \vdash \neg\varphi(x) \rightarrow Pr_{Th}(\ulcorner \neg\varphi(\dot{x}) \urcorner) \tag{2.1}$$

By (a slight generalization of) Lemma 2.2.9 we also have:

$$\mathbf{PA} + Con(Pr_{Th}) \vdash \neg Pr_{Th}(\ulcorner \varphi(\dot{x}) \urcorner) \vee \neg Pr_{Th}(\ulcorner \neg\varphi(\dot{x}) \urcorner),$$

which is equivalent to:

$$\mathbf{PA} + Con(Pr_{Th}) \vdash Pr_{Th}(\ulcorner \varphi(\dot{x}) \urcorner) \rightarrow \neg Pr_{Th}(\ulcorner \neg\varphi(\dot{x}) \urcorner). \tag{2.2}$$

Combining 2.1 and 2.2 and contraposition leads to:

$$\mathbf{PA} + Con(Pr_{Th}) \vdash Pr_{Th}(\ulcorner \varphi(\dot{x}) \urcorner) \rightarrow \varphi(x)$$ □

The following result relates reflection principles for higher complexity classes:

Theorem 2.3.3. Let **Th** be an axiomatizable extension of **PA**. Over **PA**, RFN_{Σ_n} and $RFN_{\Pi_{n+1}}$ are deductively equivalent.

Proof. Take $\varphi(y,x)$ to be a Σ_n-formula. Then $\forall y\varphi(y,x) \in \Pi_{n+1}$. It is a fact that within **PA** we have provable closure under numerical substitution: $\mathbf{PA} \vdash Pr_{Th}(\ulcorner \forall x\varphi(x) \urcorner) \leftrightarrow \forall x Pr_{Th}(\ulcorner \varphi(\dot{x}) \urcorner)$. So also:

$$\mathbf{PA} + RFN_{\Sigma_n} \vdash Pr_{Th}(\ulcorner \forall y\varphi(y,\dot{x}) \urcorner) \rightarrow \forall y Pr_{Th}(\ulcorner \varphi(\dot{y},\dot{x}) \urcorner).$$

By RFN_{Σ_n} we have:

$$\mathbf{PA} + RFN_{\Sigma_n} \vdash \forall y Pr_{Th}(\ulcorner \varphi(\dot{y},\dot{x}) \urcorner) \rightarrow \forall y\varphi(y,x)).$$

Combining the foregoing we arrive at the desired result:

$$\mathbf{PA} + RFN_{\Sigma_n} \vdash Pr_{Th}(\ulcorner \forall y\varphi(y,\dot{x}) \urcorner) \rightarrow \forall y\varphi(y,x)).$$ □

Using the partial truth predicates of Corollary 4.1.3.1, we can show that the restricted reflection principles are finitely axiomatizable over **PA**.

Lemma 2.3.4. Let Th be an axiomatizable extension of **PA**. Then the schema RFN_{Π_n} is deductively equivalent to the universal instance:

$$\psi := \forall x[Pr_{Th}(\ulcorner T_{\Pi_n}(\dot{x}) \urcorner) \rightarrow T_{\Pi_n}(x)],$$

where T_{Π_n} is the partial truth predicate for formulas in Π_n. The same result holds for the schema RFN_{Σ_n} and T_{Σ_n}.

Proof. Clearly RFN_{Π_n} implies $\forall x[Pr_{Th}(\ulcorner T_{\Pi_n}(\dot{x})\urcorner) \rightarrow T_{\Pi_n}(x)]$. For the converse, observe that by HBL-1 and the definition of the partial truth predicates we have:

$$\mathbf{PA} \vdash \forall \bar{x} Pr_{Th}(\ulcorner \varphi(\dot{\bar{x}}) \leftrightarrow T_{\Pi_n}(\varphi(\dot{\bar{x}}))\urcorner)$$

So we can infer that for $\varphi(x) \in \Pi_n$, with x being the only free variable:

$$\mathbf{PA} \vdash Pr_{Th}(\ulcorner \varphi(\dot{x})\urcorner) \rightarrow Pr_{Th}(\ulcorner T_{\Pi_n}(\varphi(\dot{x}))\urcorner)$$

From which follows, with ψ denoting the universal instance:

$$\mathbf{PA} + \psi \vdash Pr_{Th}(\ulcorner \varphi(\dot{x})\urcorner) \rightarrow T_{\Pi_n}(\ulcorner \varphi(x)\urcorner),$$

and by the definition of the partial truth predicate we obtain $\mathbf{PA} + \psi \vdash Pr_{Th}(\ulcorner \varphi(\dot{x})\urcorner) \rightarrow \varphi(x)$, as required. □

Note that the proof does not go through for $n = 0$. This is because $T_{\Delta_0} \in \Delta_1$, whence RFN_{Δ_0} does not imply $\forall x[Pr_{Th}(\ulcorner T_{\Delta_0}(\dot{x})\urcorner) \rightarrow T_{\Delta_0}(x)]$ (although the converse still holds).

Corollary 2.3.4.1. For $n \geq 1$, the schemas RFN_{Π_n} and RFN_{Σ_n} are finitely axiomatizable over **PA**.

An important theorem, proved in [KL68] (incidentally, one of the loci classici for work on reflection principles) relates uniform reflection over the weak theory of elementary arithmetic **EA** to full induction over **EA**. The theory **EA** is a weak theory, amounting to **PA** with induction for Δ_0-formulas only. It derives its name from the fact that every primitive recursive function can be represented by a Δ_0-formula.

Theorem 2.3.5.

$$\mathbf{EA} + RFN(\mathbf{EA}) \equiv \mathbf{EA} + Ind(L_{PA}) \equiv \mathbf{PA}.$$

Illuminating discussion on the relation between reflection principles and induction, as well as the connection to the foundational programs pursued by Kreisel and Feferman, can be found in [Dea14]. For the purposes of the later results on reflection principles over truth theories, we are interested in the proof of a slightly weaker result, for the arithmetical theory $\mathbf{I\Sigma_1}$. The theory $\mathbf{I\Sigma_1}$ is formulated in the same language as **PA**, i.e. $L_{PA} = L_{I\Sigma_1}$. It is axiomatized by Robinson arithmetic **Q** as well as the induction schema for Σ_1-formulas only (hence it is slightly stronger than **EA**). The axioms of **Q** correspond to axioms 1-7 in our axiomatization of **PA**, as well as the axiom $\forall y[y = 0 \vee \exists x(S(x) = y)]$. The usual formalization of Gödel coding, provability and partial truth predicates (see theorem 4.1.3.1), is already possible within $\mathbf{I\Sigma_1}$, so that the apparatus we have developed in this chapter carries over.

Before we proceed to the theorem statement and proof, we state two lemmas that can be found in [HP98] as respectively $I.2.52$ and the combination of $III.3.17$ and $V.5.19$. The first lemma shows that, contrary to **PA** which is not finitely axiomatizable, restricting the induction schema leads to a finitely axiomatizable theory.

Lemma 2.3.6. For $n > 0$, the theories $\mathbf{I\Sigma_n}$, i.e. Q + induction over Σ_n-formulas, are finitely axiomatizable.

The proof is easiest when considering the sequent calculus proof system rather than a hilbert-style proof system. A sequent calculus differs from other proof systems in that it makes use of conditional validities, called sequents. A sequent is of the form $A_1, \ldots, A_n \vdash B_1, \ldots, B_n$ meaning that the conjunction of A_i implies the disjunction of B_i. As is usual, several rules are present for deriving new valid sequents from previous ones. The cut-rule is of particular interest, because it corresponds to the usage of a lemma in a proof. It is given by:

$$\textbf{Cut:}\ \frac{\Gamma \vdash A, \Delta \quad \Gamma', A \vdash \Delta'}{\Gamma, \Gamma' \vdash \Delta, \Delta'}.$$

Note the role of A as a lemma that can be 'cut' from the resulting sequent. What is peculiar about the cut-rule is that it is the only rule which violates the condition that every formula present in a sequent is already present in a sequent higher up the proof tree. A classic result in proof theory is the cut-elimination theorem due to Gentzen [Gen35] . It states that any proof can be transformed in a proof that does not rely on the cut-rule. This proof then has the property that any formula occurring in a sequent must have occurred higher up the proof tree. It is this property which we will exploit in the proof. The second lemma states that, similar to provability, cut-free provability in a sequent-calculus, denoted as $CFPr_{Th}(x)$, can be formalized within **PA**.

Lemma 2.3.7.

$$\mathbf{PA} \vdash \forall x[Pr_{PA}(x) \leftrightarrow CFPr_{PA}(x)].$$

Theorem 2.3.8.

$$\mathbf{I\Sigma_1} + RFN(\mathbf{I\Sigma_1}) \equiv \mathbf{I\Sigma_1} + Ind(L_{PA}) \equiv \mathbf{PA}.$$

Proof sketch. It is well known that $Q{+}Ind(L_{PA}) \equiv \mathbf{PA}$, and hence $I\Sigma_1{+}Ind(L_{PA}) \equiv \mathbf{PA}$. It remains to prove that $\mathbf{I\Sigma_1} + RFN(\mathbf{I\Sigma_1}) \equiv \mathbf{I\Sigma_1} + Ind(L_{PA})$. First we show that:

$$\mathbf{I\Sigma_1} + RFN(\mathbf{I\Sigma_1}) \vdash \mathbf{I\Sigma_1} + Ind(L_{PA}).$$

Working within $\mathbf{I\Sigma_1} + RFN(\mathbf{I\Sigma_1})$, assume that $\varphi(0)$ and $\forall x[\varphi(x) \rightarrow \varphi(S(x))]$. The standard provability predicate $Pr_{I\Sigma_1}$ fulfills the HBL derivability conditions so that by HBL-1, HBL-2, and closure under numerical substitution it holds that:

$$\mathbf{I\Sigma_1} \vdash Pr_{I\Sigma_1}(\ulcorner\varphi(0)\urcorner)$$
$$\mathbf{I\Sigma_1} \vdash \forall x[Pr_{I\Sigma_1}(\ulcorner\varphi(\dot{x})\urcorner) \rightarrow Pr_{I\Sigma_1}(\ulcorner\varphi(S(\dot{x}))\urcorner)].$$

Note that $Pr_{I\Sigma_1}(\ulcorner\varphi(\dot{x})\urcorner)$ is Σ_1, since both the provability predicate and Feferman's dot notation are Σ_1. Hence, by Σ_1-induction in $\mathbf{I\Sigma_1}$, it is provable that $\forall x Pr_{I\Sigma_1}(\ulcorner\varphi(\dot{x})\urcorner)$.

By $RFN(\mathbf{I\Sigma_1})$, it follows that $\forall x \varphi(x)$, which is what we set out to prove.
For the other direction, we need to show that:

$$\mathbf{I\Sigma_1} + Ind(L_{PA}) \vdash \mathbf{I\Sigma_1} + RFN(\mathbf{I\Sigma_1}).$$

Since $\mathbf{I\Sigma_1} + Ind(L_{PA}) \equiv \mathbf{PA}$, it suffices to work in $\mathbf{PA}$. First, we show that the following partial global reflection principle for $\mathbf{I\Sigma_1}$ is provable within $\mathbf{PA}$:

$$\forall x[(Sent_{\Sigma_n}(x) \wedge Pr_{I\Sigma_1}(x)) \to T_{\Sigma_n}(x)].$$

To show this, assume that d is a proof of $\varphi(\bar{x})$ by the lights of $Pr_{I\Sigma_1}$, that is, a hilbert-system style proof. By Lemma 2.3.6, we have that $\mathbf{I\Sigma_1}$ can be axiomatized by a single sentence (using conjunction) I_1. Hence there is a proof d' from I_1 to $\varphi(\bar{x})$. It follows by lemma 2.3.7 that there is a cut-free sequent-calculus proof d" of I_1 to $\varphi(\bar{x})$. The idea is to show, within $\mathbf{PA}$, that:

1. It holds that $T_{\Sigma_n}(\ulcorner I_1 \urcorner)$, for some n;
2. If all cut-free sequent-calculus proofs of depth m fall under T_{Σ_n}, all cut-free sequent-calculus proofs of depth $m+1$ fall under T_{Σ_n};
3. By applying induction to the formalized statements of the previous two properties, we have that $T_{\Sigma_n}(\ulcorner \varphi(\dot{\bar{x}}) \urcorner)$.

We do not go the extra mile of formalizing these steps. Note however that that the sub-formula property of cut-free sequent-calculus proofs is instrumental in the proof. The sub-formula property states that every formula occurring in the proof is a sub-formula of the last sequent in the proof. This implies that the complexity of each formula occurring in the proof is bounded by a given maximum n (by virtue of the finite axiomatization of $I\Sigma_1$), which makes it possible to use a bounded partial truth predicate.

Finally, combining the partial global reflection principle with Corollary 4.1.3.1 which says that $\forall \bar{x}[\varphi(\bar{x}) \leftrightarrow T_{\Sigma_n}(\ulcorner \varphi(\dot{\bar{x}}) \urcorner)]$, $RFN(\mathbf{I\Sigma_1})$ follows. □

CHAPTER 3

History of Reflection Principles

Reflection principles go back a long time, to the work of Kreisel, Levy and Feferman in the early 60's [KL68], [Fef62]. These reflection principles, and the theories that result from adjoining them to base theories, were studied principally with certain foundational aims in mind. Understanding the philosophical import of Kreisel and Feferman's work requires going back even further, to the early 20's, when Hilbert laid out his programme. The next section is devoted to giving a short summary of his programme, and the historical context in which it was embedded. It is necessarily simplified; a more nuanced and thorough overview can be found in [Zac07].

3.1 Hilbert's Programme

Hilbert's programme was an encompassing proposal to offer a foundation which could ground mathematics and mathematical practice. This came at a time where an awareness grew that the intuition of the working mathematician could lead one astray, as the discovery of Russell's paradox proved. When Hilbert's programme is mentioned, usually in the context of the incompleteness theorems, it is often summarized as an attempt at axiomatizing mathematics, and more importantly, proving the consistency of this axiomatization. It is less often mentioned that this proof of consistency was to be done using only *finitary* methods, whereby the consistency of mathematics would be justified by the obvious correctness of these methods. The emphasis on a finitist approach to consistency proofs is to a large extent a reaction of Hilbert to the intuitionist mathematics of Brouwer, and later the apostatic Weyl, which famously rejects the principle of the excluded middle and the notion of a completed infinite totality [Iem20]. The consequence is that much of classical mathematics, and practice, is simply not available anymore. Hilbert on the other hand took a broadly formalist, and instrumentalist position with regards to mathematics. In his lecture on the concept of infinity he states:

> No, when aside from the proof of consistency the question of justification of a [newly introduced] measure should have a sense, then it is only this one, whether the measure is accompanied by corresponding success. Indeed, success is necessary, it is here too the highest authority to which everyone bows [Hil26].[1]

Recourse to the fruitfulness and consistency of axiomatic systems would not be enough to convince the adepts of intuitionism however. Taking up the challenge to make the axiomatic systems in question incontrovertible required delineating a secure part of mathematics which could serve as the bedrock for all other parts of the field. This aim was fulfilled by distinguishing between the 'real' and 'ideal' parts of mathematics [Zac07]. The real part of mathematics consists of finitistically meaningful statements, which belong to *contentual number theory*. Ideal mathematics on the other hand is not meaningful from a finitistic point of view, but is useful. An example of an ideal principle is the principle of excluded middle for infinite totalities, for instance, as applied to the infinite set of natural numbers, that each number either has a certain property or a number exists which lacks the property. Contentual number theory is privileged in so far that it is, in a Kantian vein, intuitively accessible to us:

> [. . .] Mathematics has at its disposal a content which is secured independent from all logic, and so can never be grounded solely through logic. Hence, the efforts of Frege and Dedekind had to fail. Rather, certain extra-logical concrete objects are already represented to us as immediate intuitions prior to all thought, forming the precondition for the application of logical inferences, and the usage of logical operations [Hil26].[2]

Side-stepping for now the question of what exactly contentual number theory is supposed to be, Hilbert's proposal was to justify adjoining the ideal part of mathematics, with its instrumental value, to the real part, i.e. contentual number theory. Aside from the required fruitfulness, the consistency of this expanded theory needs to be proved, using only finitary principles of reasoning. The reason Hilbert could believe that this was possible is that consistency is a claim about finite objects, namely that no proof leads to a contradiction.

[1] Nein, wenn über den Nachweis der Widerspruchsfreiheit hinaus noch die Frage der Berechtigung zu einer Maßnahme einen Sinn haben soll, so ist es doch nur die, ob die Maßnahme von einem entsprechenden Erfolge begleitet ist. In der Tat, der Erfolg ist notwendig; er ist auch hier die höchste Instanz, der sich jedermann beugt. (Own Translation)

[2] [...] daß die Mathematik über einen unabhängig von aller Logik gesicherten Inhalt verfügt und daher nie und nimmer allein durch Logik begründet werden kann, weshalb auch die Bestrebungen von Frege und Dedekind scheitern mußten. Vielmehr ist als Vorbedingung für die Anwendung logischer Schlüsse und für die Betätigung logischer Operationen schon etwas in der Vorstellung gegeben: gewisse, außer-logische konkrete Objekte, die anschaulich als unmittelbares Erlebnis vor allem Denken da sind. (Own Translation)

Although Hilbert never explicitly made it an aim of his foundationalist programme, there is an interesting link between a consistency proof and conservativity. A consistency proof of the ideal theory does not only establish that it is free from contradictions, it also shows that if a real statement is proven in the ideal theory, it is already provable in the real theory, by finitary reasoning. In this sense, ideal mathematics is justified beyond mere consistency, as the ideal reasoning can always be replaced with finitary reasoning for real mathematics. An informal argument (adapted from [Smo77]) illustrates how these are related. Take **R** and **I** to be respectively formalizations of real and ideal mathematics, and let $\varphi := \forall x(f(x) = g(x))$ be a real statement, provable in ideal mathematics. It is important that φ is a Π_1 formula, since the result does not hold for formulas of higher complexity, as we shall see in the discussion of a formal counterpart to the argument.

Under a given encoding this means that there is a derivation d such that:

$$\mathbf{I} \vdash Pr_I(\ulcorner d \urcorner, \ulcorner \varphi \urcorner).$$

Since derivations are finite strings of symbols, they are real concrete objects and so fall under the purview of **R**, implying that also:

$$\mathbf{R} \vdash Pr_I(\ulcorner d \urcorner, \ulcorner \varphi \urcorner).$$

Now suppose that in fact φ does not hold, so that $f(a) \neq g(a)$ for some a. Then by the principle of excluded middle (which is finitistically acceptable in this case since φ is a real statement) $\neg\varphi$ holds, and so we have that there is a derivation c so that:

$$\mathbf{R} \vdash Pr_I(\ulcorner c \urcorner, \ulcorner \neg\varphi \urcorner).$$

By assumption, **R** proves the consistency of **I**, implying that

$$\mathbf{R} \vdash \neg[Pr_I(\ulcorner c \urcorner, \ulcorner \neg\varphi \urcorner) \wedge Pr_I(\ulcorner d \urcorner, \ulcorner \varphi \urcorner)],$$

from which it follows that $\mathbf{R} \vdash \varphi$. This informal argument has a formal counterpart in Theorem 2.3.2. Assume that **PA** formalizes real mathematics (which however most finitists would deny), and a stronger axiomatizable extension **Th** formalizes ideal mathematics. For any given Π_1-formula $\forall x\varphi(x)$, if $\mathbf{Th} \vdash \forall x\varphi(x)$, then by the equivalence over **PA** of $Con(Pr_{Th})$ and $Rfn_{\Pi_1}(Th)$, we have that:

$$\mathbf{PA} + Con(Pr_{Th}) \vdash Pr_{Th}(\ulcorner \forall x\varphi(x) \urcorner) \to \forall x\varphi(x)$$

Since by HBL-1 it already holds that $\mathbf{PA} \vdash Pr_{Th}(\ulcorner \forall x\varphi(x) \urcorner)$, it follows that $\mathbf{PA} \vdash \forall x\varphi(x)$.

Hilbert never set out generally what the finitistically valid principles are [Zac07]. The principle of induction, and primitive recursion applied to finite totalities were considered finitistically valid, but it was never claimed that these principles exhausted finitary reasoning. These principles are formalizable as primitive recursive arithmetic (**PRA**), where the only functions are those definable by primitive recursion, and only quantifier-free formulas are substitutable in the induction schema . The ambiguity of whether this is also an upper limit on finitary reasoning made it possible for Kleene to assert as late as 1952 of Gentzen's consistency proof of **PA**, which uses transfinite induction up to ϵ_0 (exactly what this means is explained in the next section) over the finitary theory **PRA**, that "to what extent the Gentzen proof can be accepted as securing classical number theory [...] [depends] on how ready one is to accept induction up to ϵ_0 as a finitary method" [Kle52]. As it stands, considering such principles of transfinite induction as finitarily acceptable is a minority position, and the incompleteness theorems are understood as showing that no finitary consistency proofs of the axiomatic theories that Hilbert considered are possible.

3.2 Turing's Ordinal logic

The germ of reflection principles can already be found in Turing's PhD thesis [Tur36]. Taking the incompleteness results as a point of departure, rather than an end of the road, he aimed to construct ever more complete theories by adjoining consistency statements. In his own words, the incompleteness result "indicates means whereby from a system L of logic a more complete system L' may be obtained." We follow the exposition in [RS20] of his results. The idea is simple enough. We take an evidently correct, but incomplete theory $\mathbf{T_0}$ as starting point, adjoin $Con(Pr_{T_0})$ to obtain $\mathbf{T_1} \supseteq \mathbf{T_0}$ and so on. This process can be iterated into the *transfinite*, i.e. beyond finite indexes. The correct way to do this is to make use of *ordinal numbers*, which are a generalization of the natural numbers, where the fact that the natural numbers are ordered is not ignored, in contrast to the generalisation of cardinals. We now give a short introduction to ordinals as relevant for the theories we will consider.

3.2.1 Ordinals and Their Representation

Ordinals were first introduced by Cantor (in fact this is part of what Hilbert refers to when he states that "No one shall expel us from the paradise that Cantor has created for us." [Hil26]), and the theory of ordinals has since developed in step with the needs of logic to refer to every larger 'infinities'. For our purposes, we only need relatively small ordinals, which we will motivate heuristically. A thorough introduction to the subject, developed within set theory, can be found in chapter 7 and 8 of [End77].

Similar to how cardinal numbers represent the equivalent sets under a bijection, ordinals represent the equivalent sets under *order-preserving* bijections. More precisely, we are considering well-orderings, which are total orders that are well-founded. As a reminder, a total order is defined as follows:

Definition 3.2.1. A total order $(<, A)$ is a pair of a binary relation $<$ on a non-empty set A for which holds:

- $\forall x \in A[\neg x < x]$ (Irreflexivity)
- $\forall x \in A, \forall y \in A, \forall z \in A[x < y \land y < z \to x < z]$ (Transitivity)

- $\forall x \in A, \forall y \in A[x < y \vee y < x \vee y = x]$ (Trichotomy)

A total order $(<, A)$ is well-founded if every non-empty subset of (A) has a least element, i.e. satisfies:

- $\forall X[X \subseteq A \wedge X \neq \emptyset \rightarrow \exists x \in X, \forall y \in A[y < x \rightarrow y \notin X]$.

Now, an order-preserving bijection f on two well-orderings $(A_1, <_1)$ and $(A_2, <_2)$, is a bijection for which holds that $f(a_1) <_2 f(a_2)$ if $a_1 <_1 a_2$, where a_1 and a_2 are elements of A_1. This induces an equivalence relation on well-orderings, and we identify ordinals with the equivalence classes. For technical reasons this definition is not tenable in practice, and instead we identify ordinals with a representative of the equivalence class. The representative ordinals we will consider are defined within set-theory. Before we get to the formal definition, we discuss ordinals completely informally. The set we will consider is the set of the natural numbers, and by varying the order over the natural numbers, we will obtain different ordinals.

The finite ordinals are those with which we become familiar at an early age: $0, 1, 2, \ldots$ The first infinite ordinal is:

$$\omega := 0 < 1 < 2 < \ldots,$$

which is the pair of the natural numbers with the canonical ordering. Now consider the following ordering of the natural numbers, with $0 > n$ for each $n \in \mathcal{N}$:

$$\omega + 1 := 1 < 2 < \cdots < 0.$$

Clearly ω can not be isomorphic to $\omega + 1$, since ω has no last element, whereas $\omega + 1$ does, but it is isomorphic to the initial segment of $\omega + 1$. This is reflected by the notation, suggesting that $\omega + 1$ has one more element than ω. Now, one can easily go further, by 'gluing' ω to itself:

$$\omega + \omega := \omega * 2 = 0 < 2 < \cdots < 2n < \cdots < 1 < 3 < \ldots 2n + 1 < \ldots.$$

Using the lexicographic ordering over tuples of natural numbers (which can be coded as single natural numbers), one can introduce exponentiation:

$$\omega^2 := (0, 1) < (0, 2) < \cdots < (1, 0) < (1, 1) < \cdots < (2, 0) < (2, 1) < \ldots$$

We will need larger ordinals still, but first define (set-theoretic) ordinals formally.

Definition 3.2.2. A set A is *transitive* iff every element B of A is also a subset of A.

The reason behind calling such a set transitive is that if $C \in B$ and $B \in A$, then $C \in A$.

Definition 3.2.3. An *ordinal* α is a set which is both transitive, and well-ordered by the membership relation $\in$.

The finite ordinals are then given as:

$$\begin{aligned}
&0 = \{\} = \emptyset \\
&1 = \{0\} = \{\emptyset\} \\
&2 = \{0, 1\} = \{\{\emptyset\}, \emptyset\} \\
&3 = \{0, 1, 2\} = \{\{\{\emptyset\}, \emptyset\}, \{\emptyset\}, \emptyset\} \\
&\quad \dots
\end{aligned}$$

It is an axiom of set theory that $\omega = \bigcup\{n \mid n \in \mathcal{N}\}$ exists. The following lemmas are easy to prove from the definition of ordinals:

Lemma 3.2.1. If α is an ordinal, then the *successor ordinal* $suc(\alpha) = \alpha + 1 := \alpha \cup \{\alpha\}$ is also an ordinal.

Lemma 3.2.2. Let A be a set of ordinals. Then $\beta = \bigcup\{\alpha \mid \alpha \in A\}$ is an ordinal.

As an example, all finite ordinals are successor ordinals, and can be constructed from 0. However, ω is not a successor ordinal. It is an example of a *limit ordinal*, for which no ordinal α exists such that $suc(\alpha)$ is that ordinal.

It can be shown that the set-theoretic definition of ordinals is indeed the representative of an equivalence class, as every well-ordered set has a unique order-preserving isomorphism to exactly one ordinal. This gives rise to the following definition:

Definition 3.2.4. The order type $Ord(<, A)$ of a well-ordered set $(<, A)$ is the unique ordinal α for which there is an order-preserving isomorphism from $(<, A)$ to α.

In the informal examples, we have suggestively written down ordinals using addition, multiplication, and exponentiation. This notation is justified by the possibility to define arithmetic operations over ordinals.

Definition 3.2.5. Let α and β be ordinals. The set A is defined as $\alpha \,\Sigma\, \beta := \alpha \cup \beta'$, where β' is the set β with all elements $b \in \beta$ for which $b \in \alpha$ renamed to $b' \notin \alpha$. Define the well-order $<$ on A as the extension of the respective well-orders of α and β with the property that for each $b \in \beta'$ and $a \in \alpha$: $b' > a$. Then $\alpha + \beta := Ord(\alpha \,\Sigma\, \beta, <)$.

Couching the previously informal example in terms of this definition:

$$\omega + 1 = Ord(\mathcal{N} \,\Sigma\{0\}, <) = Ord(0 < 1 < \cdots < 0').$$

Note that addition on ordinals is not commutative:

$$1 + \omega = Ord(\{0\} \,\Sigma\, \mathcal{N}, <) = Ord(0' < 0 < 1 < \dots) = \omega.$$

This definition vindicates the notation $Suc(\alpha) = \alpha + 1$. Similarly, we can define multiplication over ordinals.

Definition 3.2.6. Let α and β be ordinals. The set A is defined as the Cartesian product $\alpha \times \beta$. Define a well-order $<$ on $(a, b) \in A$ as follows: $(a, b) < (a', b')$ iff $b < b'$ or $b = b'$ and $a < a'$. Then $\alpha * \beta := Ord(\alpha \times \beta, <)$.

Formalizing a previously informal example once again:

$$\omega * 2 = Ord(\omega \times \{0,1\}, <) = ((0,0) < (1,0) < (2,0) \dots (0,1) < (1,1) < (2,1) < \dots)$$

Also multiplication fails to be commutative:

$$2 * \omega = Ord(\{0,1\} \times \omega, <) = ((0,0) < (1,0) < (0,1) < (1,1) < (0,2) < (1,2) < \dots) = \omega$$

Exponentiation is easier to define recursively, which has the drawback that it requires proving that transfinite recursion over the ordinals is well-defined. We will not show this, but details can be found in the beginning of [End77, Chapter 8].

Definition 3.2.7. For a non-zero ordinal α and ordinal β, exponentiation α^β is defined recursively as:

- $\alpha^0 = 2$;
- $\alpha^{suc(\beta)} = \alpha^\beta * \alpha$;
- $\alpha^\lambda = \bigcup\{\alpha^\beta | \beta \in \lambda\}$, when λ is a limit ordinal.

For example, $\omega^2 = \omega * \omega$, and $\omega^\omega = \bigcup\{\omega^n | n \in \omega\}$. We now come to the first epsilon ordinal ϵ_0, which we encountered in the previous section. It is defined as:

$$\epsilon_0 = \bigcup\{\omega, \omega^\omega, \omega^{\omega^\omega}, \dots\}.$$

It has the interesting fixed-point property that $\epsilon_0 = \omega^{\epsilon_0}$. Also the other operations have fixed points:

- ω^n is a fixed point for addition from the left, as for all ordinals $\alpha < \omega^n$, $\alpha + \omega^n = \omega^n$.
- ω^{ω^n} is a fixed point for multiplication from the left, as for all ordinals $\alpha < \omega^{\omega^n}$, $\alpha * \omega^{\omega^n} = \omega^{\omega^n}$.
- ϵ_0 is a fixed point (the least) for exponentiation, as for all ordinals $\alpha < \epsilon_0$, $\alpha^{\epsilon_0} = \epsilon_0$.

This perspective suggest another way to construct new ordinals: define a well-behaved operation on the ordinals (i.e. a *normal* function), then the fixed points of this operation will themselves be ordinals. It's been a bit of a hike along peaks of ever larger ordinals, but we now scale our last peak using this newest tool in the backpack. We define the binary Veblen function $\alpha, \beta \to \varphi(\alpha, \beta)$ as follows:

- $\varphi(0, \alpha) = \omega^\alpha$
- For $\beta > 0$: $\varphi(\beta, \alpha)$ is the α-th element of the sequence of fixed points $\forall \gamma < \beta : [\delta = \varphi(\gamma, \delta)]$, the sequence being ordered along increasingly larger ordinals.

By definition $\varphi(1, 0)$ is the first fixed point of $\delta = \omega^\delta$, which we saw to be ϵ_0. The larger epsilon ordinals can be obtained as $\varphi(1, 1) = \epsilon_1$, $\varphi(1, 2) = \epsilon_2$ etc. The largest ordinal we will require in the remainder of this chapter is the Feferman-Schütte ordinal Γ_0, which is defined as the smallest ordinal α such that $\alpha = \varphi(\alpha, 0)$.

So far we have only considered ordinals as sets. The theories we will consider in this chapter are mostly first-order arithmetic theories, in which the natural numbers (numerals) are essentially the domain under consideration. Nevertheless, it is possible to code ordinals larger than ω in such a way that they can be 'reasoned' with within arithmetic, similarly to how **PA** is able to reason about its own consistency. For a given (recursive) ordinal α, we can define a recursive relation R representing an ordering $<_R$ such that $(<_R, \mathcal{N})$ is a well-ordering which is represented by α. In practice, this is usually accomplished by means of an *ordinal notation system*, an example of which we will see in the next section. The important principle of *transfinite induction* generalizes induction along the natural numbers to induction up to a given ordinal as follows:

$$TI(\alpha, \varphi) := \forall x, y[(xRy \to \varphi(x)) \to \varphi(y)] \to \forall x \varphi(x).$$

The statement that **PA** can't prove the principle of transfinite induction up to ϵ_0 then really means that it can't show that the natural encoding R of ϵ_0 corresponds to a well-ordering. Similar to the issues we encountered in formalizing provability within **PA**, questions of naturalness also apply to the coding of ordinals, as shown by Kreisel, and discussed in [Rat99].

3.2.2 Ordinal logics

We already discussed the basic idea behind the ordinal logics of Turing, namely adjoining the consistency statement of a given theory to obtain stronger theories. If we index our theories by ordinals, it is easy to see how to obtain $\mathbf{T}_{suc(\alpha)}$ from $\mathbf{T}_\alpha$. Perhaps less obvious is the case of $\mathbf{T}_\lambda$, where λ is a limit ordinal, since we can never reach λ by the successor operation. However, given that we have expressed our trust in $\mathbf{T}_\alpha$, for each $\alpha < \lambda$, it is reasonable to express our trust in the union of all theories $\mathbf{T}_\alpha$. Formally:

$$\begin{aligned} &\mathbf{T}_{\alpha+1} = \mathbf{T}_\alpha + Con(Pr_{T_\alpha}) \qquad \text{For } \alpha + 1 \text{ a successor ordinal.} \\ &\mathbf{T}_\lambda = \bigcup_{\alpha < \lambda} T_\alpha \qquad \text{For } \lambda \text{ a limit ordinal.} \end{aligned}$$

A few complications arise due to the fact that we require the theories thus obtained to be recursively axiomatizable, and for **PA** to 'recognize' them as such. Moreover, we have seen that a proof predicate (and also the derived consistency statement) can be encoded in such a way that the incompleteness result does not even apply. The solution is to use Kleene's $\mathcal{O}$ ordinal notation system, which contains a notation (natural number) for each recursive ordinal, and to encode the axioms of each theory by a Σ_1-formula. It is this latter condition that leads to natural consistency statements under a given natural encoding (for a discussion of this issue, see [Fef62]). In order to effectively represent ordinals in our arithmetic theories, we use Kleene's $\mathcal{O}$ ordinal notation system, given by:

Definition 3.2.8. We use the notation $suc(a) = 2^a$ for successor ordinals, and $lim(e) = 3 \cdot 5^e$ for limit ordinals. The ordinals $|a|$ represented by their notation $a \in \mathcal{O}$, with the notations partially ordered by $<_{\mathcal{O}}$, are defined simultaneously as follows:

1. The ordinal $|0|$ is represented by $0 \in \mathcal{O}$.
2. If the ordinal notation a representing ordinal $|a|$ is in $\mathcal{O}$, then the notation $suc(a) \in \mathcal{O}$, $a <_{\mathcal{O}} suc(a)$, and $|suc(a)| = |a| + 1$.
3. If e is the index of a computable function, then $\{e\}(n)$ represents the result of applying e to n. If $\{e\}(n) <_{\mathcal{O}} \{e\}(n+1)$ for every natural number n, then $lim(e) \in \mathcal{O}$ and $|lim(e)| = Sup(\{|\{e\}(n)| \mid n \in \mathbb{N}\}$.
4. $<_{\mathcal{O}}$ is transitive.

Note that $<_{\mathcal{O}}$ is a partial order since there are infinitely many notations for each limit ordinal (item 3 of the definition), leading to infinitely many branches in the continuing successor ordinals. Only the finite ordinals are uniquely denoted.

We can now state the preceding discussion formally. We consider a consistency progression based on $\mathbf{T}$, which is a primitive recursive function $n \to Ax_n(x)$ mapping each natural number to a Σ_1-formula $Ax_n(x)$ representing the axioms of $\mathbf{T}_n$ in such a way that the following holds:

- $\mathbf{PA} \vdash \mathbf{T}_0 = \mathbf{T}$;
- $\mathbf{PA} \vdash \mathbf{T}_{suc(a)} = \mathbf{T}_a \cup Con(Pr_{T_a})$;
- $\mathbf{PA} \vdash \mathbf{T}_{lim(e)} = \bigcup_x \mathbf{T}_{\{e\}(x)}$.

With these conditions satisfied, the following theorem obtains:

Theorem 3.2.3. For any true (in the standard model) Π_1-sentence φ, a notation a_φ can be constructed such that $|a_\varphi| = \omega + 1$ and $\mathbf{T}_{a_\varphi} \vdash \varphi$, where $\mathbf{T}_{a_\varphi}$ is a theory in a consistency progression based on a recursively axiomatizable theory $\mathbf{T}$ extending $\mathbf{PA}$. Moreover, the function $\varphi \to a_\varphi$ is primitive recursive.

At first blush this theorem seems to make good on the promise of generating ever more complete theories. Since the natural consistency statement for a given theory **T** is a Π_1-sentence, we can indeed 'complete' the theory **T**. However, analyzing the proof shows that the difficulty is hidden away in the construction of a_φ. For a given true Π_1-sentence $\psi := \forall x \varphi(x)$, the proof defines a computable function indexed by e such that:

$$\{e\}(n) = \begin{cases} n_{\mathcal{O}} & \text{if } \varphi(\underline{k}) \text{ is true for every } k \leq n \\ suc(lim(e)) & \text{otherwise,} \end{cases}$$

where $n_{\mathcal{O}}$ stands for the ordinal notation of the natural number n in $\mathcal{O}$. The computable function so defined is then used to construct the ordinal notation a_ψ such that $|a_\psi| = \omega + 1$. But this implies that coming to know that $T_{a_\psi} \vdash \psi$ is of no value in theorem discovery, since it already presupposes that we are able to determine that $a_\psi \in \mathcal{O}$, which requires recognizing that $\varphi(x)$ is true for every n, by the first condition of the definition of $\{e\}(n)$.

Turing also considered the question of whether completeness could be achieved with respect to true sentences of higher complexity, Π_2-sentences to be exact. He conjectured that this could be done using recursive progressions based on adjoining, in our terminology, the Rfn_{Π_2} principle. However, this conjecture was shown to be false by Feferman [Fef62], as a direct corollary of the following theorem.

Theorem 3.2.4. Let $(\mathbf{T}_a)_{a \in \mathcal{O}}$ be a recursive progression based on a recursively axiomatizable theory **T** extending **PA** by adjoining the (unrestricted) local reflection principle. That is, $\mathbf{T}_{suc(a)} = \mathbf{T}_a \cup Rfn(T_a)$. Then it holds that $\bigcup_{a \in \mathcal{O}} \mathbf{T}_a \subseteq \mathbf{T}+$ all true Π_1-sentences.

Proof. Denoting $\mathbf{T}+$ all true Π_1-sentences by $\mathbf{T}^*$, we will show by induction over $a \in \mathcal{O}$ that $\mathbf{T}_a \subseteq \mathbf{T}^*$. The case where $a = 0$ is trivial, since $\mathbf{T} = \mathbf{T}_0$. For the induction step where $b = suc(a)$, assume that $\mathbf{T}_a \subseteq \mathbf{T}^*$. Proving that $\mathbf{T}_b \subseteq \mathbf{T}^*$ amounts to showing that $\mathbf{T}^* \vdash Rfn(T_a)$.

Take an arbitrary sentence φ. In the first case, $Pr_{T_a}(\ulcorner \varphi \urcorner)$ is true, in which case $\mathbf{T}_a \vdash \varphi$. By induction hypothesis, it holds that $\mathbf{T}^* \vdash \varphi$ which entails $\mathbf{T}^* \vdash Pr_{T_a}(\ulcorner \varphi \urcorner) \to \varphi$. For the second case, assume that $Pr_{T_a}(\ulcorner \varphi \urcorner)$ is false. In that case, $\neg Pr_{T_a}(\ulcorner \varphi \urcorner)$ is a true Π_1-sentence (Since Pr_{T_a} is a Σ_1-formula), and by definition $\mathbf{T}^* \vdash \neg Pr_{T_a}(\ulcorner \varphi \urcorner)$, which entails $\mathbf{T}^* \vdash Pr_{T_a}(\ulcorner \varphi \urcorner) \to \varphi$.

Finally, for the induction step where $b = lim(e)$ is a limit ordinal notation, suppose that $\mathbf{T}_a \subseteq \mathbf{T}^*$ for all $a <_{\mathcal{O}} b$. It then follows by our definition of the limit ordinal notation that:

$$\mathbf{T}_b = \bigcup_x \mathbf{T}_{\{e\}(x)} = \bigcup_{a <_{\mathcal{O}} b} \mathbf{T}_a \subseteq \mathbf{T}^*.$$

□

Corollary 3.2.4.1. There is a true Π_2-sentence which is not provable in $\bigcup_{a \in \mathcal{O}} \mathbf{T}_a$.

Proof. We show that there is a true Π_2-sentence which is not provable in $\mathbf{T}^*$, and a fortiori, is not provable in $\bigcup_{a \in \mathcal{O}} \mathbf{T}_a$. Because $\mathbf{T}^*$ contains all true Π_1-sentences, it is not recursively axiomatizable, and so Ax_{T^*} is of Π_1-complexity, which given the standard coding of provability leads to its provability predicate being Σ_2. By the diagonal lemma there is a sentence φ such that:

$$\mathbf{T}^* \vdash \varphi \leftrightarrow \neg Pr_{T^*}(\ulcorner \varphi \urcorner).$$

By construction, φ is of Π_2 complexity. It holds that $\mathbf{T}* \nvdash \varphi$, by a similar argument as is used in the proof of the first incompleteness theorem (φ asserts its own unprovability). □

Although Turing's conjecture was shown to be false, his approach was substantially generalized by Feferman in his 1962 paper [Fef62]. In this paper he makes use of a new type of progressions, so-called *autonomous progressions* . He was indebted to Kreisel for the idea of it, which will be the subject of the next section.

3.3 Kreisel's Analysis of Finitism

While finitism could not, as a philosophical position, be the bedrock for all of mathematics that Hilbert envisioned, it still remained a position of interest for its own sake. For example, Tait has argued that finitist reasoning represents a kind of lower limit if one is to reason at all about numbers:

> Rather, the special role of finitism consists in the circumstance that it is a minimal kind of reasoning presupposed by all nontrivial mathematical reasoning about numbers. [...] Thus finitism is fundamental to mathematics even if it is not a foundation in the sense Hilbert wished. [Tai81]

A large part of Kreisel's research was devoted to the study of 'informal rigour' in analyzing common notions (e.g. [Kre67], [Kre87]). As opposed to Hilbert's conception of proof theory being the study of formal systems first and foremost, with the hard part being the development of the correct tools of formalization, his position was that the actual difficulty to be tackled was the relation between the intuitive notion of proof and its formalization [Kre68]. It is in this context that his analysis of finitary proof has to be understood. For Kreisel, the incompleteness theorems did not necessarily preclude the correct formalization of the informal notion of finitist proof, since it is not a given that one's informal notion includes that every formula is either provable or refutable [Kre58]. In a later, more mature proposal he couches the formalization of an informal concept in terms of reflection:

> The process of recognizing the validity of such [proof] principles (including principles for defining new concepts, that is, formally, of extending a given language) is here conceived as a process of reflection; reflecting on the given concepts, reflecting on this process of reflection, and so forth [Kre70].

In other words, one takes a first limited stab at producing the formal counterpart to an informal notion, and then unfolds the implicit content within through reflection on this initial formalisation.

How is this applied to the notion of finitist proof? Kreisel takes it that if Pr_{Σ_μ} has been recognized by finitist means to be the provability predicate of a partial formalization of finitist mathematics $\mathbf{\Sigma}_\mu$, and $Pr_{\Sigma_\mu}(\ulcorner\varphi(0^x)\urcorner)$ has been established, where x is a free variable, and $\ulcorner\varphi(0^x)\urcorner$ is the function of x which returns $\ulcorner\varphi(S(\ldots S(0))\urcorner$ with x occurrences of S, then $\varphi(x)$ has been finitistically established. Taking **PRA** to be the initial partial formalisation of finitism, he argues that the result of this progression through formal systems leads to **PA**, so that the finitist theorems are co-extensive with those of **PA**. With respect to the discussion of the preceding section, the most interesting point is that this progression is understood more strictly than Turings 'wide' ordinal logic. Where Turing associated a formal system to each ordinal in a path through $\mathcal{O}$ extensionally, in Kreisel's analysis each system $\mathbf{\Sigma}_\mu$ has to *prove* that $suc(\mu) \in \mathcal{O}$ in order for $\mathbf{\Sigma_{suc(\mu)}}$ to be part of the progression. Kreisel is rather terse on philosophical motivation for what Feferman later named autonomous progressions. A more thorough motivation for studying these progressions can be found in [Fef62], where Feferman gives a kind of phenomenological analysis of the workings of an idealized mathematician. The mathematician in question is working with a collection of formal systems such that the theory $\mathbf{T_d}$, where $d \in \mathcal{O}$, is defined in terms of the generation procedure of Section 3.2.2. By subscribing to the generation procedure (whether it is adding consistency or a reflection principle) they should find $\mathbf{T_d}$ an acceptable theory if $\mathbf{T_0}$ was acceptable. However, as they continue applying the generation procedure it might not be possible for them to determine whether $d \in \mathcal{O}$, and so they'll be unable to continue unless an oracle is available. If we consider the formal systems in question as representing the limits of their mathematical prowess, then no such oracle is available, and $d \in \mathcal{O}$ is a statement to be proved within a previous system in the progression.

The upshot of restricting oneself to autonomous progressions is that – as opposed to the progressions considered by Turing – any form of completeness is out of the question (which fits nicely with Kreisel's objection to completeness being necessary for formalizing finitary provability). After proving an existence theorem for autonomous progressions, Feferman shows that the autonomous progression is restricted to a recursively enumerable subset $\mathcal{O}' \subset \mathcal{O}$. It is easy to see that if $\mathbf{A} = \bigcup_{d\in\mathcal{O}'} \mathbf{T}_d$, then $\varphi \in \mathbf{A}$ if and only if there exists a d such that $d \in \mathcal{O}'$ and $\mathbf{PA} \vdash Pr_{T_d}(\ulcorner\varphi\urcorner)$. But this implies that $\mathbf{A}$ is recursively enumerable, since $\mathcal{O}'$ is recursively enumerable and so is Ax_{T_d} by construction. By the incompleteness theorems not even $Con(Pr_A)$ is provable in $\mathbf{A}$.

3.4 Feferman's Predicativism

We have already had cause to mention several results of Feferman's 1962 paper. As we saw in the previous section, Turing's conjecture that Π_2-completeness could be achieved using progressions based on the local reflection principle was settled in the negative. Feferman showed that the situation is drastically different if the uniform reflection principle is used to generate succeeding theories.

Theorem 3.4.1. Let $(T_a)_{a \in \mathcal{O}}$ be a recursive progression based on **PA** and the uniform reflection principle. Then for every true $\varphi \in Sent_{PA}$ there exists an $a \in \mathcal{O}$ such that $\mathbf{T}_a \vdash \varphi$.

Similarly to Turing's Π_1-completeness result, this theorem is of little epistemological value, since recognizing that $a \in \mathcal{O}$ for a given $\mathbf{T}_a \vdash \varphi$ is at least as hard as recognizing that φ is true.

Feferman went on to study progressions further in service of a reevaluation of the philosophical position of predicativism. We will limit ourselves to the minimal exposition of predicativism necessary to understand Feferman's contribution and the role of reflection principles therein. A much more expansive discussion of the history of predicativism can be found in [Fef09]. Predicativism is one answer to the issues raised by the discovery of Russell's paradox (as Hilbert's finitism was one answer). The paradox itself is well known; consider the set which contains all sets that do not contain themselves, then the contradiction that the set contains itself if and only if it does not contain itself follows. Poincaré drew attention to other paradoxes which did not involve set-theoretic notions, but all seemed to make use of the *vicious circle principle*. For example, in Russell's paradox the set in question is defined in terms of itself: $S = \{x \mid x \notin x\}$. Any definition of an entity in terms of the class to which the entity belongs is potentially problematic. The culpable axiom schema in naive set theory is unrestricted comprehension:

$$\forall \bar{a} \exists X \forall x (x \in X \leftrightarrow \varphi(x, \bar{a})).$$

In words, any formula φ can be used to define the set X, even when φ refers to X itself. The solution followed in **ZF** set-theory is to get rid of the axiom of comprehension, while making up for the loss by introducing other restricted axioms of set-definition. An impredicative notion still occurs, namely in the axiom schema of separation:

$$\forall \bar{a} \forall Y \exists X \forall x [x \in X \leftrightarrow (x \in Y \wedge \varphi(x, \bar{a}, Y)].$$

Since the formula φ could still quantify over the totality of all sets (including the set X to be defined), this is an impredicative definition.

It's important to realize that the rift between mathematics as built on impredicative set theory and predicativism is a philosophical one and not a matter of avoiding paradox. Using **ZFC** set-theory as basis for mathematics is uncontroversial, as virtually no one expects further contradictions to lurk in the axioms (even if we have to take certain independence results in stride). However, the usage of impredicative definitions does seem to commit one to a platonistic ontology of mathematics [Fef64]. By assuming the validity of the axiom schema of separation, we assume the meaning of the defining formula φ to be well-determined. Since this formula could refer to the totality of all sets, this can only be if these sets have an independent existence, regardless of whether we are able to define them. On the other hand, predicativism as defended by Feferman takes only the natural numbers as a given – and contrary to constructivism, as a completed totality – and all other mathematical objects to be defined in terms of the natural numbers, or objects constructed previously. This naturally leads to a ramified hierarchy (similar to Russell's ramified theory of types), where objects can be classified according to the stage at which they were constructed. While this hierarchy is natural from the predicativist point of view, it is unwieldy to the working mathematician: imagine doing analysis with real numbers of different degrees. The first major work on the technical possibilities of predicativism can be found in Weyl's *Das Kontinuum* [Wey18] – a youthful 'sin' before his conversion to intuitionism. As Weyl realized the inacceptability of introducing degrees of the reals, he restricted himself to the first stage, using comprehension axioms of the form:

$$\forall y \forall z \ldots \forall Y \forall Z \ldots \exists S \forall x [x \in S \leftrightarrow \varphi(x, y, z, \ldots, Y, Z, \ldots)].$$

In the comprehension axiom schema, the capital letters represent set variables and the formula φ is an arithmetical formula (i.e. not containing bound set variables) which notably does not refer to the set S to be defined. By restricting the sets under consideration to the arithmetically definable ones (of degree 0), no ramification results. Surprisingly, this system of *arithmetic analysis* includes most of classical analysis, after replacing the usual impredicative definitions with arithmetical ones.

Due to the different philosophical developments in intuitionism and Hilbert's programme, Weyl's predicativism went largely ignored until the 1950's. Feferman picks up the thread in [Fef64], in that he pursues a full explication of predicative mathematics, necessarily having to deal with the degrees inherent in a ramified hierarchy. He does this by looking at autonomous recursive progressions of two different kinds, both based on the uniform reflection principle. The first kind is one based on a ramified theory. The set variables $X, Y, Z \ldots$ are now replaced with variables $X^a, Y^a, Z^a \ldots$ of degree a. Every set variable in a formula must have a specific degree associated with it, and the formulas of the new language are called *graded* formulas. The degree $d(\varphi)$ of the formula φ is then given by the maximum of all $a + 1$ and b, where a is the degree of a bound set variable and b is the degree of a free set variable. We now define the ramified theory $\mathbf{R_c}$ indexed by the ordinal notation c in the autonomous recursive progression based on the uniform reflection principle. As before, for each theory $\mathbf{R_c}$ we have available a standard provability predicate $Pr_{R_c}(x)$ which is defined in terms of the Σ_1-formula $Ax_c(x)$ representing the axioms of $\mathbf{R_c}$.

Definition 3.4.1. The theory $\mathbf{R_c}$ is a recursively enumerable theory defined by:

Log The basic logical axioms for numerical and set variables, as well axioms for identity;

Ind For arbitrary graded formulas φ with $d(\varphi) \leq_{\mathcal{O}} c$ the induction axiom

$$\varphi(0) \wedge \forall x(\varphi(x) \rightarrow \varphi(x+1)) \rightarrow \forall x \varphi(x);$$

RC For each $a \leq_{\mathcal{O}} c$ and each graded formula φ with $d(\varphi) \leq_{\mathcal{O}} a$ the ramified comprehension axiom:

$$\exists X^a \forall x[x \in X^a \leftrightarrow \varphi(x)],$$

with X^a not free in φ;

Ref The appropriate reflection principle for each preceding theory R_a with $suc(a) <_{\mathcal{O}} c$ and formula φ with $d(\varphi) \leq a$:

$$\forall x Pr_{R_a}(\ulcorner \varphi(\dot{x}) \urcorner) \rightarrow \forall x \varphi(x);$$

Lim The appropriate limit generalization rules for each $a = lim(e)$ with $a \leq_{\mathcal{O}} c$ and $suc(d) \leq_{\mathcal{O}} a$:

$$\forall z \leq_{\mathcal{O}} a Pr_{R_d}(\ulcorner \forall X^z \varphi(X^z) \urcorner) \rightarrow \forall X^a \varphi(X^a)$$

Note that in the definition the axiom *Lim* fulfills two roles at once: It is both a reflection principle, and a limit generalization principle, in a form we have not seen previously. As before, the progression introduced here is supposed to unfold the concept of predicativism, with the notion of constructing mathematical objects in stages carried out into the transfinite.

Aside from the R_c progression, Feferman also considers an *unramified* progression, which he argues can be understood to capture predicativism just as well. This progression H_c always includes the hyperarithmetical comprehension rule (HCR):

$$\frac{\forall x[\varphi(x) \leftrightarrow \psi(x)]}{\exists X \forall x[x \in X \leftrightarrow \varphi(x)]},$$

where set variables only occur universally quantified in φ (Π^1_1-formula) and only occur existentially quantified in ψ (Σ^1_1-formula). Similar to the arithemetical hierarchy definition, this implies that both φ and ψ are Δ^1_1 formulas. On the face of it, this rule is not predicatively justifiable, since φ contains universal quantification over sets. We only sketch the argument for its predicative validity, which is based on a return to the notion of the 'potential' rather than actual totality of all sets. Instead of attaching degrees to sets related to the stage in which they have been constructed, we allow quantification over all sets within a restricted class of formulas only. This class of formulas is defined with respect to a given collection of sets in such a way that their 'meaning' is stable even if the collection were expanded. Let's make this concrete. For an arbitrary second-order formula φ and collection of sets $\mathcal{M}$ we define $\varphi_{\mathcal{M}}$ as the formula φ with all set variable quantifiers ranging over sets in $\mathcal{M}$ only. Then a formula φ is called *definite relative to* $\mathcal{M}$ iff for every $\mathcal{N} \supseteq \mathcal{M}$ it holds that $\forall x[\varphi_{\mathcal{M}}(x) \leftrightarrow \varphi_{\mathcal{N}}(x)]$. Now, for a given collection of sets $\mathcal{M}$ which has been constructed in a predicatively justifiable way, it seems perfectly justifiable to expand it with sets that have been defined using formulas that are definite relative to $\mathcal{M}$. This process of expansion can be carried out into the transfinite in the usual way. As it turns out, the fixed point of this *predicative set construction* process, starting from the arithmetically definable sets, coincides with the collection of hyperarithmetical sets, which are precisely the sets that are Δ^1_1-definable. The HCR-rule is then the proof-theoretic analogue of this result.

The autonomous progression H_c, using the uniform reflection principle, is then based on the system $\mathbf{H}$ which is given by the *Log* axioms, as well as induction for arbitrary second-order formulas, and the HCR-rule. Now, Feferman goes on to show that for both the progressions R_c and H_c the least ordinal that cannot to be shown to be an ordinal within a theory $\mathbf{R}_a$ or $\mathbf{H}_a$ within the respective progressions, i.e. is not autonomous with respect to the progression, is Γ_0, now known as the Feferman-Schütte ordinal. Moreover, he shows that the theories $\cup_{a \leq_{\mathcal{O}} \Gamma_0} \mathbf{R}_a$ and $\cup_{a \leq_{\mathcal{O}} \Gamma_0} \mathbf{H}_a$ are equivalent in so-far that they are intertranslatable:

Theorem 3.4.2.

- With any $a \leq_{\mathcal{O}} \Gamma_0$ and φ such that $\mathbf{H}_a \vdash \varphi$ we can associate $c, d \leq_{\mathcal{O}} \Gamma_0$ such that $\mathbf{R}_d \vdash \varphi^{(c)}$, where $\varphi^{(c)}$ is the result of replacing every set variable X in φ by X^c, the set variable of degree c.
- With any $a \leq_{\mathcal{O}} \Gamma_0$ and φ such that $\mathbf{R}_a \vdash \varphi$ we can associate $b, c \leq_{\mathcal{O}} \Gamma_0$ such that $\mathbf{H}_c \vdash \forall X \varphi_{\mathcal{M}_b(X)}$, where $\mathcal{M}_b(X)$ is a collection based on the predicative set construction process starting from X, at stage b.

In conclusion, what Feferman, and independently Schütte [Sch65] showed is that Γ_0 is the least well-ordering which is impredicative with respect to provability, thus 'calibrating' the extent of predicative mathematics. Moreover, due to the equivalence sketched between $\cup_{a \leq_{\mathcal{O}} \Gamma_0} \mathbf{R}_a$ and $\cup_{a \leq_{\mathcal{O}} \Gamma_0} \mathbf{H}_a$, the problem of ramified objects to mathematical practice is also avoided.

CHAPTER 4

Truth Theories

Although much is made of the philosophical and mathematical import of the Gödel incompleteness theorems, a different set of incompleteness results, due to Tarski [Tar33], deserves to be at least as well known. These limiting results show that any sufficiently expressive system cannot represent truth within the system, or in other words, represent their own semantics. We should be careful to once again distinguish what we can achieve from what we cannot. As we saw before, the incompleteness results do not preclude the representability of very many syntactical notions after all. Similarly, we will see that a more limited arithmetical truth is within our grasp. Finally, we will introduce the main truth theories with which this thesis will be concerned, and discuss in which way they do, and do not live up to expectations.

4.1 Tarski's Undefinability Theorem

The undefinability theorem is nothing more than a formalization of the liar sentence in natural language:

$$\text{Liar} \;:=\; \text{“Liar is false”}\;.$$

The liar sentence asserts its own falsehood, leading to paradox, since if it is true, it is false, and vice versa. First, we prove a syntactical result, which shows that within a theory one can't *define* a truth predicate T which fulfills the condition that $T(\ulcorner\varphi\urcorner) \leftrightarrow \varphi$, i.e. the T-schema .

Theorem 4.1.1. Tarski's Undefinability Theorem Let **Th** be a consistent theory containing negation for which the syntactical Diagonal Lemma 2.2.4 holds. Then there is no formula $\varphi(x) \in L_{Th}$ (with the intended meaning 'is true') such that for all sentences $\psi \in L_{Th}$:

$$\mathbf{Th} \vdash \varphi(\ulcorner\psi\urcorner) \leftrightarrow \psi$$

Proof. For a contradiction, assume that $\varphi(x) \in L_{Th}$, and for all sentences $\psi \in L_{Th}$: $\mathbf{Th} \vdash \varphi(\ulcorner \psi \urcorner) \leftrightarrow \psi$. The Diagonal Lemma applies to **Th**, so we can fix a sentence γ such that:

$$\mathbf{Th} \vdash \gamma \leftrightarrow \neg\varphi(\ulcorner \gamma \urcorner)$$

But by assumption we also have that:

$$\mathbf{Th} \vdash \varphi(\ulcorner \gamma \urcorner) \leftrightarrow \gamma$$

From which we derive that $\mathbf{Th} \vdash \varphi(\ulcorner \gamma \urcorner) \leftrightarrow \neg\varphi(\ulcorner \gamma \urcorner)$, and a contradiction follows immediately. □

The previous result might, falsely, give the impression that the undefinability of truth has something to do with provability. In fact, the concept of truth can't even coherently be *expressed* in language, with no reference to provability. In other words, even if we relaxed the requirement that our theory be recursively axiomatizable, there is no truth-property available in a first-order language. The lack of a truth predicate is not an epistemic limitation (if only we could calculate better ...), but an ontological one.

Theorem 4.1.2. No formula $\varphi(x) \in L_{PA}$ exists which expresses the property of arithmetical truth, that is $\mathbb{N} \models \psi$ iff $\mathbb{N} \models \varphi(\ulcorner \psi \urcorner)$.

Proof. Assume for a contradiction that such a formula $\varphi(x) \in L_{PA}$ exists. By the semantic Diagonal Lemma there is a sentence γ such that

$$\neg\varphi(\ulcorner \gamma \urcorner) \leftrightarrow \gamma,$$

is true. But since φ is supposed to express the property of arithmetical truth it must also be true that:

$$\varphi(\ulcorner \gamma \urcorner) \leftrightarrow \gamma,$$

from which a contradiction follows. □

While the undefinability theorem puts a damper on our hopes for a straightforward theory of truth, namely the theory itself, we can still make a few positive claims. These definability results require a lot of bookkeeping and formalisation to prove, so we will content ourselves with sketching the argument. First notice that an arbitrary finite set $\Gamma = \{\varphi_1, \ldots, \varphi_n\}$ of L_{PA} formulas has a perfectly well defined truth predicate, namely:

$$T_\Gamma(x) := \bigvee_{i=1}^{n} x = \ulcorner \varphi_i \urcorner \wedge \varphi_i.$$

It is straightforward that indeed $T_\Gamma(\ulcorner \psi \urcorner) \leftrightarrow \psi$.

This is not the best we can do, in fact arithmetical truth can be defined for ever larger classes of the arithmetical hierarchy. The most laborious step is to give, inside **PA**, a definition of $Sat_{\Delta_0}(x, y) \in \Delta_1$, that is, a satisfaction relation where x is a (coded) formula, and y is a (coded) variable assignment . Once a satisfaction relation is defined, we can define a truth predicate as $T_\Gamma(\ulcorner\varphi\urcorner) \leftrightarrow \forall y Sat_\Gamma(\ulcorner\varphi\urcorner, y)$, where Γ is a class of formulas. The different codings that follow are adapted from Kaye's 'Models of Peano Arithmetic' [Kay91]. First we need to see how to evaluate terms, which is in structure very similar to the definition of $Tm(x)$.

Definition 4.1.1. $Valseq(y, s, t)$ is the formula given by:

$$\begin{aligned} Valseq(y, s, t) = {} & Termseq(s) \wedge len(t) = len(s) \wedge \\ & \forall i{<}len(s) : \\ & (s)_i = \ulcorner 0 \urcorner \wedge (t)_i = 0 \vee \\ & \exists j{\le}s\ ((s)_i = \ulcorner v_j \urcorner \wedge (t)_i = (y)_j) \vee \\ & \exists j{<}i\ ((s)_i = \ulcorner S((s)_j) \urcorner \wedge (t)_i = S((s)_j)) \vee \\ & \exists j, k{<}i\ ((s)_i = \ulcorner ((s)_j + (s)_k) \urcorner \wedge (t)_i = (t)_j + (t)_k) \vee \\ & \exists j, k{<}i\ ((s)_i = \ulcorner ((s)_j \cdot (s)_k) \urcorner \wedge (t)_i = (t)_j \cdot (t)_k). \end{aligned}$$

And $Val(x, y) = z$ is the formula:

$$\exists s, t\ Valseq(y, s \cap (x), t \cap z) \vee (\neg Tm(x) \wedge z = 0).$$

Intuitively, $Val(x, y) = z$ means that z is the value of the (coded) term x, given a variable assignment y. Defining $Sat_{\Delta_0}(x, y)$ is then a matter of coding the truth compositionality of a formula under a given assignment. We restrict ourselves to the cases equality of terms and conjunction by way of example.

Definition 4.1.2. $Satseq_{\Delta_0}(s, t)$ is a formula encoding the compositionality of truth. We only show the conditions for equality of terms and conjunction, highlighted in blue and red respectively:

$$\begin{aligned} & Form_{\Delta_0}(s) \wedge \\ & \forall l{<}len(t) \exists i, z, w{\le}t[(t)_l = \langle i, z, w \rangle \wedge i{<}len(s) \wedge w \le 1 \wedge \\ & \{\exists u, u'{\le}s[Tm(u) \wedge Tm(u') \wedge (s)_i = \ulcorner u = u' \urcorner \wedge \\ & (w = 1 \leftrightarrow Val(u, z) = Val(u', z))] \vee \\ & \exists j, k{<}i[(s)_i = \ulcorner ((s)_j \wedge (s)_k) \urcorner \wedge \\ & \exists l_1, l_2{<}l\ \exists w_1, w_2{\le}1((t)_{l_1} = \langle j, z, w_1 \rangle \wedge (t)_{l_2} = \langle k, z, w_2 \rangle \wedge \\ & (w = 1 \leftrightarrow w_1 = 1 \wedge w_2 = 1))]\} \vee \\ & \cdots] \end{aligned}$$

Finally, $Sat_{\Delta_0}(x, y)$ is given by:

$$\exists s, t\{Satseq_{\Delta_0}(s \cap (x), y) \wedge \exists l{<}len(t)[(t)_l = \langle len(s), y, 1 \rangle)]\}.$$

The intuitive meaning of $Satseq_{\Delta_0}(s,t)$ is that t is a code for a sequence of triples $\langle i,z,w\rangle$, with i being an index (smaller than the length of s), and $(s)_i$ being a formula whose 'truth value' is w, given a variable assignment z. Checking that all the compositional properties hold, as well as that $Sat_{\Delta_0}(x,y) \in \Delta_1$ would take us too far, but can be proved [Kay91, Chapter 9].

Now we continue to climb up along the arithmetical hierarchy. Take $Sat_{\Sigma_0}, = Sat_{\Pi_0} = Sat_{\Delta_0}$, and proceed recursively:

Definition 4.1.3.

$$\begin{aligned} Sat_{\Sigma_{n+1}}(x,y) = \exists s,t[&len(t) = len(s) > 0 \land Form_{\Sigma_{n+1}}(x) \land (s)_{len(s)} = x \\ &\land (t)_{len(t)} = y \land \forall i{<}len(s)(i > 0 \to \exists k{\le}s\exists z{\le}t((s)_i = \ulcorner \exists v_k (s)_{i-1} \urcorner) \\ &\land (t)_{i-1} = Subst((t)_i, z, k)) \land Sat_{\Pi_n}((s)_0, (t)_0)]. \end{aligned}$$

Essentially, the formula 'checks' if a variable assignment exists for all variables quantified over in the outermost existential quantifier, and then makes use of the satisfiability predicate defined at an earlier stage. The case for $Sat_{\Pi_{n+1}}$ is similar.

Theorem 4.1.3 (Partial Satisfaction Definability). We use the notation $[\dot{\bar{x}}]$ to represent code of the variable assignment corresponding to $\bar{x}$ when quantified over. For arbitrary $\varphi \in \Sigma_n$, where φ is a formula of L_{PA}:

$$\mathbf{PA} \vdash \forall \bar{x}[\varphi(\bar{x}) \leftrightarrow Sat_{\Sigma_n}(\ulcorner \varphi(\bar{x}) \urcorner, [\dot{\bar{x}}])].$$

Similarly we have for arbitrary $\varphi \in \Pi_n$, where φ is a formula of L_{PA}:

$$\mathbf{PA} \vdash \forall \bar{x}[\varphi(\bar{x}) \leftrightarrow Sat_{\Pi_n}(\ulcorner \varphi(\bar{x} \urcorner, [\dot{\bar{x}}])].$$

The previous theorem for satisfaction implies a corresponding result for partial truth predicates:

Corollary 4.1.3.1 (Partial Truth Definability). For arbitrary $\varphi \in \Sigma_n$, where φ is a formula of L_{PA}:

$$\mathbf{PA} \vdash \forall \bar{x}[\varphi(\bar{x}) \leftrightarrow T_{\Sigma_n}(\ulcorner \varphi(\dot{\bar{x}}) \urcorner)].$$

Similarly we have for arbitrary $\varphi \in \Pi_n$, where φ is a formula of L_{PA}:

$$\mathbf{PA} \vdash \forall \bar{x}[\varphi(\bar{x}) \leftrightarrow T_{\Pi_n}(\ulcorner \varphi(\dot{\bar{x}}) \urcorner)]$$

The complexity of T_{Δ_0} is Δ_1, while the complexity of T_{Σ_n} and T_{Π_n} is Σ_n and Π_n respectively.

4.2 Disquotational Theories

The results of the previous section illustrate what is and what is not possible with regards to truth, within the confines of standard Peano arithmetic (or true arithmetic as in Theorem 4.1.2). To go further, we extend Peano arithmetic – our base theory, providing us with the means to encode syntax – by adding a truth predicate. This truth predicate is axiomatized in such a way that it captures, or at least approaches, the desired informal notion of truth. This raises the question what the informal notion of truth is that we want to capture. One answer – which will occupy us for the remainder of this section – can be found in Tarski's 1933 paper [Tar33]. His answer is that the truth predicate should fulfill a certain *material adequacy condition*. Before we make this condition explicit, we provide our setting some formal footing :

Definition 4.2.1. L_T is the language obtained by adding a new one-place predicate T to the language L_{PA}.

Having added a new predicate, Peano arithmetic needs to be reformulated within the new language:

Definition 4.2.2. PAT is the theory in the language L_T containing the logical axioms in L_T, the axioms of **PA**, and all instances of the induction schema for formulas in L_T.

A truth-theory with **PAT** as a base-theory is then any recursively axiomatizable theory extending **PAT**, within the language L_T. In practice, a truth-theory will only add axioms containing the truth predicate, since it is the concept of truth which we want to formalize. In our setting, Tarski's adequacy condition can be formulated as follows:

Definition 4.2.3. A truth-definition is materially adequate for L_{PA} if and only if the Tarski-biconditional $T(\ulcorner\varphi\urcorner) \leftrightarrow \varphi$ is provable within the truth-theory for every sentence $\varphi \in L_{PA}$. $T(\ulcorner\varphi\urcorner) \leftrightarrow \varphi$ will also be referred to as the 'T-schema'.

Note that the adequacy condition makes a distinction between the *object-language* L_{PA}, and the truth theory formulated in a *meta-language* L_T. Whether or not this adequacy condition is sufficient depends in parts on one's aims: If we are interested in a truth theory in a restricted setting only, the necessity of a meta-language does not seem to be a drawback. If, on the other hand, we consider it necessary to give an account of truth within the object-language itself, taking the object-language to be English for example, then this adequacy condition is too narrow. We can give the more ambitious attempt, to formulate a truth theory where object- and meta-language coincide, a try. Clearly, **PAT** is not yet a truth-theory, given that the predicate T has no axioms defining it yet (it does not extend **PAT**). A naive approach to fulfill the adequacy condition is to simply add the desired biconditionals, for each formula in L_T to **PAT**. Unfortunately, this approach immediately founders, as Tarski's undefinability theorem applies to any recursively axiomatizable extension of **PA**. We will not consider more sophisticated approaches to formulate an *untyped* theory of truth, but instead turn towards the more humble aim of fulfilling the adequacy condition with L_{PA} as object-language.

If instantiating the Tarski-biconditionals for each sentence of L_T is too much to ask, perhaps doing so for each sentence of L_{PA} fares better.

Definition 4.2.4. The theory **TB**('Tarski Biconditionals') is given by all the arithmetical substitutions of the T-schema, as well as the axioms of **PAT**:

$$Ax(\mathbf{TB}) = Ax(\mathbf{PAT}) \cup \{T(\ulcorner\varphi\urcorner) \leftrightarrow \varphi \mid \varphi \in Sent_{PA}\}.$$

Why do we expect this theory to fare any better? Returning to Tarski's undefiniability theorem, we see that it hinges on the existence of a liar-sentence φ, for which $\neg T(\ulcorner\varphi\urcorner) \leftrightarrow \varphi$. Inconsistency derives from trying to apply the truth-predicate to this sentence, which already asserts its own falsehood, which forces one to consider the status of $T(\ulcorner\neg T(\varphi)\urcorner)$. By allowing the truth-predicate to apply only to arithmetical sentences, in which the truth predicate does not occur, no self-referential paradoxes can rear their head. We now show that **TB** is consistent, by showing that it has a model.

Theorem 4.2.1. **TB** has a model, namely an expanded standard model $\mathcal{M} := \langle\mathbb{N}, \mathcal{E} = \{\ulcorner\varphi\urcorner \mid \mathbb{N} \models \varphi, \text{ and } \varphi \in Sent_{PA}\}\rangle$, where $\mathcal{E}$ is the extension of the truth predicate.

Proof. We have to show that the axioms of TB are modelled in $\mathcal{M}$. We only concern ourselves with the Tarski-biconditionals, and the induction axiom schema. We have for an arbitrary $\varphi \in Sent_{PA}$:

$$\mathcal{M} \models T(\ulcorner\varphi\urcorner) \Leftrightarrow \mathbb{N} \models \varphi.$$

But since φ is an arithmetical sentence, the extension of the truth predicate is irrelevant to the evaluation of φ:

$$\mathbb{N} \models \varphi \Leftrightarrow \mathcal{M} \models \varphi.$$

Putting this together we have that indeed:

$$\mathcal{M} \models T(\ulcorner\varphi\urcorner) \leftrightarrow \varphi.$$

Now consider the induction axiom schema, for all $\varphi(x) \in L_T$: :

$$\varphi(0) \wedge \forall x\,[\varphi(x) \rightarrow \varphi(S(x))] \rightarrow \forall x\varphi(x).$$

We only need to show that the instances of the schema for which φ contains a truth predicate are modelled by $\mathcal{M}$. As we have just shown that $\mathcal{M} \models T(\ulcorner\varphi\urcorner) \leftrightarrow \varphi$, it suffices to replace every instance of $T(\ulcorner\psi\urcorner$ in φ by ψ. The resulting instance of the induction axiom schema will then not contain the truth predicate, and be modelled by $\mathcal{M}$ since already $\mathbb{N}$ models the instance. □

Corollary 4.2.1.1. **TB** is consistent.

A consistent truth theory, which fulfills Tarski's material adequacy condition: not too shabby! In other respects however, **TB** does not measure up to the standard. What it lacks is *truth-theoretic* strength. While the precise sense in which this notion needs to be taken differs among authors (see for example [Hal14, p.73]), the idea is roughly the following: when reasoning with truth, some principles are taken be to evident, and license inferences. For example, the law of excluded middle in classical logic seems to apply equally well to sentences themselves, as assertions of the truth of these sentences. The previous remark suggests that we need to be able to 'quantify over sentences' (remember that this is still ordinary quantification, but the Gödel encoding let's us express such notions). We will use the following abbreviations:

$$\forall x(Sent_{PA}(x) \to \dots) \text{ as } \forall \varphi \in Sent_{PA}(\dots) \text{ and}$$
$$\exists x(Sent_{PA}(x) \wedge \dots) \text{ as } \exists \varphi \in Sent_{PA}(\dots).$$

Following the example, it would be a sign of **TB**'s truth-theoretic strength if it were the case that it proved the truth of the principle of excluded middle:

$$\mathbf{TB} \vdash \forall \varphi \in Sent_{PA} : T(\ulcorner \varphi \vee \neg\varphi \urcorner).$$

Notice that for individual formulas, the principle holds always. Since for an arbitrary φ we have that $\mathbf{TB} \vdash \varphi \vee \neg\varphi$, the respective Tarski-biconditional leads us to conclude that also $\mathbf{TB} \vdash T(\ulcorner \varphi \vee \neg\varphi \urcorner)$. What **TB** turns out to be unable to do is prove the generalization to all formulas, which is a corollary of the following theorem (due to Halbach [Hal01a]):

Theorem 4.2.2. Let $\varphi \in L_{PA}$ be a formula for which holds that $\mathbf{TB} \vdash \forall x[\varphi(x) \to T(x)]$, then it also holds that there is a natural number n such that $\mathbf{TB} \vdash \forall x[\varphi(x) \to T_n(x)]$, where T_n is a partial truth-predicate (as in Corollary 4.1.3.1).

Proof. By assumption, there is a (finite) derivation d of $\forall x[\varphi(x) \to T(x)]$ within **TB**. Define $Rank(\psi)$ to be the minimum level of the arithmetical hierarchy in which ψ, or a formula logically equivalent to it, falls. That is, if $\psi \in \Gamma_n$ and $\psi \notin \Gamma_{n-1}$, where $\Gamma_n \in \{\Sigma_n, \Pi_n\}$, then $Rank(\psi)$ is n.

Take n to be the natural number given by $1 + Max(\{Rank(\psi) \mid [T(\ulcorner \psi \urcorner) \leftrightarrow \psi] \in d\}$. We can now construct a second derivation d', with all occurrences of T replaced by the partial truth-predicate T_n. By definition of the partial truth-predicate T_n, all steps in the derivation d' are still theorems of **TB**. In particular, $\mathbf{TB} \vdash \forall x[\varphi(x) \to T_n(x)]$. □

A corollary of the theorem is that **TB** cannot prove the truth of the law of excluded middle in general.

Corollary 4.2.2.1.

$$\mathbf{TB} \nvdash \forall x[\exists \varphi(x = \ulcorner \varphi \vee \neg\varphi \urcorner) \to T(x)]$$

Proof. Assume for a contradiction that indeed $\mathbf{TB} \vdash \forall x[\exists \varphi(x = \ulcorner \varphi \vee \neg\varphi \urcorner) \to T(x)]$. Then, by the previous theorem, $\mathbf{TB} \vdash \forall x[\exists \varphi(x = \ulcorner \varphi \vee \neg\varphi \urcorner) \to T_n(x)]$. This implies that there is a bounded truth-predicate which suffices to track the truth of all instances of the law of excluded middle. But this cannot be the case, since the formulas in question can be of arbitrary high level in the arithmetical hierarchy. □

Also compositional truth-principles exceed the strength of **TB**. For example:

Theorem 4.2.3.

$$\mathbf{TB} \nvdash \forall \varphi \in Sent_{PA} : T(\ulcorner \neg\varphi \urcorner) \leftrightarrow \neg T(\ulcorner \varphi \urcorner)$$

Proof. We show that the set $Z := \Gamma \cup \{\neg\forall\varphi(T(\ulcorner\neg\varphi\urcorner) \leftrightarrow \neg T(\ulcorner\varphi\urcorner))\}$, where Γ is a finite subset of the axioms of **TB**, has a model. By compactness, also $\mathbf{TB} \cup \{\neg\forall\varphi(T(\ulcorner\neg\varphi\urcorner) \leftrightarrow \neg T(\ulcorner\varphi\urcorner))\}$ has a model. The result then follows by soundness.

Once again we consider an expanded standard model of Peano arithmetic, with the truth predicate T_Γ interpreted as $\{\ulcorner\varphi\urcorner \mid \mathbb{N} \models \varphi \text{ and } [T(\ulcorner\varphi\urcorner) \leftrightarrow \varphi] \in \Gamma\}$. By construction, $\langle \mathbb{N}, T_\Gamma \rangle \models \Gamma$. Also, for any sentence ψ for which neither the sentence itself, nor its negation, occurs in a Tarski-biconditional in Γ, we have that $\langle \mathbb{N}, T_\Gamma \rangle \models \neg T(\psi)$ and $\langle \mathbb{N}, T_\Gamma \rangle \models \neg T(\neg\psi)$. Hence, $\langle \mathbb{N}, T_\Gamma \rangle \models Z$. □

So far we have seen that while **TB** formalizes the disquotationalist intuition, and is consistent, it does not prove the truth of the principle of excluded middle, or compositional truth-principles. In fact, any other desirable principle can also be shown to be out of reach for **TB** by similar arguments. Another issue is that the general T-schema has perhaps been restricted too much (by only allowing arithmetical formulas in it) in order for **TB** to be consistent. By being typed it avoids the liar-paradox, but also does not allow us to express any higher-order truths, i.e. sentences of the form $T(\ulcorner T(\varphi)\urcorner)$, in the language. In natural language however, this does occur, and usually unproblematically so. If one's aim is to capture part of the role the truth predicate plays in our language, the typing restriction is unsatisfying.

A classic result due to McGee [McG92] offers us a Pyrrhic solution to both issues. The motivation is the following: while the unrestricted T-schema, applying to sentences in L_T, is inconsistent, and the T-schema of **TB** too restricted, perhaps a different class of sentences to instantiate the T-schema with might do the trick.

Lemma 4.2.4. For an arbitrary L_T sentence φ, it holds that there exists a sentence ψ such that $\mathbf{PAT} \vdash \varphi \leftrightarrow (T(\ulcorner\psi\urcorner) \leftrightarrow \psi)$.

Proof. Define $\gamma(x) := T(x) \leftrightarrow \varphi$. By the diagonal lemma there exists a sentence ψ such that $\psi \leftrightarrow \gamma(\ulcorner\psi\urcorner)$. Hence $\psi \leftrightarrow (T(\ulcorner\psi\urcorner) \leftrightarrow \varphi)$ and the result follows. □

Theorem 4.2.5. Let Δ be a set of **PAT**-consistent sentences. Then there is a set of instances of the T-schema Γ such that:

1. $\Gamma \vdash_{\mathbf{PAT}} \Delta$.

2. Γ is **PAT**-consistent.

3. Γ is maximal.

4. $\Gamma \cup \mathbf{PAT}$ is complete.

Proof. By the previous lemma, for any sentence $\varphi \in \Delta$, there exists a T-schema instance $T(\ulcorner \psi_\varphi \urcorner \leftrightarrow \psi_\varphi)$ which is provably equivalent to it. By the consistency of Δ, the set $\Gamma_0 := \{T(\ulcorner \psi_\varphi \urcorner) \leftrightarrow \psi_\varphi \mid \varphi \in \Delta\}$ is consistent over **PAT**.

Now consider the set $S := \{\Theta \mid \Gamma_0 \subseteq \Theta$ and Θ is consistent over $\mathbf{PAT}\}$, where Θ consists only of instances of the T-schema. Since S is a family of sets ordered by partial inclusion, it consists of chains, that is, totally ordered subsets of S. Each of these chains has a maximal element, namely the union of all sets within the chain (being consistent still, and consisting only of instances of the T-schema). Hence Zorn's lemma applies, and S has at least one maximal element Γ.

Clearly the first three conditions of the theorem are satisfied by Γ. It remains to show the fourth condition. Take any $\varphi \in L_T$. Since Γ is consistent, either $\Gamma \cup \mathbf{PAT} \cup \{\varphi\}$ or $\Gamma \cup \mathbf{PAT} \cup \{\neg\varphi\}$ is consistent. So either $\Gamma \cup \mathbf{PAT} \cup \{T(\ulcorner \psi_\varphi \urcorner) \leftrightarrow \psi_\varphi\}$ or $\Gamma \cup \mathbf{PAT} \cup \{T(\ulcorner \psi_{\neg\varphi} \urcorner) \leftrightarrow \psi_{\neg\varphi}\}$ is consistent. Since Γ is maximal, exactly one of these biconditionals is an element of Γ. But then either $\Gamma \cup \mathbf{PAT} \vdash \varphi$ or $\Gamma \cup \mathbf{PAT} \vdash \neg\varphi$ □

Superficially, the theorem seems to be the solution to our troubles with **TB**. Any desired truth-theory — including for example the principle of the excluded middle, and compositionality principles — is equivalent to a set of instances of the T-schema. We can even be greedy and expand this set to a maximal set of instances, so that unlike with **TB**, where the T-schema is restricted to sentences of L_{PA}, we have not excised any biconditional unless strictly necessary. And finally, no typing restrictions are in place.

But it is a Pyrrhic victory for the disquotationalist. The disquotationalist takes the T-schema to be natural and basic: it is the full explication of our notion of truth, which ought to derive the principles of truth we hold to be evident. *Choosing* a desired truth theory, and deriving from that the specific set of T-schema substitutions we must accept, amounts to putting the cart before the horse. Moreover, many of the mutually incompatible maximal sets of Tarski-biconditionals seem to be anything but a natural theory of truth. A set of false arithmetical sentences will also be entailed by some maximal set of biconditionals. If anything then, the T-schema itself is not basic. It is up to the disquotationalist to argue why a given set of instances of the T-schema is basic without justifying the choice by alluding to desired principles.

4.3 Compositional Truth

If the previous section makes one thing clear, it is that the disquotationalist intuition, as formalized by **TB**, is insufficient to capture completely what we mean by truth. This section will be about a different approach. Rather than aiming to fulfill the material adequacy condition directly by adding T-schema instances, we take inspiration from **TB**'s failure to derive plausible truth-theoretic principles. Similar to Tarski's definition of truth in a model, we instead add the evident truth-theoretic principles to **PAT**, resulting in the *compositional truth theory* **CT** .

Definition 4.3.1. **CT** is the truth-theory consisting of:

CT-1 Ax(**PAT**);

CT-2 $\forall s, t \in Tm^c : T(\ulcorner s = t \urcorner) \leftrightarrow Val(s) = Val(t)$;

CT-3 $\forall \varphi \in Sent_{PA} : T(\ulcorner \neg\varphi \urcorner) \leftrightarrow \neg T(\ulcorner \varphi \urcorner)$;

CT-4 $\forall \varphi, \psi \in Sent_{PA} : T(\ulcorner \varphi \wedge \psi \urcorner) \leftrightarrow T(\ulcorner \varphi \urcorner) \wedge T(\ulcorner \psi \urcorner)$;

CT-5 $\forall v, \varphi(x)\{Sent_{PA}(\ulcorner \forall v \varphi(v) \urcorner) \rightarrow [T(\ulcorner \forall v \varphi(v) \urcorner) \leftrightarrow \forall x T(\ulcorner \varphi(\dot{x}) \urcorner)]\}$.

As a reminder, in the last axiom we have used Feferman's dot notation :

$$\forall x T(\ulcorner \varphi(\dot{x}) \urcorner) := \forall x \exists n, z[n = Numeral(x) \wedge z = Sub(\ulcorner \varphi(x) \urcorner, x, n) \wedge T(z)]$$

Notice that in this way x is truly a free variable in $\ulcorner \varphi(\dot{x}) \urcorner$, whereas $\ulcorner \varphi(x) \urcorner$ would merely be a numeral. The following dual sentences to the axioms **CT-4** and **CT-5** can be proved straightforwardly making use of axiom **CT-3**:

CT-4' $\forall \varphi, \psi \in Sent_{PA} : T(\ulcorner \varphi \vee \psi \urcorner) \leftrightarrow T(\ulcorner \varphi \urcorner) \vee T(\ulcorner \psi \urcorner)$;

CT-5' $\forall v, \varphi(x)\{Sent_{PA}(\ulcorner \exists v \varphi(v) \urcorner) \rightarrow [T(\ulcorner \exists v \varphi(v) \urcorner) \leftrightarrow \exists x T(\ulcorner \varphi(\dot{x}) \urcorner)]\}$.

As it turns out, **CT** is consistent by virtue of the existence of the same model we saw earlier:

Theorem 4.3.1. **CT** has a model, namely $\mathcal{M} := \langle \mathbb{N}, \mathcal{E} = \{\ulcorner \varphi \urcorner \mid \mathbb{N} \models \varphi$, and $\varphi \in Sent_{PA}\}\rangle$, where $\mathcal{E}$ is the extension of the truth predicate.

Corollary 4.3.1.1. **CT** is consistent.

We saw earlier by way of Theorem 4.2.3 that the truth-compositional principles of **CT** are not derivable in **TB**. We have framed this problem as **TB** lacking the requisite 'truth-theoretic' strength . By definition **CT** is truth-theoretically strong, and as the following theorem and its corollary ([Hal14, p.66]) go to show, also satisfies the material adequacy condition:

Theorem 4.3.2. For each formula $\varphi(\bar{x})$:

$$\mathbf{CT} \vdash \forall \bar{x}[T(\ulcorner \varphi(\dot{\bar{x}}) \urcorner) \leftrightarrow \varphi(\bar{x})].$$

Proof. By induction on the complexity of $\varphi(\bar{x})$.

Base case: Consider the case where $\varphi(\bar{x})$ is of the form $t_1(\bar{x}) = t_2(\bar{x})$. We need to show that:

$$\mathbf{CT} \vdash \forall \bar{x}[T(\ulcorner t_1(\dot{\bar{x}}) = t_2(\dot{\bar{x}}) \urcorner) \leftrightarrow t_1(\bar{x}) = t_2(\bar{x})].$$

By the definition of the Feferman dot notation, this is equivalent to:

$$\mathbf{CT} \vdash \forall \bar{x}[T(Sub(\ulcorner t_1(\bar{x}) = t_2(\bar{x}) \urcorner, \ulcorner \bar{x} \urcorner, Numeral(\bar{x}))) \leftrightarrow t_1(\bar{x}) = t_2(\bar{x})].$$

Now note that $Sub(\ulcorner t_1(\bar{x}) \urcorner, \ulcorner \bar{x} \urcorner, Numeral(\bar{x}))$ is the code of a closed term, so that by **CT-1** the previous equation is equivalent to:

$$\mathbf{CT} \vdash \forall \bar{x}[Val(Sub(\ulcorner t_1(\bar{x}) \urcorner, \ulcorner \bar{x} \urcorner, Numeral(\bar{x}))) = Val(Sub(\ulcorner t_2(\bar{x}) \urcorner, \ulcorner \bar{x} \urcorner, Numeral(\bar{x})))$$
$$\leftrightarrow t_1(\bar{x}) = t_2(\bar{x})].$$

It is a theorem of **PA** (Cfr. [Kay91, p.121],[HP98, p.55]) that:

$$\forall \bar{x}[Val(Sub(\ulcorner t(\bar{x}) \urcorner, \ulcorner \bar{x} \urcorner, Numeral(\bar{x}))) = t(\bar{x})].$$

It follows that the base-case is equivalent to

$$\mathbf{CT} \vdash \forall \bar{x}[t_1(\bar{x}) = t_2(\bar{x}) \leftrightarrow t_1(\bar{x}) = t_2(\bar{x})],$$

which holds trivially.

Negation case: By axiom **CT-2**, and the fact that

$$\mathbf{PA} \vdash \forall \bar{x}[Sent_{PA}(\ulcorner \varphi(\dot{\bar{x}}) \urcorner)],$$

it holds that:

$$\mathbf{CT} \vdash \forall \bar{x}[T(\ulcorner \neg\varphi(\dot{\bar{x}}) \urcorner) \leftrightarrow \neg T(\ulcorner \varphi(\dot{\bar{x}}) \urcorner)].$$

Since $T(\ulcorner \varphi(\dot{\bar{x}}) \urcorner)$ is of lower complexity, by the induction hypothesis, and propositional logic:

$$\mathbf{CT} \vdash \forall \bar{x}[T(\ulcorner \varphi(\dot{\bar{x}}) \urcorner) \leftrightarrow \varphi(\bar{x})]$$
$$\mathbf{CT} \vdash \forall \bar{x}[\neg T(\ulcorner \varphi(\dot{\bar{x}}) \urcorner) \leftrightarrow \neg\varphi(\bar{x})]$$

From which it follows that indeed:

$$\mathbf{CT} \vdash \forall \bar{x}[T(\ulcorner \neg\varphi(\dot{\bar{x}}) \urcorner) \leftrightarrow \neg\varphi(\bar{x})].$$

Conjunction case: By axiom **CT-3**, and the fact that

$$\mathbf{PA} \vdash \forall \bar{x}[Sent_{PA}(\varphi(\dot{\bar{x}}))],$$

it holds that:

$$\mathbf{CT} \vdash \forall \bar{x}[T(\ulcorner \varphi(\dot{\bar{x}}) \wedge \psi(\dot{\bar{x}}) \urcorner) \leftrightarrow T(\ulcorner \varphi(\dot{\bar{x}}) \urcorner) \wedge T(\ulcorner \psi(\dot{\bar{x}}) \urcorner)].$$

Since both $T(\ulcorner \varphi(\dot{\bar{x}}) \urcorner)$ and $T(\ulcorner \psi(\dot{\bar{x}}) \urcorner)$ are of lower complexity, the induction hypothesis can be applied, so that:

$$\mathbf{CT} \vdash \forall \bar{x}[T(\ulcorner \varphi(\dot{\bar{x}}) \wedge \psi(\dot{\bar{x}}) \urcorner) \leftrightarrow \varphi(\bar{x}) \wedge \psi(\bar{x})].$$

Quantification case: We want to show that:

$$\mathbf{CT} \vdash \forall \bar{x}[T(\ulcorner \forall v \varphi(v, \dot{\bar{x}}) \urcorner) \leftrightarrow \forall v \varphi(v, \bar{x})].$$

By the induction hypothesis, it holds that:

$$\mathbf{CT} \vdash \forall \bar{x} \forall v[T(\ulcorner \varphi(\dot{v}, \dot{\bar{x}}) \urcorner) \leftrightarrow \varphi(v, \bar{x})].$$

This is equivalent to:

$$\mathbf{CT} \vdash \forall \bar{x}[\forall v T(\ulcorner \varphi(\dot{v}, \dot{\bar{x}}) \urcorner) \leftrightarrow \forall v \varphi(v, \bar{x})].$$

By axiom **CT-5**, it holds that:

$$\mathbf{CT} \vdash \forall \bar{x}[T(\ulcorner \forall v \varphi(v, \dot{\bar{x}}) \urcorner) \leftrightarrow \forall v T(\ulcorner \varphi(\dot{v}, \dot{\bar{x}}) \urcorner)].$$

Combining the previous two equations we indeed obtain:

$$\mathbf{CT} \vdash \forall \bar{x}[T(\ulcorner \forall v \varphi(v, \dot{\bar{x}}) \urcorner) \leftrightarrow \forall v \varphi(v, \bar{x})].$$ □

Corollary 4.3.2.1. For all $\varphi \in Sent_{PA} : \mathbf{CT} \vdash \varphi \leftrightarrow T(\ulcorner \varphi \urcorner)$.

After this proof, the reader might have developed a specific discomfort with **CT**. We have introduced truth deflationism as a doctrine which holds that the concept of truth is 'metaphysically light', meaning that our theory of truth should not make any ontological claims. In the context of arithmetic, it would be a mistake to claim that our best theories of truth are intended to capture truth in the standard model. The whole reason of explicating the concept of truth using an axiomatic theory of truth is to give the rules of how truth functions as a logico-linguistic element, in an a priori model-agnostic way. Yet, as now should have become clear, the **CT-5** axiom prima facie seems to do more. Essentially, the axiom says that if for every *numeral* n, $\varphi(n)$ is true (in the sense of our truth theory), then we can conclude that $\forall x \varphi(x)$ is true. Semantically however, $\forall x \varphi(x)$ is true in a given model if $\varphi(x)$ is true of every *object* in the model. By the existence of non-standard models these are clearly two different claims. Is this not a way of making the concept of truth take on metaphysical weight? I think this is not the case. Peano arithmetic is clearly a theory about the natural numbers, even if it fails to single out the natural numbers as a model. Similarly, the truth-theory **CT** takes the domain of discourse to be the natural numbers. On the linguistic level, we intend to talk about properties of the natural numbers, and the truth thereof. Of course, **CT** is equally unable to single out the natural numbers as model, as the following lemma shows:

Lemma 4.3.3. There exist non-standard models of **PA** that can be expanded to a model of **CT**.

Proof. Since $\langle \mathbb{N}, T = \{\ulcorner\varphi\urcorner \mid \mathbb{N} \models \varphi, \text{ and } \varphi \in Sent_{PA}\}\rangle$ is a model of **CT**, by upwards Löwenheim-Skolem there must be a model of larger cardinality, say $M = \langle \mathcal{M}, T_M\rangle$ for which $M \models$ **CT**. Since **CT** $\vdash$ **PA**, this implies that $\mathcal{M} \models$ **PA**. But by definition $\mathcal{M} \neq \mathbb{N}$ so that $\mathcal{M}$ must be a non-standard model of **PA**. It follows that there are non-standard models of **PA** that can be expanded to **CT**. □

In conclusion, I think there is no need to worry for the truth deflationist about **CT**, at least in this regard. Our discourse is always a discourse about something, and in the case of **PA**, about the natural numbers. All we did by introducing **CT** is to make this even more explicit. But we remain firmly planted in language, and do not make the mistake of claiming that truth ought to simply be arithmetical truth.

Having hopefully laid any worries about **CT** to rest, we can say that **CT** is in many ways an attractive truth theory, capturing both our intuitive notions surrounding 'truth-theoretic' reasoning, and fulfilling the material adequacy condition. It was championed, if never made explicit, by Davidson in a series of papers kicked of with 'Truth and Meaning' [Dav67], and at least at some point by Field [Fie99]. So far the commitment to the 'metaphysical lightness' of truth has been explicated by reference to the material adequacy condition, but we have refrained from arguing what truth might still be *for*. This will be the topic of the next section.

4.4 Truth and its Uses

The disquotationalist endorsing **TB** is committed to explicating truth-deflationism by the Tarski-biconditionals. The Tarski biconditionals serve to quote, as when the assertion of a sentence φ is replaced by the assertion of its truth $T(\ulcorner\varphi\urcorner)$, and to disquote, when the direction is reversed. This does not mean that the disquotationalist subscribes to the eliminability of the truth predicate: truth is not redundant. A first example of how the truth predicate is useful is that of *blind ascription*. In blind ascription, the truth predicate allows one to express the content of a sentence(s), without necessarily articulating, or even being able of articulating the sentence(s) itself. As is often the case, Ramsey was one of the first to make this observation:

> In the second case in which the proposition is described and not given explicitly, we have perhaps more of a problem, for we get statements from which we cannot in ordinary language eliminate the words "true" and "false." Thus if I say "he is always right" I mean that the propositions he asserts are always true, and there does not seem to be any way of expressing this without using the word "true." [Ram27]

Blind ascriptions sometimes serve to cope with epistemic limitations, or laziness, such as when I claim that what Ramsey said about blind ascriptions is true, without being able or wanting to exactly reproduce what it is that he said. In this and other cases, the truth predicate only ranges over a finite set of sentences. On the other hand, the truth predicate can range over an infinite collection of sentences when it is used in *generalizations*. This particular usage has first been endorsed by Quine:

> We can generalize on 'Tom is mortal', 'Dick is mortal', and so on, without talking of truth or of sentences; we can say 'All men are mortal'. We can generalize similarly on 'Tom is Tom', 'Dick is Dick', '0 is 0', and so on, saying 'Everything is itself'. When on the other hand we want to generalize on 'Tom is mortal or Tom is not mortal', 'Snow is white or snow is not white', and so on, we ascend to talk of truth and of sentences, saying 'Every sentence of the form 'p or not p' is true', or 'Every alternation of a sentence with its negation is true'. [Qui86]

In contrast with the case of blind ascriptions, rather than laziness, a genuine limitation is involved. Truth increases our expressive power by allowing us to condense the infinitely many instances of the schema 'p or not p' in the single sentence 'Every sentence of the form 'p or not p' is true'. At least, that is the idea. Formally, the truth predicate needs to be understood within the context of the truth-theory in which it occurs. As we have see in theorem 4.2.3, **TB** is unable to prove these kinds of generalization, whereas in **CT** these principles are taken to be axioms. So, at least along this instrumental axis, **CT** wins out over **TB**.

As pointed out by both Shapiro [Sha98] and Ketland [Ket99], a third purpose of truth-theories is to support our *truth-theoretic reasoning*. The salient example for our purposes is reasoning about **PA** as follows:

1. The axioms of **PA** and logic are true.
2. The rules of inference in a given proof-system for **PA** preserve truth.
3. Hence, all theorems of **PA** are true.

Clearly, this amounts to an argument for the *global reflection principle* $Pr_{PA}(\ulcorner\varphi\urcorner) \rightarrow T(\ulcorner\varphi\urcorner)$. By the T-schema, this means that we can conclude the local reflection principle

$Rfn(PA)$ as well. Now consider $Con(Pr_{PA})$. As we have seen in Theorem 2.3.2, even $Rfn_{\Pi_1}(Th)$ suffices to derive $Con(Pr_{PA})$. Hence our truth-theoretic reasoning leads us to accept the consistency of **PA**. But here's the rub: by Gödel's second incompleteness theorem we know that **PA** does not prove its own consistency. So whichever truth-theory will support this kind of reasoning will have to go beyond **PA** in arithmetical power, that is, the truth-theory in question is *not conservative* over **PA**. In so far as one is willing to understand the 'metaphysical lightness' of truth as not going beyond the base theory, such a truth-theory is not deflationist. In the words of Shapiro:

> To be sure, there is no consensus today on the interlocked notions of logical consequence, semantic content, metaphysical strength, and metaphysical possibility; and some philosophers are convinced that most of these notions are obscure and bankrupt. This might give the deflationist some room to maneuver, and it leaves the issue hard to adjudicate. Nevertheless, it seems that in some sense or other, the deflationist is committed to the conservativeness of truth. Deflationism presupposes that there is some sense of 'consequence' according to which truth is conservative. [Sha98]

The question is of course whether this informal discussion is reflected in the truth-theories we have so far considered. As it turns out, this is in fact the case, with **CT** being able to formally capture the truth-theoretic reasoning we have given informally. The next chapter is devoted to exploring these *non-conservativity* results.

CHAPTER 5

Conservativeness

Two notions of conservativity have been argued for as desirable for a truth theory to have. We first define these in full generality:

Definition 5.0.1. Let $\mathbf{Th_1}$ and $\mathbf{Th_2}$ be theories defined in languages $L_1 \subseteq L_2$.

- $\mathbf{Th_2}$ is **syntactically conservative** over $\mathbf{Th_1}$ if and only if $\mathbf{Th_1} \subseteq \mathbf{Th_2}$ and for every $\varphi \in L_1$: $\mathbf{Th_2} \vdash \varphi$ implies that $\mathbf{Th_1} \vdash \varphi$.
- $\mathbf{Th_2}$ is **semantically conservative** over $\mathbf{Th_1}$ if and only if every model of $\mathbf{Th_1}$ can be expanded to a model of $\mathbf{Th_2}$.

These notions do not coincide. Although semantic conservativity implies syntactic conservativity, the converse does not hold. A simple example not involving truth serves to illustrate this fact [Cie17].

Example 5.0.1. Consider the language L^c_{PA}, the extension of the language L_{PA} by a newly introduced constant symbol c. The theory $\mathbf{PA^c}$ is then axiomatized in L^c_{PA} as follows:

$$Ax(\mathbf{PA^c}) = Ax(\mathbf{PA}) \cup \{c \neq \underline{n} \mid n \in \mathbb{N}\}$$

The theory $\mathbf{PA^c}$ is syntactically, but not semantically conservative over $\mathbf{PA}$.

Proof. Assume for a contradiction that $\mathbf{PA^c} \vdash \psi$ but not $\mathbf{PA} \vdash \psi$. Then $\mathbf{PA} + \neg\psi$ is consistent. By this consistency $\mathbf{PA} \cup \{\neg\psi\}$ has a model M. We will now show that $\mathbf{PA^c} \cup \{\neg\psi\}$ has a model as well. Consider an arbitrary finite subset S of $\mathbf{PA^c} \cup \{\neg\psi\}$. Since S is finite it contains a formula $c \neq \underline{n}_{max}$ for some largest n_{max}. Hence, by interpreting c as the natural number $n_{max} + 1$, M is a model for S. By compactness also $\mathbf{PA^c} \cup \{\neg\psi\}$ has a model, contradicting the assumption.

To show that $\mathbf{PA^c}$ is not semantically conservative over **PA**, it suffices to observe that there is no n in the standard model which could interpret c such that each formula of $\{c \neq \underline{\mathrm{n}} \mid n \in \mathbb{N}\}$ is modelled. □

5.1 Disquotational Truth

A first observation, due to Halbach [Hal01b], is that **TB**, and a fortiori **CT**, is not even conservative over first-order logic, that is, over an empty base theory.

Theorem 5.1.1. **TB** is not syntactically conservative over first-order logic with identity.

Proof. Consider the sentences $\forall x(x = x)$ and $\forall x(x \neq x)$. Since the background theory is just first-order logic with identity, unlike in **PA** we do not have the means to perform coding of sentences within the background theory. If we nevertheless add a Tarski-biconditional for each sentence, with a newly introduced constant symbol c_i for each sentence (the 'name' of the sentence), we have the following corresponding Tarski-biconditionals:

- $T(c_1) \leftrightarrow \forall x(x = x)$
- $T(c_2) \leftrightarrow \forall x(x \neq x)$

Since we can derive $\forall x(x = x)$, and $\forall x(x \neq x)$ is refutable, it follows that $T(c_1)$ and $\neg T(c_2)$. But this implies that $c_1 \neq c_2$ and hence that $\exists x \exists y(x \neq y)$. Since this is not a theorem of first-order logic with identity, **TB** is not conservative. □

We see that at least over pure logic our truth-theories are ontologically productive, even if minimally so, by proving the existence of two objects. However, it is difficult to argue why this notion of conservativity should be relevant for truth deflationists. As noted in the proof, pure logic with identity is not even expressive enough to formalize syntactical notions. It is hard to see what the truth theory **TB** over logic means if we can't even express that truth or falsity are properties of sentences, rather than freshly introduced constant symbols. Moreover, any of the settings in which we would like to have a theory of truth, be it a scientific field, natural language, or even the test-bed of **PA**, will already entail the existence of at least two objects. Of more interest are the conservativity results over a sufficiently expressive theory such as **PA**.

At least on one account we have reason to be satisfied as deflationist: **TB** is a syntactically conservative truth theory over **PA**.

Theorem 5.1.2. **TB** is syntactically conservative over **PA**.

Proof. Assume $\mathbf{TB} \vdash \varphi$ where $\varphi \in L_{PA}$, with a derivation d. Since d is finite, there is an upper bound to the complexity of the formulas occurring in d. Take n to be the natural number given by $1 + Max(\{Rank(\psi) \mid [T(\ulcorner\psi\urcorner) \leftrightarrow \psi] \in d\})$. Construct the derivation d' by replacing every occurrence of the truth predicate by a partial truth predicate of rank n. This derivation d' is a proof within **PA**, and the conclusion φ remains the same, so that also $\mathbf{PA} \vdash \varphi$. □

The stricter condition of semantic conservativity is not fulfilled by **TB**, as shown independently by Cieśliński and Ëngstrom (unpublished) [Cie17][Str13]. First we define the notion of a prime model.

Definition 5.1.1. For a model $M \models \mathbf{PA}$, the prime model $K(M)$ is the model whose universe consists of all elements a for which a formula $\varphi(x) \in L_{PA}$ exists such that $M \models \exists! x\varphi(x)$, and a is this unique element. These elements are also called the elements definable in M.

An important idea we will need is the coding of a set by an element of a model. Coded sets play an important role in the model theory of **PA**, see for example [Kay91, Chapter 11].

Definition 5.1.2. Let $M \models \mathbf{PA}$, and let $a \in M$. Then the set coded by a in M is the set $S \subseteq \mathbb{N} := \{n \in \mathbb{N} \mid M \models prime(n)|a\}$, where $prime(n)$ is the function mapping n to the n^{th} prime.

The previous definition justifies the condensed notation $x \in a$, standing for '$M \models prime(x)|a$.'

The following lemma is the stepping stone to showing the non-conservativity of **TB**. [Cie17, p.97-98]

Lemma 5.1.3. The following conditions are equivalent for a non-standard model M of **PA**:

1. The model M can be expanded to a model of **TB**.
2. The set of true arithmetical sentences in M, $Th(M)$, is coded in M.

Proof. For the direction from 2 to 1, assume that a is the specific code of $Th(M)$ in M. Then expand M to M' by defining the introduced truth predicate $T := \{x \mid x \in M \text{ and } M \models x \in a\}$. Then $(M, T) \models \mathbf{TB}$. In particular, as T is defined by a formula with parameter a in M, it is inductive.
For the opposite direction assume we have model M which can be expanded to a model M' of **TB**. For every finite natural number k, we can code the set of true arithmetical sentences in M' with code smaller than k:

$$\text{For all } k \in \mathbb{N} : M' \models \exists s \forall x[x \in s \leftrightarrow (x < k \wedge T(x)]$$

By the Overspill Lemma (Lemma 2.1.2), there exists a non-standard $a \in M'$ – and hence also $a \in M$ – such that

$$\text{For all } k \in \mathbb{N} : M' \models \exists s \forall x[x \in s \leftrightarrow (x < a \wedge T(x)]$$

Picking such a particular s we have found the code for $Th(M)$ in M, since every true formula is coded as a natural number, and for every natural number n: $a > n$. □

Theorem 5.1.4. **TB** is not semantically conservative over **PA**.

Proof. Consider an arbitrary non-standard model M of **PA**. Then $K(M)$, the prime model of M, cannot be expanded to a model of **TB**. Assume for a contradiction that such an expansion exists. Then by Lemma 5.1.3 there is an $a \in K(M)$ such that a codes $Th(K(M))$. Since $K(M)$ is a prime model, there is a formula ψ such that $\mathbf{PA} \vdash \exists! x \psi(x)$ and $M \models \psi(a)$, that is, a is definable in M. So, we can define a truth predicate as follows:

$$T(x) := \exists z[\psi(z) \wedge x \in z]$$

This truth predicate applies to all $\varphi \in L_{PA}$. But the existence of this truth predicate is in contradiction with the semantic Undefinability Theorem. □

5.2 Compositional Truth

The situation is even more drastic for **CT**: even syntactical conservativity is out of the question. First we show that **CT** proves *the global reflection* principle [Hal14, p. 91-92]. Following custom in the literature, we will dispense with corner quotes in order to keep the formulas easy to parse. It should still be clear in context when formulas or the codes of formulas occur.

Theorem 5.2.1.

$$\mathbf{CT} \vdash \forall \varphi \in Sent_{PA} : Pr_{PA}(\varphi) \rightarrow T(\varphi)$$

Proof. For the purposes of the proof, it is convenient to assume that our first-order logic is axiomatized by a Hilbert-style system, with only modus ponens and generalization as inference rules. Except for the instances of the induction schema, there are only finitely many axioms of **PA**. By Corollary 4.3.2.1, for each φ of these axioms $\mathbf{CT} \vdash T(\ulcorner \varphi \urcorner)$. Next, consider the following instance of the induction schema of **CT**:

$$[T(Sub(x,v,0)) \wedge \forall y(T(Sub(x,v,\dot{y})) \rightarrow T(\varphi(Sub(x,v,S(\dot{y})))))] \rightarrow \forall y T(Sub(x,v,\dot{y}))$$

By applying the quantifier axiom **CT-5** (notice that this leads to the conditional on $Sent_{PA}(\forall x \varphi(x))$), as well as **CT**-2, and **CT**-4, we have within CT:

$$Sent_{PA}(\forall vx) \rightarrow [T(Sub(x,v,0)) \wedge T(\forall v(x \rightarrow Sub(x,v,S(v))))] \rightarrow T(\forall vx).$$

Now we can bring out the truth predicate,

$$Sent_{PA}(\forall vx) \rightarrow T\{[Sub(x, v, 0) \wedge \forall v(x \rightarrow Sub(x, v, S(v)))] \rightarrow \forall vx\},$$

which shows that in **CT** all closed instances of the induction schema are true. Note that it was instrumental for the derivation that the truth predicate is inductive. Since we work in a Hilbert-style axiomiatization, several logical axioms have to be shown to be true. We consider the axiom $\varphi \rightarrow (\psi \rightarrow \varphi)$ by way of example. Start with the following instance, with both φ and ψ variables rather than fixed formulas:

$$T(\varphi) \rightarrow (T(\psi) \rightarrow T(\varphi))$$

by propositional reasoning we derive in **CT**:

$$\neg T(\varphi) \vee (\neg T(\psi) \vee T(\varphi))$$

By application of the compositional truth axioms:

$$Sent_{PA}(\varphi) \wedge Sent_{PA}(\psi) \rightarrow T(\varphi \rightarrow (\psi \rightarrow \varphi))$$

Since both φ and ψ are variables, we obtain by two applications of generalisation the desired result within **CT**:

$$\forall\varphi\forall\psi[Sent_{PA}(\varphi) \wedge Sent_{PA}(\psi) \rightarrow T(\varphi \rightarrow (\psi \rightarrow \varphi))]$$

The other logical axioms are similarly shown to be true.

We now prove the theorem by induction on proof length, in the sense of Pr_{PA}. For this purpose, we introduce a provability predicate $Prv(x, y)$, expressing that y is provable with a proof with length at most x. Since we are interested in sentences only, and proofs in **PA** can contain open formulas, we have to be a bit careful in considering only the universal closure of these formulae. For this we will use the – provably recursive in **PA** – universal closure function $Cl(\varphi)$ which yields a sentence only if φ is an open formula (and for the sake of definiteness, 0 otherwise). Now consider the following instance of the induction axiom:

$$\begin{aligned}
&\forall\varphi(Prv_{PA}(0, \varphi) \rightarrow T(Cl(\varphi))) \wedge \\
&\forall x[\forall\varphi(Prv_{PA}(x, \varphi) \rightarrow T(Cl(\varphi))) \rightarrow \forall\varphi(Prv_{PA}(S(x), \varphi) \rightarrow T(Cl(\varphi)))] \rightarrow \\
&\forall x\forall\varphi(Prv_{PA}(x, \varphi) \rightarrow T(Cl(\varphi)))
\end{aligned}$$

The base case of the induction instance has been provided for by showing all the axioms to be true within **CT**. What remains is to show that inference preserves truth. Since we conveniently have only modus ponens and generalisation as inference rules, it remains to show that these two rules preserve truth. The case of generalisation is trivial, since by induction hypothesis if **CT** $\vdash \varphi(x)$, then already **CT** $\vdash T(Cl(\varphi(x))$. For modus ponens we apply the compositional axioms of **CT** and derive in **CT**:

$$\begin{aligned}\forall\varphi,\psi[&Sent_{PA}(Cl(\varphi)) \wedge Sent_{PA}(Cl(\varphi \to \psi))\\ &\to T(Cl(\varphi) \wedge T(Cl(\varphi \to \psi))\\ &\to \forall x T(\varphi(\dot{x})) \wedge \forall x T(\varphi(\dot{x}) \to \psi(\dot{x}))\\ &\to \forall x[T(\varphi(\dot{x})) \wedge (\varphi(\dot{x}) \to \psi(\dot{x}))]\\ &\to \forall x[T(\varphi(\dot{x})) \wedge (\neg\varphi(\dot{x}) \vee \psi(\dot{x}))]\\ &\to \forall x T(\psi(\dot{x}))]\end{aligned}$$

Note that we have for simplicity assumed that φ and ψ both have only one free variable, which is the same for both. The general case does not offer any complications. By the previous derivation, the following is a theorem of **CT**:

$$\forall\varphi,\psi[Sent_{PA}(Cl(\varphi)) \wedge Sent_{PA}(Cl(\varphi \to \psi)) \to (T(Cl(\varphi)) \wedge T(Cl(\varphi \to \psi))) \to T(Cl(\psi))],$$

which shows that modus ponens does indeed preserve truth. □

Note that in the proof we used the inductiveness of the truth predicate both to show that **CT** derives that all induction axioms are true, and to formalise that truth is preserved by the inference rules. The proof is nothing more than the formal counterpart of the *truth-theoretic reasoning* discussed at the end of the previous chapter.

Corollary 5.2.1.1. **CT** $\vdash Con(PA)$.

Proof. Instantiate the global reflection principle with $\ulcorner 0 = 1 \urcorner$:

$$Pr_{PA}(\ulcorner 0 = 1 \urcorner) \to T(\ulcorner 0 = 1 \urcorner)$$

Since **PA** $\vdash 0 \neq 1$ we have by Corollary 4.3.2.1, and axiom **CT-3** that also **CT** $\vdash \neg T(\ulcorner 0 = 1 \urcorner)$. But then it follows that $\neg Pr_{PA}(\ulcorner 0 = 1 \urcorner)$, by modus tollens. In other words, **CT** proves the consistency of Peano arithmetic. □

Corollary 5.2.1.2. **CT** is not syntactically conservative over **PA**.

Proof. **CT** proves the consistency of **PA**, which by the second incompleteness theorem is something **PA** cannot do. □

It follows that **CT** is not semantically conservative over **PA**, since non-standard models of **PA** in which $\neg Con(PA)$ hold can not be expanded to a model of **CT**.

5.3 How Conservative is Conservative Enough?

An important question is which notion of conservativity is the most pressing for the truth-deflationist. It has been argued that the stricter notion of semantic conservativity should be accepted as a criterion by which to evaluate truth theories [Str13]. Most authors, including Shapiro and Ketland [Sha98, Ket99], have used the syntactic notion of conservativity, with an explicit rebuttal of semantic conservativity as being appropriate by Cieśliński [Cie17]. It is worth summarizing his rebuttal, as it is insightful for the larger project of axiomatic truth theories.

What is the motivation for requiring semantic conservativity of our truth theories? In the words of Strollo:

> Not only would we operate at a linguistic level by adding a [truth predicate] and interpreting it with a suitable extension, we would also need to intervene into the domain by changing and shaping it. In this sense, and in open contrast with the deflationist claim, the property of truth would enter reality as a robust ingredient and an exhaustive inventory of the world should include mention of such a thing as truth.[Str13]

As Cieśliński observes, the deflationist claim mentioned has to be made more explicit in order to evaluate whether truth enters as a 'robust ingredient'. It is helpful to consider a different approach to truth theories, as exemplified by Kripke's semantic theory of truth [Kri75]. In this theory, the language L_{PAT} is considered, where the truth predicate T is a partial one. It is interpreted by two sets T^+ and T^-, being respectively the extension and anti-extension of T. The predicate T then holds of sentences that belong to T^+, does not hold of sentences belonging to T^-, and is undetermined for all other sentences. Given a model $M = (K, T^+, T^-)$, where K is a model of **PA**, formulas in the language are evaluated by a satisfaction relation $\models_{sk}$ along the lines of strong Kleene logic [Hor11, p.118].

Definition 5.3.1. The satisfaction relation $\models_{sk}$ is defined inductively as follows, given a model $M = \langle U, I, \alpha \rangle$.

- For an atomic formula $P(t_1, \ldots, t_n)$, $M \models_{sk} P(t_1, \ldots, t_n)$ iff $\langle {t_1}^I, \ldots, {t_n}^I \rangle \in P^+$.
- For an atomic formula $P(t_1, \ldots, t_n)$, $M \models_{sk} \neg P(t_1, \ldots, t_n)$ iff $\langle {t_1}^I, \ldots, {t_n}^I \rangle \in P^-$.
- $M \models_{sk} \varphi \wedge \psi$ iff $M \models_{sk} \varphi$ and $M \models_{sk} \psi$.
- $M \models_{sk} \neg(\varphi \wedge \psi)$ iff either $M \models_{sk} \neg\varphi$ or $M \models_{sk} \neg\psi$.
- $M \models_{sk} \forall x \varphi(x)$ iff $M' \models_{sk} \varphi$ for each $M' = \langle U, I, \alpha' \cup \{x \leftarrow c\} \rangle$, with $c \in U$. The variable assignment α' is given by α restricted to the free variables of $\forall x \varphi(x)$.

- $M \models_{sk} \neg\forall x\varphi(x)$ iff $M' \models_{sk} \neg\varphi$ for at least one $M' = \langle U, I, \alpha' \cup \{x \leftarrow c\}\rangle$, with $c \in U$. The variable assignment α' is given by α restricted to the free variables of $\forall x\varphi(x)$.
- $M \models_{sk} \neg\neg\varphi$ iff $M \models_{sk} \varphi$.

Given this satisfaction relation, Kripke singles out the least fixed point of the following monotonic operator:

$$\Gamma(T^+, T^-) = (\{\varphi \in Sent_{PAT} \mid (\mathbb{N}, T^+, T^-) \models_{sk} \varphi\},\\ \{\varphi \in Sent_{PAT} \mid (\mathbb{N}, T^+, T^-) \models_{sk} \neg\varphi\}).$$

Intuitively, the intended model is constructed in stages, where each stage adds the sentences true in the previous model to the extension of the truth predicate, and similarly for the anti-extension. This truth theory has several virtues: it is untyped and is thus able to express higher-order truths, it leaves the liar sentence indeterminate, and in a partial sense the Tarski-biconditionals hold unrestricted in the least fixed-point model.

Horsten gives three arguments for why, despite its virtues, a semantic theory such as Kripke's is unsatisfying [Hor11]. All three of the arguments circle around the property of *universality*: it is desirable for our language (which includes the concept of truth) to be universal, in the sense that every new concept can be defined within the language. No meta-language ought to be necessary, since no meta-language is on offer for natural language. The arguments are:

1. Ideally we want a definition of truth for a natural language. As Tarski's undefinability theorem shows, if we aim for a definitional approach a meta-language is necessary. Which meta-language is available to define the notion of truth in natural language? An axiomatic approach does not suffer from this issue since the truth axioms can be stated in the same language as the language for which truth is characterized.
2. Even if one were to define the notion of truth for a given language by singling out an intended model, or class of models for it, this singling out itself has to occur in a more expressive language. How could one avoid the pressure to also define the notion of truth for the more expansive meta-language – leading to infinite regress?
3. Another problem for the semantic approach is the scope of objects that can be discussed in natural language. In particular, every set belongs to the domain of discourse in natural language. But this implies that the domain of natural language does not itself form a set. Singling out an intended model for a natural language will be impossible.

These three arguments together, while not making the semantic approach untenable outright, do go a long way to explain why most truth-deflationists have, and continue to adopt an axiomatic approach. Of course, this does not mean that in the investigation of truth theories we should not use any model-theoretic reasoning, or results and techniques of model theory. The constraint is only that no models or class of models are singled out as being intended.

Given that one adopts an axiomatic approach, keeping in mind the objections against the semantic approach, does semantic conservativity still strike us as a necessary condition for our truth theories? Cieśliński offers two possible ways of arguing for semantic conservativity, and rebuts them in turn. A first line of argument would go as follows: The truth theories that expand on **PA**, while ostensibly neutral with respect to the models underlying it, ought not to exclude the expansion of the standard model. If they do exclude this expansion, the meaning of our arithmetical terms and sentences would be wrong. We need semantic conservativity to safeguard that the (expanded) standard model is retained. This argument can be dismissed due to referring explicitly to an intended model: one can't both argue for an axiomatic theory, while still wanting one's notion of truth to be truth in the standard model.

The second argument takes the opposite tack: it is precisely to be model-agnostic that one should support semantic conservativity. All models are to be taken on equal footing, and no models should be lost by the addition of a truth predicate and truth axioms. But this seems too strong of a claim as well: there are models which we intuitively take to be 'wrong'. For example, by the second incompleteness theorem there are models of **PA** in which $\neg Con(Pr_{PA})$ holds. Using **PA**, we do take the theory to be consistent, and models in which a generalized consistency statement does not hold strike us as wrong, even if they still model the theory. This does not imply that as deflationists we intend the models which do model the consistency statement as the models of our truth predicate. In axiomatizing a truth theory we take the truth axioms to be sufficient to characterize our notion of truth. If it so happens that certain non-standard models of **PA** are not expandable to a model of this truth theory, this should not be taken as a mark against the theory.

The preceding discussion has shown that semantic conservativity is too strong of a condition to expect our truth-theories to fulfill. We have seen in the previous chapter that Shapiro [Sha98] and Ketland [Ket99] have argued for syntactic conservativity as being at least part of what it means to be a deflationist about truth. If that is so, we know by Theorem 5.2.1, and its corollary, that **CT** is unacceptable to the deflationist. On the other hand, the discussion at the end of the previous chapter made it clear that for the purpose of expressing generalizations and formalizing our informal truth-theoretic reasoning **CT** is indispensable. We are stuck between Chylla and Charybdis.

Which options remain open to the deflationist? A first option is to take the result at face-value and look for an improved or completely different truth-theory. This is unlikely to ever be found, since our informal truth-theoretic reasoning is supposed (among other purposes) to show the consistency of the base-theory. By the incompleteness theorems this implies the non-conservativity of the truth-theory in question.

A second option is to deny that syntactic conservativity is, even partially, the correct explication of deflationism. Both Horsten [Hor11] and Halbach [Hal01b] have argued that non-conservativity over the base theory is intrinsic to the best truth theories on offer. Halbach argues that the deflationist commitment is a commitment to circumscribing the *use* of truth and nothing else. Specifically, truth serves to express and even prove generalizations such as the distribution of truth over conjunction. Nothing more lies within the purview of truth. That a truth-theory like **CT** then engenders certain substantial results is something the deflationist should take on board. Horsten similarly argues that the deflationist understanding of truth is one where truth serves to express certain generalizations. In addition he underlines the "role of the concept of truth as an inferential tool." The concept of truth serves to buttress truth-theoretical reasoning such as the argument for the consistency of **PA**. The comparison with negation he uses as illustration is insightful for understanding his view:

> It is known from the discussion between constructive and classical mathematics that some purely positive arithmetical existence statements can (as far as we know) only be proved by relying on the law of excluded middle. In other words, some mathematical statements not containing the concept of negation can only be proved by making use of the concept of negation. The deflationist emphasizes that the notion of truth is in this respect similar to that of negation. Truth, like education, is not so much a putting in as a drawing out. It helps to draw out implicit commitments of theories that we have postulated. [Hor11, p. 93]

Particularly interesting is the claim that a truth-theory "draws out implicit commitments of theories." Where Horsten seems to suggest that our truth-theories are required to derive (bring out) statements such as the Gödel sentence (G_{PA}) and $Con(Pr_{PA})$, Tennant [Ten02] has argued that truth-theories are in no way needed to make the argument for the truth of G_{PA} or $Con(Pr_{PA})$ comprehensible. Implicit in our acceptance of **PA** lies the acceptance of a reflection principle, the aforementioned implicit commitment of our theory, which allows us to reconstruct the argument with no reference to a truth-predicate. Tennant concludes that a substantial notion of truth is not necessary to ground this particular kind of reasoning, and that it remains open to a truth deflationist to advance conservative theories of truth.

Tennant was happy to show that there is a deflationary licit way to argue for sentences such as G_{PA} by using the implicit commitments we have in the form of an appropriate reflection principle over the base theory. Cieśliński takes the argument further by claiming that our implicit commitment is not just to **PA**, but to the truth-theory [Cie10]. That is, we are not only in a position to accept a reflection principle for **PA**, expressing its soundness, but also a reflection principle which applies to sentences or formulas in the extended language, containing the truth predicate. A third option is thus available to the deflationist: take the conservativity requirement seriously for one's basic truth theory, meaning that the basic truth theory ought to be conservative over the underlying base theory. Similarly as for **PA**, we have an implicit commitment to this basic truth theory by our acceptance of its theorems. Expressing this commitment through the acceptance of a reflection principle will then (hopefully) allow us to recover the deductive power needed to not only express, but also prove as valid certain truth-theoretic generalizations and reasoning.

This third option is the option we intend to explore for the remainder of this thesis. However, the story we have told so far is much too summary to be apt for defending it as viable. In particular, an argument needs to be given as to what the implicit commitment to a theory entails, and how we are justified in accepting a reflection principle through it. As Halbach has pointed out: “the transition from a theory to a reflection principle for that theory requires an argument” [Hal01a]. Before we do that, it is helpful to have seen the lay of land first. The next chapter will state and prove the known technical results on reflection principles for **TB**, **CT**, and its variants.

CHAPTER 6

Reflecting On Truth

In this chapter we will consider the impact of adding suitable reflection principles to truth-theoretically weak theories. These theories have the benefit of being arithmetically conservative over their base-theory **PA**, so that they evade the criticism of not capturing the alleged explication of truth-deflationism. On the other hand, we will see that the addition of reflection principles increases the strength of these weak truth-theories, not only arithmetically, but also truth-theoretically. That is, the theories so obtained are non-conservative over **PA**, and prove compositional truth principles which were underivable in the weak truth-theory.

6.1 Compositional Truth

We will start with the truth theory $\mathbf{CT}^-$, which is the same as the truth theory **CT**, except for the induction schema which is not extended to formulas of $\mathbf{L_{PAT}}$ but is constrained to apply only to arithmetical sentences. From the proof of the syntactic non-conservativity of **CT** (see Theorem 5.2.1), where the induction schema played an important role, we might expect $\mathbf{CT}^-$ to be syntactically conservative over **PA**. This is indeed the case.

Theorem 6.1.1. $\mathbf{CT}^-$ is syntactically conservative over **PA**.

Proving this is far from trivial and we refer the interested reader to the literature. A proof of the syntactic conservativity of a closely related theory due to Kotlarski et al. [KKHL81] in 1981 was convoluted and has recently been improved upon by a model-theoretic argument due to Visser and Enayat [EV15]. A proof of the syntactic conservativity of $\mathbf{CT}^-$ based on cut-elimination is due to Leigh [Lei15].

We have seen in our discussion of reflection principles (see theorem 2.3.2) that even over the restricted class of Π_1-formulas, the addition of both local and uniform reflection principles for Pr_{PA} are sufficient to derive the consistency of **PA**. By the incompleteness theorems this implies that the new theory exceeds **PA** in arithmetical strength. There are two reasons for considering even weaker reflection principles. First, if we are to argue for the philosophical innocence of adding a reflection principle to a weak truth-theory, weaker reflection principles will plausibly be easier to defend. Second, from a logical point of view, it is interesting to develop more understanding of exactly where the boundary from an arithmetically conservative extension to an arithmetically non-conservative extension lies. Speaking loosely, this would allow us to determine exactly what it is in the reflection principle that goes beyond **PA**.

Formally, we will now consider reflection principles where the provability predicate occurring in it corresponds to a weaker theory. The objects of the reflection principles are sentences and formulas in L_{PA}. For example, the local reflection principle which expresses the soundness of first-order logic, is given by the following schema :

$$Rfn(\emptyset) : Pr_{\emptyset}(\ulcorner\varphi\urcorner) \rightarrow \varphi, \qquad \varphi \text{ is a sentence in } L_{PA}.$$

In this formulation $Pr_{\emptyset}$ represents provability from first-order logical axioms only (and *not* the arithmetical axioms). Since we now work in a truth-theory with a well-defined truth predicate, a stronger reflection principle than the uniform reflection principle is open to us, the global reflection principle:

$$GR(Th) : \forall\varphi \in Sent_{PA}[Pr_{Th}(\ulcorner\varphi\urcorner) \rightarrow T(\ulcorner\varphi\urcorner)].$$

Our first result will be that accepting the global reflection principle for first-order logic over $\mathbf{CT}^-$ is already sufficient to go beyond **PA** [Cie10]. The proof is complicated by the fact that in $\mathbf{CT}^-$ it is, unlike **CT**, not the case that truth distributes over conjunction:

$$\mathbf{CT}^- \nvdash \forall\varphi_1, \ldots \varphi_n \in Sent_{PA} : T(\ulcorner\varphi_1 \wedge \cdots \wedge \varphi_n\urcorner) \leftrightarrow T(\ulcorner\varphi_1\urcorner) \wedge \cdots \wedge T(\ulcorner\varphi_n\urcorner)[\text{Cie10}].$$

The reason for this failure is the lack of extended induction for the truth predicate, which is necessary to generalize **CT-4**.

Theorem 6.1.2. $\mathbf{CT}^- + GR(\emptyset) \vdash GR(PA)$.

Proof. Working in $\mathbf{CT}^- + \forall\varphi \in Sent_{PA} : Pr_{\emptyset}(\ulcorner\varphi\urcorner) \rightarrow T(\ulcorner\varphi\urcorner)$, fix a φ such that $Pr_{PA}(\ulcorner\varphi\urcorner)$. We have to show that $T(\ulcorner\varphi\urcorner)$. Assume that d is a proof of φ in **PA** letting $(W, \alpha_0, \ldots \alpha_n)$, be a tuple of the axioms used in d. Here W is the conjunction of all non-inductive axioms of **PA** present in d, and α_i is an instance of the induction schema. It holds that:

$$\emptyset \vdash (W \wedge \alpha_0 \wedge \cdots \wedge \alpha_n) \rightarrow \varphi.$$

By $GR(\emptyset)$ and truth-compositionality in $\mathbf{CT}^-$ it follows that:

$$\mathbf{CT}^- + GR(\emptyset) \vdash T(\ulcorner W \wedge \alpha_0 \wedge \cdots \wedge \alpha_n\urcorner) \rightarrow T(\ulcorner\varphi\urcorner).$$

So it suffices to show that $\mathbf{CT}^- + GR(\emptyset) \vdash T(\ulcorner W \wedge \alpha_0 \wedge \cdots \wedge \alpha_n \urcorner)$ to complete the proof. It is easy to see that similar to Corollary 4.3.2.1, $\mathbf{TB}^-$ is a sub-theory of $\mathbf{CT}^-$, so that $\mathbf{CT}^- \vdash W \leftrightarrow T(\ulcorner W \urcorner)$. Since $\mathbf{CT}^- \vdash W$, it suffices to show that (by axiom **CT-4**):

$$\mathbf{CT}^- + GR(\emptyset) \vdash T(\ulcorner \alpha_0 \wedge \cdots \wedge \alpha_n \urcorner).$$

For a contradiction, assume that $\neg T(\ulcorner \alpha_0 \wedge \cdots \wedge \alpha_n \urcorner)$. Since already in first-order logic:

$$\emptyset \vdash \neg(\alpha_1 \wedge \cdots \wedge \alpha_n) \rightarrow \neg\alpha_1 \vee \cdots \vee \neg\alpha_n, \tag{6.1}$$

by $GR(\emptyset)$ and compositionality it follows that $T(\ulcorner \neg\alpha_0 \vee \cdots \vee \neg\alpha_n \urcorner)$. Assume that each instance α_i of the induction schema corresponds to a formula β_i in the sense that α_i is of the form:

$$\alpha_i = [\beta_i(0) \wedge \forall y(\beta_i(y) \rightarrow \beta_i(y+1))] \rightarrow \forall x \beta_i(x).$$

Denote by $\gamma(x)$ the formula:

$$\begin{gathered}\{[\beta_0(0) \wedge \forall y(\beta_0(y) \rightarrow \beta_0(y+1))] \wedge \neg\beta_0(x)\} \vee \dots \\ \vee \{[\beta_n(0) \wedge \forall y \beta_n(y) \rightarrow \beta_n(y+1))] \wedge \neg\beta_n(x)\}.\end{gathered}$$

Then we have by first-order logic that:

$$\emptyset \vdash (\neg\alpha_0 \vee \cdots \vee \neg\alpha_n) \rightarrow \exists x \gamma(x),$$

and by $GR(\emptyset)$ and compositionality, we have that also $T(\ulcorner \exists x \gamma(x) \urcorner)$. By axiom **CT-5** it follows that $\exists a T(\ulcorner \gamma(\dot{a}) \urcorner)$. However, it holds that for all a, $\emptyset \vdash \neg\gamma(a)$. To see this, notice that by an easy application of induction that for all β_i and for all a:

$$\emptyset \vdash [\beta_i(0) \wedge \forall y(\beta_i(y) \rightarrow \beta_i(y+1))] \rightarrow \beta_i(a).$$

By propositional reasoning, $\neg\gamma(a)$ is equivalent to:

$$\begin{gathered}\{[\beta_0(0) \wedge \forall y(\beta_0(y) \rightarrow \beta_0(y+1))] \rightarrow \beta_0(a)\} \wedge \dots \\ \wedge \{[\beta_n(0) \wedge \forall y(\beta_n(y) \rightarrow \beta_n(y+1))] \rightarrow \beta_n(a)\}.\end{gathered}$$

As every member of this conjunction is provable in logic, also the conjunction itself is provable in logic, so that indeed for all a: $\emptyset \vdash \neg\gamma(a)$. By $GR(\emptyset)$ we have that for all a $\emptyset \vdash \forall a T(\ulcorner \neg\gamma(\dot{a}) \urcorner)$, yielding the desired contradiction. □

Since we know that $\mathbf{CT}^-$ is syntactically conservative over **PA**, and $\mathbf{CT}^- + GR(\emptyset) \vdash GR(PA)$ leads immediately to the non-conservativity over **PA** of $\mathbf{CT}^- + GR(\emptyset)$, it follows that $\mathbf{CT}^- \nvdash GR(\emptyset)$. This means that a conservative compositional truth theory is already too weak to prove the truth of first-order logic.

The addition of a global reflection principle to $\mathbf{CT}^-$ is quite 'stable', in the sense that the theory obtained is robust with regards to which theory the global reflection principle expresses the soundness of. This is the content of the following corollary:

Corollary 6.1.2.1. The following theories are equivalent:

A) $\mathbf{CT}^- + GR(\emptyset)$;

B) $\mathbf{CT}^- + GR(PA)$;

C) $\mathbf{CT}^- + GR(T)$.

We have abused notation a little by using $GR(T)$ to stand for $\forall\varphi \in L_{PA} : [Pr_T(\ulcorner\varphi\urcorner) \to T(\ulcorner\varphi)\urcorner]$, where $Pr_T(\ulcorner\varphi\urcorner)$ is a formula of L_T, meaning informally that 'φ is the last element in a sequence d, where each element is either true (in the sense of the T-predicate), a logical axiom (and *not* an arithmetical axiom), or follows from previous elements by an inference rule'. In other words, $GR(T)$ expresses the notion that whatever is provable from true premises is also true.

Proof. That A) implies B) is the result of theorem 6.1.2. That C) implies A) is evident, since a proof from no premises is also a proof from true premises. It remains to show that B) implies C).

Consider a proof d of ζ from true premises. Working in $\mathbf{CT}^- + GR(PA)$, we want to show that also $T(\ulcorner\zeta\urcorner)$. Assume d has the premises $\{\varphi_0, \ldots, \varphi_n\}$, where by construction of d either $\varphi_i \in LogAx$, that is an axiom of first-order logic, or $T(\ulcorner\varphi_i\urcorner)$ holds. Now since we have $GR(PA)$, all the logical axioms are true, so that we have that for each φ_i, $T(\ulcorner\varphi_i\urcorner)$. On the other hand, since

$$\emptyset \vdash (\varphi_0 \wedge \cdots \wedge \varphi_n) \to \zeta,$$

and logic is true, we have that $T(\ulcorner(\varphi_0 \wedge \cdots \wedge \varphi_n) \to \zeta\urcorner)$. By compositionality it hence suffices to show that $T(\ulcorner\varphi_0 \wedge \cdots \wedge \varphi_n\urcorner)$ to conclude the proof.

Define $\psi(x)$ as:

$$(x = 0 \to \varphi_0) \wedge (x = 1 \to (\varphi_0 \wedge \varphi_1)) \wedge ... \wedge (x = n \to (\varphi_0 \wedge ... \wedge \varphi_n)).$$

We will show that $T(\ulcorner\forall x\psi(x)\urcorner)$, which implies that $T(\ulcorner\psi(n)\urcorner)$, which in turn implies $T(\ulcorner\varphi_0 \wedge \cdots \wedge \varphi_n\urcorner)$.
Since **PA** is true, it is sufficient to show that $T(\ulcorner\psi(0) \wedge \forall x[\psi(x) \to \psi(x+1)]\urcorner)$. Notice that $\psi(x)$ is an arithmetical formula, so that this move is warranted. For the base case, it is easy to see that $T(\ulcorner 0 = 0 \to \varphi_0\urcorner)$. It is also the case that

$$\mathbf{PA} \vdash 0 \neq 1 \wedge \ldots 0 \neq n,$$

whence we know by $GR(PA)$ that $T(\ulcorner 0 \neq 1 \wedge \ldots 0 \neq n\urcorner)$. In addition, it's a first-order theorem that:

$$(0 \neq 1 \wedge \cdots \wedge 0 \neq n) \to [(0 = 1 \to (\varphi_0 \wedge \varphi_1)) \wedge \cdots \wedge (0 = n \to (\varphi_0 \wedge \cdots \wedge \varphi_n))].$$

By $GR(PA)$ and compositionality it follows that $T(\ulcorner\psi(0)\urcorner)$. For the induction step assume that $T(\ulcorner\psi(m)\urcorner)$. It is again a theorem of first-order logic that:

$$\psi(m) \to [m = m \to (\varphi_0 \wedge \cdots \wedge \varphi_m)], \tag{6.2}$$

and $T(\ulcorner\varphi_0 \wedge ... \wedge \varphi_m\urcorner)$ follows. Now we have by compositionality, and the fact that $T(\ulcorner\varphi_i\urcorner)$ for each $\varphi_i \in \{\varphi_0, \ldots, \varphi_n\}$, that $T(\ulcorner(\varphi_0 \wedge ... \wedge \varphi_m) \wedge \varphi_{m+1}\urcorner)$. It follows that:

$$T(\ulcorner m+1 = m+1 \to (\varphi_0 \wedge \cdots \wedge \varphi_m \wedge \varphi_{m+1})\urcorner).$$

Defining

$$\chi := \bigwedge_{i \leq n, m+1 \neq i} m+1 \neq i,$$

we have that, similarly as in the base-case, $\mathbf{PA} \vdash \chi$, and hence, $T(\ulcorner\chi\urcorner)$. As in 6.2, it is a first-order theorem that:

$$\begin{aligned}&\chi \to [(m+1 = 0 \to \varphi_0)\\ &\wedge \ldots (m+1 = m \to (\varphi_0 \wedge \cdots \wedge \varphi_m)) \wedge (m+1 = m+2 \to (\varphi_0 \wedge \cdots \wedge \varphi_{m+2}))\\ &\wedge \ldots (m+1 = n \to (\varphi_0 \wedge \cdots \wedge \varphi_n))].\end{aligned}$$

By $GR(PA)$ and compositionality it follows that $T(\ulcorner\psi(m+1)\urcorner)$. Now, since we have established that $\forall x T(\ulcorner\psi(x)\urcorner)$, it follows that in particular $T(\ulcorner\psi(n)\urcorner)$. It is a first-order theorem that

$$\psi(n) \to (n = n \to (\varphi_0 \wedge \cdots \wedge \varphi_n)),$$

whence it follows that $T(\ulcorner\varphi_0 \wedge \cdots \wedge \varphi_n\urcorner)$ as desired. □

A natural question is the relation of the foregoing 3 equivalent axiomatizations to **CT**. We have seen in theorem 5.2.1 that **CT** proves the global reflection principle $GR(PA)$, which led to its syntactic non-conservativeness. Until recently (for example in [Cie10]), the following result due to Kotlarski [Kot86] was quoted:

Theorem 6.1.3.

$$\mathbf{CT}^- + GR(PA) = \Delta_0\text{-}\mathbf{CT}.$$

In the theorem statement, Δ_0-**CT** stands for $\mathbf{CT}^-$ with the addition of induction in the extended language for Δ_0-formulas only. Recently however, a flaw in the proof has been observed [WŁ17a], with a correct proof announced but as yet not published [WŁ17b]. As Corollary 6.1.2.1 and the preceding theorem show, the addition of only a minimal amount of induction, or alternatively a weak reflection principle, is sufficient to produce a theory which is not arithmetically conservative, or as vividly described by Enayat [WŁ17b], which crosses the Tarski boundary.

Finally, we have so far considered reflection principles where the objects are sentences of the base-theory. If the sentences involved are in the extended language, containing the truth predicate, **CT** (with full induction for the extended language) can be recovered [Cie10]. The relevant reflection principle, now over the extended language, is:

$$RFN^T(\emptyset) : \forall \bar{x}[Pr_\emptyset(\varphi(\ulcorner \dot{\bar{x}} \urcorner)) \to \varphi(\bar{x})], \qquad \varphi \text{ is a formula in } L_{PAT}.$$

Adding this schema to $\mathbf{CT}^-$ recovers **CT**.

Theorem 6.1.4. [Cie10]

$$\mathbf{CT}^- + RFN^T(\emptyset) = \mathbf{CT}.$$

Proof. The aim is to show that all instances of induction can be proved within $\mathbf{CT}^- + RFN^T(\emptyset)$. Assume then that $\varphi(0) \land \forall x(\varphi(x) \to \varphi(S(x)))$, where $\varphi \in L_{PAT}$. It holds already by first order logic that, for every numeral n:

$$[\varphi(0) \land \forall x(\varphi(x) \to \varphi(S(x))] \to \varphi(n).$$

It follows by the construction of the provability predicate, and the definition of Feferman's dot notation that:

$$\mathbf{CT}^- \vdash \forall y Pr_\emptyset(\ulcorner [\varphi(0) \land \forall x(\varphi(x) \to \varphi(S(x)))] \to \varphi(\dot{y}) \urcorner).$$

By $RFN_\emptyset$ we obtain:

$$\mathbf{CT}^- \vdash \forall y[[\varphi(0) \land \forall x(\varphi(x) \to \varphi(S(x)))] \to \varphi(y)].$$

After shifting quantifiers:

$$\mathbf{CT}^- \vdash [\varphi(0) \land \forall x(\varphi(x) \to \varphi(S(x)))] \to \forall x \varphi(x).$$ □

6.2 Disquotational Truth

Aside from $\mathbf{CT}^-$, a natural starting point for a conservative truth theory is $\mathbf{TB}^-$. The question arises again: How strong is the theory resulting from the addition of suitable reflection principles? All the following results concern reflection in the extended language L_{PAT} rather than L_{PA}. For technical reasons that will become clear later, we consider $\mathbf{I\Sigma_1}$ as our background theory, rather than **PA** as has been the case so far. Since $\mathbf{I\Sigma_1}$ is expressive enough to code all syntactical notions used, we can still formulate the truth theories as before. The theories with $\mathbf{I\Sigma_1}$ as background theory will be identified with a subscript $\mathbf{Th}_{\mathbf{I\Sigma_1}}$, while the theories with **PA** as background theory will be left subscript-less. A first result is that uniformly reflecting on $\mathbf{TB}^-_{\mathbf{I\Sigma_1}}$ recovers the truth theory **UTB** , standing for 'uniform Tarski biconditionals'.

Definition 6.2.1. The theory **UTB** is given by the axioms of PAT, as well as the uniform Tarski biconditional schema:

$$Ax(UTB) = Ax(PAT) \cup \{\forall \bar{x}[T(\ulcorner \varphi(\dot{\bar{x}}) \urcorner) \leftrightarrow \varphi(\bar{x})] \mid \varphi(\bar{x}) \in Form_{PAT}\}$$

Theorem 6.2.1.

$$\mathbf{TB^{-}_{I\Sigma_1}} + RFN^{T}(TB^{-}_{I\Sigma_1}) \vdash \mathbf{UTB}.$$

Proof. The first part of the proof is showing that the uniform Tarski biconditionals are recovered by $RFN^{T}(TB^{-}_{I\Sigma_1})$. Fix $\varphi(x_1, \dots, x_n) \in Form_{PAT}$. In $\mathbf{TB^{-}_{I\Sigma_1}}$, for each tuple of numerals $\langle n, \dots, m\rangle$, we have that:

$$\mathbf{TB^{-}_{I\Sigma_1}} \vdash \varphi(n, \dots, m) \leftrightarrow T(\ulcorner\varphi(n, \dots, m)\urcorner).$$

Under a natural formalisation of provability in $\mathbf{TB^{-}_{I\Sigma_1}}$ we then have that:

$$\mathbf{TB^{-}_{I\Sigma_1}} \vdash \forall x_1, \dots, x_n Pr_{TB^{-}_{I\Sigma_1}}(\ulcorner[\varphi(\dot{x}_1, \dots, \dot{x}_n) \leftrightarrow T(\varphi(\dot{x}_1, \dots, \dot{x}_n))]\urcorner).$$

Applying $RFN^{T}(TB^{-}_{I\Sigma_1})$ yields the result. For the second part, we need to prove that induction for the extended language results. Again, fix an arbitrary $\varphi(x) \in Form_{PAT}$. By an easy case of (external) induction we have that for every numeral n:

$$\mathbf{TB^{-}_{I\Sigma_1}} \vdash \varphi(0) \wedge \forall x[\varphi(x) \to \varphi(S(x))] \to \varphi(n).$$

The formalised counterpart is then:

$$\mathbf{TB^{-}_{I\Sigma_1}} \vdash \forall y Pr_{TB^{-}_{I\Sigma_1}}(\ulcorner[\varphi(0) \wedge \forall x(\varphi(x) \to \varphi(S(x)))] \to \varphi(\dot{y})\urcorner).$$

Applying $RFN^{T}(TB^{-}_{I\Sigma_1})$ once again, finishes the proof. □

Notice that the theory **UTB** contains induction both for formulas of arbitrary complexity (in contrast to the arithmetical base theory $\mathbf{I\Sigma_1}$), as well as formulas containing the truth predicate (in contrast to $\mathbf{TB^{-}_{I\Sigma_1}}$, which does not extend induction to the truth predicate). Once one has applied reflection to a base theory, a natural continuation is to consider reflecting further on the new theory so produced. This line of work goes back to Turing and his ordinal logics [Tur36], as well as Feferman [Fef62] where transfinite progressions of the theories reflected on are considered. For notational convenience, we will denote the theory $\mathbf{TB^{-}_{I\Sigma_1}}$ as $\mathbf{TB^{0}_{I\Sigma_1}}$, the reflected theory $\mathbf{TB^{-}_{I\Sigma_1}} + RFN^{T}(TB^{-}_{I\Sigma_1})$ as $\mathbf{TB^{1}_{I\Sigma_1}}$ and so on. The following result due to Halbach [Hal01a] will help us climb another step in the reflective progression from $\mathbf{TB^{-}_{I\Sigma_1}}$.

Theorem 6.2.2. The theories **CT** and $\mathbf{UTB^{1}_{I\Sigma_1}}$ are equivalent.

We prove the corresponding directions in turn.

Lemma 6.2.3.

$$\mathbf{UTB^{1}_{I\Sigma_1}} \vdash \mathbf{CT}$$

Proof. The proof is a matter of checking that each of the compositional truth axioms is derivable in $\mathbf{UTB^0_{I\Sigma_1}} + RFN^T(UTB^0_{I\Sigma_1})$. We limit ourselves to the cases of conjunction(**CT-4**) and the universal quantifier (**CT-5**), with the other cases being similar. For the first case, we need to prove that

$$\mathbf{UTB^0_{I\Sigma_1}} + RFN^T(UTB^0_{I\Sigma_1}) \vdash \forall \varphi, \psi \in Sent_{PA} : T(\ulcorner \varphi \wedge \psi \urcorner) \leftrightarrow T(\ulcorner \varphi \urcorner) \wedge T(\ulcorner \psi \urcorner).$$

Fixing sentences φ and ψ, by the respective Tarski-biconditionals and propositional reasoning it is easy to show that

$$\mathbf{UTB^0_{I\Sigma_1}} \vdash T(\ulcorner \varphi \wedge \psi \urcorner) \leftrightarrow T(\ulcorner \varphi \urcorner) \wedge T(\ulcorner \psi \urcorner),$$

from which follows:

$$\mathbf{UTB^0_{I\Sigma_1}} \vdash \forall \varphi, \psi \in Sent_{PA} : Pr_{UTB_{I\Sigma_1}}(\ulcorner T(\varphi \wedge \psi) \leftrightarrow T(\varphi) \wedge T(\psi) \urcorner).$$

By $RFN^T(UTB^0_{I\Sigma_1})$ the first case follows.

The second case is only slightly more difficult to prove. Fix a formula $\varphi(x)$ and variable v such that $\forall v \varphi(v)$ is a sentence. We reason in $\mathbf{UTB^0_{I\Sigma_1}}$:

$$\begin{array}{lll}
(1) & \mathbf{UTB^0_{I\Sigma_1}} \vdash \forall v[T(\ulcorner \varphi(\dot{v}) \urcorner) \leftrightarrow \varphi(v)] & \mathbf{UTB^0_{I\Sigma_1}} \text{ axiom} \\
(2) & \mathbf{UTB^0_{I\Sigma_1}} \vdash \forall v T(\ulcorner \varphi(\dot{v}) \urcorner) \leftrightarrow \forall v \varphi(v) & \text{Logic and 1} \\
(3) & \mathbf{UTB^0_{I\Sigma_1}} \vdash T(\ulcorner \forall v \varphi(v) \urcorner) \leftrightarrow \forall v \varphi(v) & \mathbf{UTB^0_{I\Sigma_1}} \text{ axiom} \\
(4) & \mathbf{UTB^0_{I\Sigma_1}} \vdash T(\ulcorner \forall v \varphi(v) \urcorner) \leftrightarrow \forall v T(\ulcorner \varphi(\dot{v}) \urcorner) & \text{Logic and 2,3} \\
(5) & \mathbf{UTB^0_{I\Sigma_1}} \vdash T(\ulcorner \forall v \varphi(v) \urcorner) \leftrightarrow \forall x T(\ulcorner \varphi(\dot{x}) \urcorner) & \text{Variable renaming}
\end{array}$$

Since v and $\varphi(x)$ were arbitrary, we derive:

$$\mathbf{UTB^0_{I\Sigma_1}} \vdash \forall v, \varphi(x) Pr_{UTB^0_{I\Sigma_1}}(\ulcorner T(\forall v \varphi(v)) \leftrightarrow \forall x T(\varphi(x)) \urcorner),$$

from which by $RFN^T(UTB^0_{I\Sigma_1})$ the result follows. □

Lemma 6.2.4.

$$\mathbf{CT} \vdash \mathbf{UTB^0_{I\Sigma_1}}$$

Proof sketch. The reason for preferring $\mathbf{I\Sigma_1}$ as a base theory in this section can now be given. Similar to the proof in theorem 2.3.8, we have to show that $RFN^T(UTB^0_{I\Sigma_1})$ can be derived in **CT** by arguing for a partial global reflection principle. First notice that $\mathbf{UTB^0_{I\Sigma_1}}$ is a subtheory of **CT**, and moreover that the finitely axiomatizable $\mathbf{I\Sigma_1}$, as well as the finitely many compositional truth axioms are sufficient to derive it. Making the necessary changes in the provability and partial truth predicates in **CT** to include the compositional axioms yields $RFN^T(UTB^0_{I\Sigma_1})$, as in the proof of theorem 2.3.8. □

If we now take the second step in the reflective progression over $\mathbf{TB^0_{I\Sigma_1}}$, we can conclude the following:

Corollary 6.2.4.1.

$$\mathbf{TB^{2}_{I\Sigma_1}} \vdash \mathbf{CT}$$

Proof. Since $\mathbf{UTB^{0}_{I\Sigma_1}}$ is a sub-theory of **UTB**, the result follows by combining theorem 6.2.1 and theorem 6.2.2. □

CHAPTER 7

Justifying Reflection

7.1 The Implicit Commitment Thesis

So far, we have remained quiet on the justification for adjoining different reflection principles to a base theory. First of all, we should, as Feferman pointed out [Fef91], distinguish between the informal notion of reflecting on the concepts one uses and the theories one subscribes to, and the formalized counterpart of reflection principles. For example, it might be that by reflecting on the axioms of set-theory, one realizes that large cardinal axioms are to be accepted, even though no amount of formalized reflection will derive them. Hence, the formalized reflection principles we have studied so far do not capture everything which occurs in reflecting on the theories one is committed to, but they do offer a lower bound. According to Feferman, the progression of theories obtained through formalized reflection are not merely acceptable when one subscribes to the base theory, but we are even *obliged* to accept their consequences :

> The notions of reflective closure introduced here are relative to a theory in the sense that they merely tell us what *ought* [my emphasis] to be accepted if one has accepted the given basic notions and schematic principles of that theory. [Fef91]

We might well ask what this normative force derives from. There is little explicit defense in the literature of the obligation an idealized mathematician is under to accept certain additional statements once they have accepted a given base theory. Indeed, Horsten has argued that these principles are merely rationally acceptable once one reflects on one's commitments [Hor]. The kind of reasoning that seems to underly the obligation in question is the following. As Horsten points out, taking a cue from Van Fraassen [Bas80, p.12-13], acceptance of a theory has both a pragmatic and doxastic component. Pragmatically accepting a theory **S** means that one is committed to using it to further one's research goals, and more prosaic goals: both cashier and engineer alike are pragmatically relying on an arithmetical theory when they determine the correct change, or determine the load-bearing properties of a bridge. Doxastically accepting an (axiomatized) theory **S** means that one *believes* the axioms, and rules of inference to be sound and complete. Presumably, this belief is grounded in justification invoking an intended model, or class of models, as when **PA** is believed to be true since its axioms are true of the natural numbers. This gives us an inkling of why one is obliged to accept statements that go beyond the base theory one has accepted. The easy case is to consider one's explicit commitments first. Say I accept the theory **PA**, in the manner described previously. If someone were to show me a proof (in **PA**) of the theorem that there are infinitely many primes, and I would reject the theorem, I'd be both pragmatically and doxastically deficient. Pragmatically, since I renege on my willingness to use the theorems of **PA**, even as new ones are introduced to me, and doxastically, since I see the correctness of every step involved in the proof yet deny the truth of the theorem. So one is indeed obliged to accept all theorems of **S** when one accepts the theory **S**. The question is then how this obligation could extend to statements which, like the reflection principles considered in this thesis, are underivable in **S**. Dean gives the following analysis of this obligation: similar to how our *explicit commitment* to **S** obliges us to accept the theorems of **S**, we are *implicitly committed* to a theory **S**+, in which the reflection principles adjoined can be derived [Dea14, p. 33,57]. This theory **S**+ has the resources to formalize the informal truth-theoretic argument of Section 4.4. Since we are implicitly committed to **S**+, we are under an obligation to accept its theorems, in particular the reflection principles for **S**. We will take a look at this argument in more detail later, but for now follow the literature by considering principles arrived through reflection to have normative force.

This obligation has come to be known as the *implicit commitment thesis* (ICT). Different variants are present in the literature, reflecting the different understandings of what it is exactly that one is committed to when one accepts a theory **S**. The most open-ended version is given by:

(open ICT) In accepting a formal theory **S** one is also committed to additional resources not available in the starting theory **S** but whose acceptance is implicit in the acceptance of **S**. [NP19]

Similarly we find:

(closed ICT) Anyone who accepts the axioms of a mathematical theory **S** is thereby also committed to accepting various additional statements Δ which are expressible in the language of **S** but which are formally independent of its axioms. [Dea14]

While superficially similar, there are two important points on which the principles differ. First, open ICT does not a priori restrict the additional resources to be in the form of statements which are to be formulated in the language of **S**, they could even be new derivation rules rather than additional axioms. In accepting a theory $\mathbf{S}$ in L_S one might be committed to a new theory $\mathbf{S}^+$ which is formulated in a different language L_{S^+}. We are *open* to expand the language in which our theory is formulated. On the other hand, it is not the case that the theories committed to through open ICT are necessarily stronger than the theories obtained through closed ICT. The closed ICT principle obliges us to accept additional statements which are independent of $\mathbf{S}$, so that one is committed to a theory $\mathbf{S}^+$ which is not conservative over $\mathbf{S}$. According to the open ICT principle however, the additional resources we are committed might well lead us to accept a theory $\mathbf{S}^+$ which is conservative over $\mathbf{S}$.

Of course, the ICT principle as such leaves many questions open, depending on the base theory $\mathbf{S}$. Exactly which resources or additional statements we are committed to when accepting a base theory $\mathbf{S}$ could depend on $\mathbf{S}$, and the epistemic status of the agent accepting the theory $\mathbf{S}$. In practice, we always find that the resources in question are expressions of the *soundness* of the underlying system $\mathbf{S}$, as when Feferman states:

> By a reflection principle we understand a description of a procedure for adding to any set of axioms A certain new axioms whose validity follow from the validity of the axioms A and which formally express, within the language of A, evident consequences of the assumption that all the theorems of A are valid. [Fef62]

We know that by the incompleteness theorems, a principle of this kind has to amount to adding new axioms, since if the system $\mathbf{S}$ were able to prove its own soundness, consistency can be proved within $\mathbf{S}$.

7.2 The Closed ICT Thesis

We start by recapitulating the arguments found in [Dea14], where the closed ICT principle is scrutinized. Essentially, Dean denies that the closed ICT thesis holds universally, due to some theories being *epistemically stable* , meaning that they are:

> [...] stable in the sense that there exists a coherent rationale for accepting this system which does not entail or otherwise oblige a theorist to accept statements which cannot be derived from its axioms. [Dea14, p. 53]

To be clear, it is not the case that an epistemically stable theory cannot be extended to include independent statements, only that there exist *epistemic positions* from which the theory is closed in terms of one's justification. One of the epistemically stable theories Dean considers is **PRA**, from the perspective of finitism. Tait has argued for the identification of the theorems of primitive recursive arithmetic (**PRA**) with the finitist theorems, in contrast with Kreisel's claim that the finitist theorems coincide with those of **PA** [Tai81]. Primitive recursive arithmetic admits different axiomatizations, but can be seen as arithmetic where only primitive recursive predicates are expressible, and induction is restricted to these predicates only. By the identification of finitist mathematics with **PRA**, accepting a reflection principle such as $RFN(PRA)$ is impossible as finitist, since the resulting theory would exceed **PRA**. This can be seen by the fact that $\mathbf{EA} \subset \mathbf{PRA}$, and we know from Theorem 2.3.5 that accepting the uniform reflection principle over **EA** obtains **PA**. Indeed, Tait has argued against Kreisel's analysis of finitism, precisely *because* he assumes the validity of a reflection principle of the form (see Section 3.3):

$$\text{From accepting } Pr_{PRA}(\ulcorner \varphi(0^x) \urcorner), \text{ accept } \varphi(x).$$

Kreisel's reasoning for the validity is familiar: If one finitistically comes to accept **PRA**, and finitistically comes to accept $Pr_{PRA}(\ulcorner \varphi(0^x) \urcorner)$, one should also accept $\varphi(x)$. But while a finitist might be able to prove every instance of a function defined by primitive recursion, the validity of **PRA** cannot be understood by them, since this requires seeing that definition by primitive recursion is valid *in general*, which requires accepting the notion of a function [Tai81, p. 545]. And for a finitist, the general notion of a function is not available, as it is in general a transfinite object. So for Tait, the reflection principle invoked by Kreisel already rests on a non-finitist base.

A second example given by Dean is the position of Isaacson, which he calls *first-orderism* [Isa87]. Isaacson argues that **PA** delineates the concept of first-order arithmetic. Attempting to extend **PA** using only 'arithmetical concepts' would not lead to a stronger theory. The justification for holding that **PA** has this status is however inherently higher-order. Namely, we see **PA** to be the first-order projection of second-order arithmetic, which singles out the natural numbers, by converting the second-order induction axiom to a first-order one. Isaacson holds that we are (sometimes) able to perceive the truth of certain statements that are independent of **PA**, such as $Con(PA)$ and $RFN(PA)$, but that this understanding rests on our antecedent higher-order understanding of the natural numbers. We are, in other words, only ready to accept certain reflection principles to the degree that we are already committed to something which goes beyond **PA**, and not by virtue of our commitment to **PA** itself.

These examples might not be enough to convince a proponent of ICT that does not hold universally. It is unclear whether Tait's finitist truly *accepts* **PRA** if they can't accept the validity of primitive recursion in general. Accepting each theorem of **PRA** individually seems intensionally different from accepting **PRA**, even if the theories so obtained are extensionally equivalent. The first-orderist position of Isaacson also seem peculiar in that it grounds our acceptance of a **PA** on a higher-order understanding, which does include the resources to determine certain statements true that are independent of **PA**. So at least in this case, the justification for accepting **PA** already includes the justification for committing to additional resources that go beyond **PA**.

These are far from knock-out counter-arguments, and in any case, it seems plausible that there are enough counter-examples that defending ICT *as is* would start to feel like playing whac-a-mole. As Dean points out, it does not seem like much of a concession to claim instead that ICT is a commitment that might fail to obtain for certain minority positions (certainly most mathematicians do not worry about helping themselves to theories that go beyond **PRA** and **PA**), but still reasonably describes the epistemic position most of us are in. He then presents a few more arguments for why even this slight reformulation is problematic. Of these, the most relevant one for us is the analysis of the usual justification for ICT. Dean isolates the truth-theoretic reasoning of Section 4.4 as playing the justificatory role for committing to reflection principles. That is, it is the embedding of a first-order theory (e.g. **PA**) in a truth-theory (e.g. **CT**), which makes it possible to formally derive the reflection principles. We know from the proof of Theorem 5.2.1 that this argument uses induction extended to formulas containing the truth predicate in an essential way. As Dean points out, if this truth-predicate is supposed to express the concept of arithmetical truth, then this is a highly non-arithmetical concept (by Tarski's undefinability theorem). Often, the induction schema is treated as open-ended: any predicate or formula can occur in it which is a property of the natural numbers. But, says Dean, it is difficult to see what the justification could be for extending the induction schema to formulas containing the truth predicate based on our *acceptance of* **PA** *alone*. After all, the truth predicate does not an express an arithmetical concept. One possible argument could be that we understand the first-order induction schema through the second-order induction schema, as Kreisel suggests:

> A moment's reflection shows that the evidence of the first order axiom schema derives from the second order schema: the difference is that when one puts down the first order schema one is supposed to have convinced oneself that the specific formulae used (in particular, the logical operations) are well defined in any structure that one considers. [Kre67]

Unfortunately, this argument for extending induction, similar as it is to Isaacson's justification to consider **PA** an epistemically stable theory of arithmetic, will not serve us to ground ICT. If we argue for ICT by referring to the derivability of the reflection principles in a truth-theory, this truth-theory itself should be justified on the basis of the justification for accepting the base theory. Since we are in particular interested in first-order theories, we can't rely on an implicit acceptance of a second-order schema to do so. We will come back to the issue of extending induction to the truth-predicate. It is my view that induction is indeed open-ended, and that this move is unproblematic. But I agree with Dean that explicating our acceptance of reflection principles by way of a formal truth-theory is suspect. It is not obvious of how we come to have a grasp of a formal concept of truth for a given base theory. We will discuss this further in section 7.4, but first take a look at a different proposal, due to Nicolai and Piazza [NP19], which makes the reliance on a formal truth-theory even more explicit.

7.3 The Open ICT Thesis

Nicolai and Piazza defend the open ICT thesis. As noted before, this thesis does not entail that on reflection over the base theory **S**, one comes to accept statements that are independent of **S**. They take it that Dean's examples (to which they add one of their own) of epistemically stable theories does preclude the acceptance of the closed ICT thesis. From the perspective of the foundational programs discussed, reflection cannot lead one to accept theorems that are not derivable within the theory. Nevertheless, Nicolai and Piazza claim that there is a *fixed semantic core* of the implicit commitment in accepting a theory **S**. This semantic core is necessarily conservative if it is to be implicit even in the case of Isaacson's first-orderism and Tait's finitism. Aside from the semantic core, there is a variable component which depends on the justification available in accepting a theory, and is empty in the case of the foundational theories of Isaacson and Tait.

What does this semantic core consist of? The authors hold that our implicit commitment to a theory is a soundness extension of the theory. They observe that when formulating a soundness extension, the notion of truth is hard to do without. As pointed out by Kreisel [KL68, p.98], the principles RFN and Rfn can be seen as stand-ins for the global reflection principle; the *intended* principle is really $\forall\varphi : Pr(\ulcorner\varphi\urcorner) \rightarrow T(\ulcorner\varphi\urcorner)$. That this strategy works at all is due to the implicit acceptance of the T-schema: $\varphi \leftrightarrow T(\ulcorner\varphi\urcorner)$. But, as we have seen, adjoining either the local or uniform reflection principle to a given base theory usually leads to a non-conservative theory. So, the authors formulate an extension of the non-finitely axiomatizable theory **S** in the language $L_S \cup \{T\}$, which expresses the soundness of **S** to the extent that this can be done in a conservative manner. First, they consider the theory $\mathbf{S}^+$, given by the axioms of **S**, as well as (a term-version of) the compositional axioms **CT**-2 to **CT**-5 of Definition 4.3.1, with induction not extended to the truth predicate. It turns out that $\mathbf{S}^+$ is not only conservative, but can't even show that all non-logical axioms of **S** are true, that is:

$$\forall\varphi[Ax_S(\varphi) \rightarrow T(\ulcorner\varphi\urcorner)]$$

is underivable in $\mathbf{S}^+$. They then consider the theory $\mathbf{CT}[\mathbf{S}]^-$ [1], which is given by:

$$\mathbf{S}^+ \cup \{\forall\varphi[Ax_S(\varphi) \rightarrow T(\ulcorner\varphi\urcorner)]\}.$$

In effect, $\mathbf{CT}[\mathbf{S}]^-$ is a theory which expresses the compositionality of truth, and the fact that we believe the non-logical axioms of **S** to be true. This theory is also conservative, at least over theories stronger than **EA**:

Theorem 7.3.1. For a recursively axiomatizable theory $\mathbf{S} \supseteq \mathbf{EA}$, $\mathbf{CT}[\mathbf{S}]^-$ is syntactically conservative over **S**. [Lei15]

Under this dynamic reading of ICT, the semantic core represents a lower bound on what we are committed to by accepting the base theory. The authors identify three possible points of objection to their proposal of the semantic core as $\mathbf{CT}[\mathbf{S}]^-$, and reply to them in turn. The first point is that the induction schema is not extended to formulas including the truth predicate. Clearly, $\mathbf{CT}[\mathbf{S}]^-$ fulfils its function by skirting

[1] The authors refer to the theory as $\mathbf{CT}[\mathbf{S}]$ in the paper. To be consistent with the notation in the thesis the '−' superscript has been added, in accordance with the lack of extended induction of the theory.

closely to non-conservativeness, without actually becoming so. We know, by virtue of Theorem 6.1.3, that even adding Δ_0-induction for truth is sufficient to turn $\mathbf{CT}[\mathbf{S}]^-$ into a non-conservative theory. The objection is responded to by invoking the distinction between the linguistic and the meta-linguistic level. On the linguistic level, we have justification for the *mathematical* principle of induction, and a proof by induction of $\forall x : 2^x > x$ states a fact about exponentiation. On the meta-linguistic level, a proof of $T(\forall x : 2^x > x)$ is a proof that every substitution of a numeral in the formula $\forall x : 2^x > x$ is true. On the linguistic level we consider properties of a function, on the meta-linguistic level properties of formulas or sentences. Hence, the justification for the mathematical principle of induction is not necessarily justification for extending the induction schema to truth, since we need justification on the meta-linguistic level. There are, I think, two issues with the argument. The first issue is external to the theory **S** being considered. As has been observed by McGee, the principle of mathematical induction seems to be unproblematic in a host of fields:

> Mathematics, including the principle of induction, is the common background to all the sciences. If the chemist wants to establish some property of plastics by induction on the length of polymer chains, she allows the introduction of chemical vocabulary in the inductions axioms, without any worry that perhaps induction is only legitimate within pure mathematics, so that it's no longer applicable when chemical concepts are involved. [...] The chemist is not reaching beyond the bounds of classical mathematics; her endeavor [is] to expand the bounds of known chemistry. [McG06, p. 111]

We might concede that there is something to the separation between the the linguistic and meta-linguistic level, but then there still needs to be an argument for the exceptionalism of the meta-linguistic level. If the chemist does not require additional justification for extending the induction principle to include chemical vocabulary, then why should the truth-theorist?

The second issue is an internal one. If indeed there were no justification to extend induction to predicates formalizing meta-linguistic concepts, the induction principle would also fail to apply to formulas containing many other coded concepts such as Pr_S, $Form_{L_S}$, $Term_{L_S}$ etc. Many proofs that have been accepted with no reservations, for example the proof of the existence of partial truth predicates (see Corollary 4.1.3.1), rely in an essential way on induction to show that the meta-linguistic concepts used are well-behaved. So, unless we are willing to concede that these results are also in need of extra justification above and beyond the justification for the theory **S** in general, and the induction principle in particular, a different argument is needed to explain why it is the concept of truth specifically which cannot be introduced in the induction schema.

The second objection the authors anticipated is the lack of plausible soundness principles such as $Con(S)$ in the semantic core, since $Con(S)$ is independent of **S**. Their reply is that the semantic core does not necessarily have to be a natural soundness extension of

S, rather it is supposed to capture the minimum resources anyone accepting **S** should also be obliged to accept. In particular, although it is natural to endorse $Con(S)$ if one accepts **S**, this endorsement will rest on additional, unavailable justification when **S** is an epistemically stable theory.

Finally, the third objection considered relates to a subtle point that is easy to miss. The semantic core includes the sentence $\forall\varphi[Ax_S(\varphi) \to T(\ulcorner\varphi\urcorner)]$, where Ax_S represents the *non-logical axioms* of **S** only. The reason for this, once again, is the aim of developing a soundness extension which is conservative. We know from Theorem 6.1.2 that global reflection for first-order logic over $\mathbf{CT}^-$ already suffices to prove the non-conservativity of the resulting theory. The theory $\mathbf{CT}[\mathbf{S}]^-$ + "all logical axioms are true" is *almost* able to prove $GR(\emptyset)$. The truth of all logical axioms, and the truth of (internalized) modus ponens (which follows from **CT**-3 and **CT**-4) is sufficient to show that $Pr^1_\emptyset(\ulcorner\varphi\urcorner) \to T(\ulcorner\varphi\urcorner)$ holds, where $Pr^1_\emptyset(\ulcorner\varphi\urcorner)$ represents provability with one application of modus ponens. This cannot be generalized straightforwardly to $GR(\emptyset)$ since we do not have the requisite induction to show the truth of iterated applications of modus ponens. The authors conjecture that $\mathbf{CT}[\mathbf{S}]^-$ + "all logical axioms are true" is in fact conservative, but lacking proof, do not extend the semantic core to include the truth of the logical axioms. This does once again raise the question of whether the semantic core is natural. Lacking $GR_\emptyset$, the semantic core does not even contain the closure of truth under first-order reasoning. How can one see one's theory **S** as true while simultaneously not holding that first-order logic is true? The authors' answer is similar to the answer to the second objection: Accepting first-order reasoning as true is a natural and desirable part of a soundness extension, but in some cases requires additional justification beyond the justification necessary to accept an epistemically stable theory.

Now, there is no question that $\mathbf{CT}[\mathbf{S}]^-$ does indeed offer a minimal set of axioms that is conservative over **S**, if one accepts that soundness extensions should be formulated by explicit usage of a truth predicate. In this way, it makes good on the claim that even for epistemically stable theories there could be obligations to accept additional statements. The question is whether the minimal justification for accepting **S** commits oneself only to $\mathbf{CT}[\mathbf{S}]^-$, that is, if it really is a lower bound. We have already argued that refraining from extending the induction schema to the truth-predicate by invoking a separation between the linguistic and meta-linguistic level is misguided. We know that extending the induction schema of $\mathbf{CT}[\mathbf{S}]^-$ would lead to $\mathbf{CT}[\mathbf{S}]$ which is not conservative over **S**, and so the open ICT thesis would have to be rejected. Nevertheless, I think a more serious worry is the lack of closure of truth under first-order reasoning. The authors take it that for epistemically stable theories additional justification is necessary beyond the justification needed to accept the theory. But surely, for any theory **S** formulated in first-order logic, accepting **S**, and formalizing this by including $\forall\varphi[Ax_S(\varphi) \to T(\ulcorner\varphi\urcorner)]$ in the semantic core, already *presupposes* that one finds first-order logic acceptable, at least with respect to the fixed language L_S. In fact, our commitment to first-order logic runs deeper than our commitment to any first-order theory **S**. If we were to find a contradiction in the theory **S**, we'd revise the non-logical principles of **S**, rather than doubt first-order logic. In other words, the justification for accepting the theory **S** includes justification for accepting first-order logic. This is even the case for the epistemically stable theories that are taken to motivate the semantic core, **PRA** as Tait's explication of finitism and Isaacon's first-orderist **PA**. Hence, there is no coherent epistemic position from which a theory **S** can be justified, but the closure of truth under first-order reasoning is not part of the semantic core.

7.4 Commitment Through truth?

As we saw, Dean understands the justification for the closed ICT thesis to be based on an inductive truth-theoretic argument. This argument in its informal form can be formalized in different systems, all stronger than **S**. Dean identifies several constraints which a truth-theory must fulfil in order for the formalized inductive argument to go through, and takes **CT** by the way of example. Similarly, Nicolai and Piazza take the soundness extension to be a formal truth-theory, taking their cue from Kreisel's view that the local and uniform reflection principle are justified in terms of the global reflection principle. Dean finds that acceptance of a truth-theory in which the inductive argument can be formalized for **S** will in general not follow from one's justification for accepting **S**, and Nicolai and Piazza propose a theory $\mathbf{CT}[\mathbf{S}]^-$ which we've argued to be an incoherent theory to accept on the basis of the justification for accepting **S**. Contrary to understanding this to be a conclusion that (some form of) ICT ought to be rejected, I think that this shows that grounding ICT by appealing to truth is the wrong road to take.

It is my view that defending the ICT thesis by explicitly or implicitly relying on a concept of truth must founder, by virtue of the weight of the conceptual baggage taken on. What is needed is an analysis of how one comes to accept a local or uniform reflection principle, with no recourse to truth. The test of this analysis will be whether it has explanatory power, by showing us *why* reflection principles are unacceptable in the cases of epistemically stable theories. An important aspect of the analysis is that Kreisel's explanatory direction from the global reflection principle to RFN and Rfn is reversed. We do not accept Rfn based on our acceptance of the global reflection principle, but accept the global reflection principle by our acceptance of Rfn, as well as a commitment to a minimal truth theory (i.e. it deriving the T-schema). This latter commitment is not a given, that is, one can be deeply suspicious of the concept of truth, informal or formal, and still accept RFN and Rfn over a theory one has accepted previously.

To support this view, I will reevaluate the common defense of reflection principles in terms of truth, as exemplified by the inductive truth-theoretic argument. There are two strands to the argument. The negative strand will show how there is an equivocation in the informal argument, which makes the argument much less transparent than it seems on a first reading. The positive strand will argue for a truth-less defense of Rfn (and RFN) by giving an account of how an idealized mathematician might come to accept Rfn (and RFN) over an accepted base theory. The case of purely pragmatic acceptance of a theory, with no belief in the truth of the theory, will be especially important. Going through the arguments, it is helpful to keep in mind the different meanings of the concept of truth which we need to disambiguate, as summarized in List 1 in the introduction to this thesis.

Starting with the negative strand, we analyze the the informal argument of section 4.4, and its translation in a formal truth-theory as Dean suggests. The informal argument relies on accepting that the axioms and inference rules of **PA** are true. Considering the different meanings of truth mentioned in List 1 , it should be clear that arithmetical truth is intended. Arithmetical truth is hyperarithmetical, and hence there is no effective way in which one can understand it, even for the idealized mathematician. By translating the informal argument to a derivation in a formal truth-theory, we equivocate between arithmetical truth and truth in a formal theory. So while it seems like we have a grasp on how we come to accept the global, and a fortiori, the local reflection principle, the argument really relies on us having a grasp on arithmetical truth to begin with, which is surely not more evident. I see two ways in which this move can be legitimate. The first way is by accepting that one intended arithmetical truth, but that the translation into a formal truth theory is legitimate. By Tarski's undefinability theorem, there is no completely faithful translation, so it has to be done in such a way that the salient aspects of arithmetical truth are preserved. This is far from an obvious task; after all we know that **TB** – which seemed eminently plausible – is unable to derive even the local reflection principle since it is conservative over **PA**. A more natural choice would be **CT**, which is indeed the example Dean gives, and which is able to derive the global reflection principle. But both **CT** and **TB** were developed as deflationary truth theories, and were explicitly not intended to formalize arithmetical truth. Even if one were willing to commit to a formal truth-theory, there would still be the question of which formal truth-theory should be chosen. While **CT** is a natural choice, there are many formal truth-theories on offer, and if the lack of unity in the field on the correct formalization of deflationary truth is anything to go by, the choice is far from obvious. The second way is to argue that one didn't intend arithmetical truth in the first place, but the formal T-predicate instead. But this only gets us to the root problem quicker: which formalization did one intend? Justifying the choice for a particular formal truth-theory is no easy task, so it should be no wonder that one's justification for a given base theory does not suffice to justify the reflection principles the ICT thesis will have us be obliged to.

I think that part of the appeal of the informal argument is that it is formalized by **CT** in a nice way when taking **PA** as the base theory, the two theories we have studied the most in this thesis. The formalisation is much less convincing when we consider different theories. In Chapter 6 we looked at the effect of adding reflection principles over theories like $\mathbf{TB}^-$ and $\mathbf{CT}^-$. The informal argument for accepting the global reflection principle for $\mathbf{CT}^-$ would then be :

1. All axioms of $\mathbf{CT}^-$ are true.
2. All rules of inference of $\mathbf{CT}^-$ are truth-preserving.

3. Hence, all theorems of $\mathbf{CT}^-$ are true.

If we were to formalize our acceptance of the global reflection principle for $\mathbf{CT}^-$, in the same way Dean did, this would amount to formulating a theory 'like' $\mathbf{CT}$ over $\mathbf{CT}^-$. The most natural way to do this would be to define the ramified truth theory $\mathbf{CT}_\mathbf{1}^-$, formulated in the language L_{T,T_1}.

Definition 7.4.1. The theory $\mathbf{CT}_\mathbf{1}^-$ is the truth-theory axiomatized by:

- The logical axioms formulated in L_{T,T_1};
- The axioms of $\mathbf{PA}$, including the induction axiom for formulas in L_{PA} only ;
- The $\mathbf{CT}$ axioms of Definition 4.3.1.
- $\forall s,t \in Tm^c : T_1(s=t) \leftrightarrow Val(s) = Val(t)$;
- $\forall \varphi \in Sent_{CT} : T_1(\ulcorner \neg\varphi \urcorner) \leftrightarrow \neg T_1(\ulcorner \varphi \urcorner)$;
- $\forall \varphi, \psi \in Sent_{CT} : T_1(\ulcorner \varphi \wedge \psi \urcorner) \leftrightarrow T_1(\ulcorner \varphi \urcorner) \wedge T_1(\ulcorner \psi \urcorner)$
- $\forall v, \varphi(x)\{Sent_{CT}(\ulcorner \forall v\varphi(v) \urcorner) \rightarrow [T_1(\ulcorner \forall v\varphi(v) \urcorner) \leftrightarrow \forall x T_1(\ulcorner \varphi(\dot{x}) \urcorner)]\}$

It is straightforward to see that this theory allows one to formalize claims of twice iterated truth. For example, "It is true that it is true that $0 = 0$" is translated as $T_1(\ulcorner T(0=0) \urcorner)$. In particular, the axioms of $\mathbf{CT}^-$ are all true in the sense of T_1. The theory $\mathbf{CT}_\mathbf{1}^-$ avoids contradiction by the way of the ramified truth-predicate. The difficulty the liar sentence presented is sidestepped, since the T_1-predicate does not apply to sentences containing T_1, but only sentences containing the 'lower' predicate T (cfr. Theorem 4.1.1).

There are two issues with this formalization. The first is that the derivation of the global reflection principle $Pr_{CT^-}(\ulcorner \varphi \urcorner) \rightarrow T_1(\ulcorner \varphi \urcorner)$ for $\mathbf{CT}^-$ is not possible, for the same reason that $\mathbf{CT}^-$ does not derive the global reflection principle for $\mathbf{PA}$, by lacking extended induction for the T_1 predicate. This could be addressed by formulating a theory $\mathbf{CT_1}$ with extended induction, but it is hard to see how one could justify extending induction to a truth predicate for the theory $\mathbf{CT_1}$, which one is implicitly committed to, but not for to the explicitly accepted theory $\mathbf{CT}^-$. Of course, this is not obviously an issue with the formalisation of the truth-theoretic argument. It might well be, as I believe, that $\mathbf{CT}^-$ simply is simply not a stable position to be in, following McGee's insistence on the open-endedness of the induction schema. The second issue is a bit pricklier. The ramified aspect of $\mathbf{CT_1}$ is simply implausible when it comes to formalizing the intuition behind the argument. The inductive argument appeals to our conception of truth, not our conception of first-order truth, second-order truth etc. And once one is implicitly committed to a theory $\mathbf{CT_1}$, why should one's implicit commitment not include the higher theory $\mathbf{CT_2}$ and so forth? As the saying goes, it is turtles all the way down. This objection and others against a ramified theory of truth are well known, and discussed in [Hor11, p.55-56]. Now, it is only fair to point out that Dean never identified **CT** as *the* formal truth-theory by which to understand the argument for reflection principles. A theory of type-free truth such as **KF** or **FS**, which we have not discussed in the thesis, would likely be a better option. But starting from a truth theory like **TB** or **CT**, justifying reflection principles by appealing to a different, type-free truth-theory seems to go beyond one's justification for the typed truth theory. Of course, Dean would likely see this example as grist to the mill for his position that ICT does not hold. In this thesis, it is taken as evidence that grounding Rfn or RFN by appealing to a concept of truth (either formal or informal) is fraught with difficulties.

The example discussed just now applies mutatis mutandi to the proposal of Nicolai and Piazza. The semantic core of the soundness extension, given by $\mathbf{CT}[\mathbf{S}]^-$, is not immediately applicable to the case where $\mathbf{S}$ is given by $\mathbf{CT}^-$. It is however straightforwardly adapted by identifying the semantic core of the soundness extension of **CT** with:

$$\mathbf{CT_1^-} \cup \{\forall\varphi[Ax_{CT^-}(\varphi) \to T_1(\ulcorner\varphi\urcorner)]\}.$$

Again, the ramified truth predicate so introduced simply does not seem a plausible formalisation of the intuition behind accepting the soundness of a theory. Nicolai and Piazza seem to implicitly recognize this limit of their proposal when they state:

> [...] nothing prevents one from asking herself what we are implicitly committed to when we are accepting the theory of truth. [...] However, since we are not interested in the theory of truth itself, but only in the boundary between acceptable and non-acceptable characterizations of the implicit commitment of the base theory, we do not consider further this possible extension of our analysis. [NP19]

I take it that the issues discussed are not not nails in the coffin for those who appeal to truth, either formal or informal, to justify the form the soundness extension for a given accepted base theory should take. What is missing however, in the papers of both Dean and Nicolai and Piazza, is a principled story of *how* justification for a base theory such as **PA** serves to justify accepting an expanded formal truth-theory and by this acceptance, certain reflection principles. We shall now attempt to give exactly such a story, which does not make use of the concept of truth, and show how it can explain not only our implicit commitments, but also the lack of implicit commitments in the case of the epistemically stable theories discussed. The closest thing to such an explanation can be found in an unpublished paper by Horsten [Hor], where he gives a compelling phenomenological analysis of how an idealized mathematician could come to know the consistency of the base theory, given their acceptance of the base theory. His analysis will be summarized in the next section.

7.5 A Phenomomenological Analysis of Reflection

Horsten couches his analysis in terms of Wright's study of *cognitive projects* [Wri04] *cognitive project*. A cognitive project is a pair of a question, and a procedure to execute in order to answer the question. Often, these cognitive projects will rely on some presuppositions being true, without which the cognitive project would not be possible. For example, the cognitive project of learning about the world around us presupposes that there is in fact an external world, and that we are not brains in a vat. A cognitive project thus entitles us to a presupposition P if:

1. There is not enough justification to believe P untrue;
2. Justifying P to be true would itself rely on further presuppositions, which are in turn no more secure than P itself.

Such an entitlement allows us to accept or trust the presupposition P, but not to *believe* P, since this would require justification we simply do not have.

Mathematics is filled to the brim with cognitive projects. The cognitive project Horsten is interested in is the project of discovering number-theoretic facts through proofs in **PA**. The idealized mathematician (who I will refer to as 'you' from now on, expressing our trust in the qualities of you, the reader!) taking part in this cognitive project accepts the theory **PA** unconditionally, and not merely instrumentally. You rely without reservations on the inference rules, and believe all theorems of **PA** through your acceptance of **PA**. In addition, you have justification for accepting **PA** but no justification for accepting the consistency of **PA**, which is essential for this cognitive project to get off the ground. However, the consistency of **PA** is a presupposition you are entitled to, since you have no justification to believe it to be untrue, and justifying the consistency of **PA** would rely on further presuppositions that are not any more secure (for example, presupposing **CT**). Horsten calls this situation the *state of innocence*. Now, we will describe the reflective process you might engage in.

The first moment of reflection is the realization that as you use the proof procedure (the second element of the cognitive project pair), you come to realize that, at least within the context of **PA**, you are essentially a machine that produces theorems of **PA**. In the state of innocence, you accepted the theorems of **PA** as they were derived, in the first moment of reflection you realize that the procedure you are using *is* deriving theorems of **PA**. Secondly, you realize (and before Gödel this was of course not an obvious realization) that this procedure can be expressed within **PA**, as a standard provability predicate Pr_{PA}. In the second moment of reflection, you manage to convince yourself that:

$$\text{For all } \varphi \in L_{PA} : \mathbf{PA} \vdash \varphi \text{ iff } \mathbf{PA} \vdash Pr_{PA}(\ulcorner\varphi\urcorner).$$

You do this by a proof of induction over proof length, on a meta-syntactical level. Horsten claims that if the minimum resources for this proof are spelled out, this can be done in a theory that is syntactically conservative over **PA**, with no reference to thick philosophical or semantical notions such as rational belief or truth. Horsten neglects to mention that this proof requires the additional presupposition that **PA** is Σ_1-sound, which is a stronger condition than presupposing mere consistency (see the discussion following Definition 2.2.4). Once this coding is in place, it is easy to see that consistency can be expressed by $\neg Pr_{PA}(\ulcorner \perp \urcorner)$, where $\perp$ can be any contradiction. In the final act of reflection, you come to realize that in your cognitive project you have been relying on the consistency of **PA**. If you were to derive a contradiction in **PA**, your cognitive project would collapse. There are two options available to you. One option is to remain agnostic on the matter of the consistency of **PA**. You are entitled to the consistency of **PA** in your cognitive project, but do not come to believe it. In this case, you instrumental acceptance of **PA** is now changed. Where before you relied on the theorems unconditionally, you now accept that there is an epistemic possibility that you are wrong to do so. This is also reflected in your doxastic acceptance. In the state of innocence you were unconditionally accepting of the axioms and inference rules of **PA**. Given that you now accept the epistemic possibility of deriving a contradiction in **PA**, it is rational to revise your unconditional belief in the axioms of **PA** to a somewhat more qualified belief, which can be spelled out by quantifying: you have a slightly lesser *degree* of belief in the axioms. The second option is to stick to your unconditional acceptance of **PA**. It would then be irrational to also hold it that there is an epistemic possibility of **PA**'s inconsistency. So you come to believe that **PA** is consistent, and through the coding you have developed, come to believe $\neg Pr_{PA}(\ulcorner \perp \urcorner)$. Note that this relies on you being in an epistemic position to *justifiably* accept **PA** unconditionally. The lesson of the analysis is that if your justification to accept **PA** is truly unconditional, then it is rational to believe $\neg Pr_{PA}(\ulcorner \perp \urcorner)$.

How does this analysis reflect on the ICT thesis? Horsten takes it that the reflection one has to engage in on one's base theory is not obligatory. Your acceptance of the theory, even unconditional, does not compel you to go through the reflective acts we have described, it only makes it rationally acceptable. You might even reflect to the extent that you revise your unconditional acceptance instead of coming to believe in the consistency of the theory. As Horsten correctly points out, this is the situation some practicing mathematicians find themselves in with regards to **ZFC**. They accept **ZFC**, at least pragmatically, but their acceptance is not unconditional due to their belief in the (slim) possibility that **ZFC** might derive a contradiction.

How does this analysis hold up? And how does it generalize to other principles we might come to accept through reflection like Rfn and RFN, or even the global reflection principle? With respect to the global reflection principle, Horsten echoes our observation that a concept of truth of a theory is not always given in the epistemic situation in which one finds oneself when one accepts a theory:

> What about the reflection process that can lead you to know a strong proof theoretic reflection principle, such as "everything that **PA** proves is true"? That reflection process is significantly more complex and requires a separate

> investigation. One issue is that you may not possess a concept of truth for arithmetical sentences at the start of your reflective journey: it is a difficult question how you come to acquire it. [Hor]

Horsten intends the analysis to require the least amount of conceptual baggage necessary, as the following quote makes clear:

> Perhaps you deeply distrust philosophy and all distinctively philosophical concepts. In particular, you may not believe that there is a concept of truth or of rational belief that you may legitimately use in your reasoning. Nevertheless, if you were to discover that **PA** is inconsistent, then as a mathematician you would (rightly) feel compelled to revise your mathematical commitments. [Hor]

I think Horsten's analysis of the process of reflection is accurate with regards to the situation and idealized mathematician described. Nevertheless, I think we can do with even less, and gain more. In the next section I will argue, along similar lines, that instrumental acceptance of a theory **S** is sufficient to come to accept its consistency, as well as the local and uniform reflection principle.

7.6 Commitment Without Truth

Before we extend Horsten's analysis, we give a bit more thought to the reflection on the formalization of provability in a theory. As we saw, deriving

$$\text{For all } \varphi \in L_{PA} : \mathbf{PA} \vdash \varphi \text{ iff } \mathbf{PA} \vdash Pr_{PA}(\ulcorner \varphi \urcorner),$$

requires the presupposition that **PA** is consistent and Σ_1-sound. Consistency and soundness are usually explained in terms of truth, so it is not immediately obvious how you come to see this without helping yourself to the concept of truth. At least consistency can be understood in purely syntactical terms. Consistency of a theory means that the theory does not derive a contradiction, that is, there is no φ for which $\mathbf{Th} \vdash \varphi \wedge \neg\varphi$, which in classical logic is equivalent to the theory not deriving every sentence φ by the principle of explosion. So, Horsten is right in pointing out that we don't need a concept of truth to reflect on our implicit commitment to consistency. The situation is different with regards to Σ_1-soundness, since this property rests on the arithmetical truth of a sentence. I think the truth-skeptic might still resist the charge of relying on a concept of truth by appealing to the equivalence between Σ_1-soundness and 1-consistency, defined as follows :

Definition 7.6.1. A system **S** in a language that expresses a numeral $\underline{n}$ for each natural number n is 1-consistent if and only if there is no Σ_1-formula $\exists x \varphi(x)$, so that $\mathbf{S} \vdash \exists x \varphi(x)$ and $\mathbf{S} \vdash \neg\varphi(\underline{n})$ for each numeral $\underline{n}$.

On the face of it, 1-consistency does not seem much easier to defend. After all, it seems that 1-consistency expresses syntactically that what we really care about are properties of the natural numbers, as represented by the numerals, and not any non-standard numbers. Don't we then require a (limited) concept of arithmetical truth for the reasoning involved in the reflective acts described? I think this doesn't have to be the case. To understand that **PA** is *about* the natural numbers in an essential way (and not about non-standard numbers) is sufficient to appreciate the property of 1-consistency as a presupposition of your cognitive project. The same reasoning holds to understand the property of ω-consistency in purely syntactical terms.

I now come to the analysis of how we could find the local and uniform reflection principle rationally acceptable if we accept a theory **S**. Contrary to Horsten, I will understand acceptance in a weaker way. It is sufficient for us to accept a theory **S** purely instrumentally, without believing in its truth. For comparison, this is how Van Fraassen (who is the source for the recognition that acceptance of a theory involves two components) characterizes scientific anti-realism:

> What does a scientist do then, according to these different positions? According to the realist, when someone proposes a theory, he is asserting it to be true. But according to the anti-realist, the proposer does not assert the theory to be true; he displays it, and claims certain virtues for it. These virtues may fall short of truth: empirical adequacy, perhaps; comprehensiveness, acceptability for various purposes. [Bas80]

For the anti-realist, accepting a scientific theory will be based on different notions of adequacy, rather than truth. The physicist who is deeply suspicious of quantum mechanics being true, will still use its predictions when determining the spectral lines of hydrogen. A different example closer to home is the theory $\mathbf{CT}[\mathbf{PA}]^-$. The theory is conservative over **PA**, but offers a non-elementary speed-up over **PA** with respect to its theorems [Fis14]. So, as a mathematician who is suspicious of the adequacy of $\mathbf{CT}[\mathbf{PA}]^-$ in formalizing the concept of truth, let alone the theory itself being true, there is instrumental value in using $\mathbf{CT}[\mathbf{PA}]^-$ to derive theorems of **PA**.

The claim is that in these cases, and other cases where a purely instrumental acceptance of a theory is involved, it is still rational to accept the local and uniform reflection principle, again purely instrumentally. Not only will the reflective process at no point require the concept of truth, you don't even need to believe the theory in question to be true, in stark contrast to the usual justification for accepting reflection principles. As in Horsten's analysis, we start from the idealized mathematician being in a state of innocence with regards to the theory they accept instrumentally. Your justification for accepting the theory is put in terms of the uses you can put the theory, not its supposed truth. In this state of innocence, you are able to reflect on the theory you have accepted. It is like stumbling upon a (Turing) machine that continuously produces theorems. You find the theorems so produced useful: the bridge you built based on some of the theorems still stands. Crucially, the machine is not a black box, its mechanism is open for you to inspect. You are able to distinguish the (extra-)syntactic elements involved, i.e. predicates, terms, axioms, formulas, and inference rules. The mechanism can even be adapted, you can add additional inputs (axioms) to the machine. Similar to Horsten's analysis you come to realize that the mechanism itself can be expressed syntactically: namely as a provability predicate $Pr_{Th}(x)$. The coding involved does not rely on knowing that **PA** is true, not even that **PA** is a theory about the natural numbers. As before, you come to the conclusion that the workings of the machine are correctly expressed by $Pr_{Th}(x)$, through proving that:

$$\text{For all } \varphi \in L_{Th} : \mathbf{Th} \vdash \varphi \text{ iff } \mathbf{Th} \vdash Pr_{Th}(\ulcorner \varphi \urcorner).$$

To do so, all that is required is to make explicit use of your presupposition that the theory is consistent, and 1-consistent, both of which are, again, syntactic notions. These presuppositions do rely on you being aware that **PA** is a theory (trying to) describe the structure of the natural numbers, since 1-consistency makes explicit usage of numerals. Moreover, you realize that the local reflection principle $Pr_{Th}(\ulcorner \varphi \urcorner) \rightarrow \varphi$ expresses your external understanding of the fact that for all $\varphi \in L_{Th}$: $\mathbf{Th} \vdash \varphi$ if $\mathbf{Th} \vdash Pr_{Th}(\ulcorner \varphi \urcorner)$. Now, given your instrumental acceptance of the theory, adding the local reflection principle is a matter of internalizing the fact that you believe the corresponding machine to be working correctly, i.e. that the theory is (1-)consistent. It is entirely rational to add the local reflection principle to the theory, but you are not obliged to do so. In some situations, it might even be preferable to abstain from doing so. The addition of the reflection principle to the theory should itself be understood instrumentally. For example, if you are interested in the theory's theorems *an sich*, it would be counter-productive to add the local reflection principle. As a mathematician studying the theorems of **EA**, it would be a mistake to add the uniform reflection principle, since then (by Theorem 2.3.5) you would be obtaining the theorems of **PA** instead.

Now, the uniform reflection principle is different in kind from the local reflection principle. In its usual form, it is specific to arithmetical theories (or extensions thereof) in a way that the local reflection principle is not. First, notice that the uniform reflection principle requires an even stronger presupposition than 1-consistency, namely ω-consistency. This presupposition is a natural one, but one can imagine a situation in which one's cognitive project does not rest on it. Secondly, the uniform reflection principle, being a kind of formalized ω-rule, makes explicit that **PA** is a theory about the natural numbers. Extending the analysis of reflection given here to different theories, which do not necessarily have a name for each object of the intended domain, would again require a different approach, in terms of satisfaction.

I've argued that the local and uniform reflection principle require only a minor amount of additional justification beyond the justification necessary to accept the consistency of one's base theory. On the other hand, I think the global reflection principle requires much more in the way of justification, contrary to Kreisel's suggestion that we motivate Rfn and RFN by virtue of our acceptance of the global reflection principle. We come to understand Rfn and RFN through a reflective process which does not require the concept of truth, nor the need to believe in the truth of the base theory. Coming to accept the global reflection principle on the other hand, requires you to have a grasp of a concept of truth which minimally includes the T-schema. As in the example of $\mathbf{CT}_{\mathbf{1}}^{-}$ (see Definition 7.4.1), it is not evident what this concept should be for a given base theory.

We've seen how the addition of local and uniform reflection principles to a theory is rationally acceptable, given one's instrumental acceptance of a theory. Moreover, this can be realized without using the concept of truth, nor does it require belief in the truth of the theory. But we shall know our analysis by its fruits. Does it explain the cases of the epistemically stable theories we have seen? Because we now have a more fine-grained analysis of the reflective process, and the justification involved therein, we can see exactly where Tait's finitism and Isaacson's first orderism lacks justification for one of the steps.

In the case of Tait's finitism , the issue is that the finitist cannot 'step outside' and reflect on his own workings as mathematician. It is not enough for the finitist to presuppose the (1-) consistency of **PRA** to realize the validity of his procedure. In order to understand why that is, let us look at how Tait understands finitist mathematics. First of, the finitist understands a few basic *constructions*. These constructions are means to build an element b of type B, denoted by $b : B$, from an element a of type A, denoted as $a : A$. The notion of construction is primitive, and is used in the place of the notion of a function, which is a non-finitary object (being in general defined over an infinite domain). Some of the constructions the finitist grasps are completely general, regardless of the objects under consideration. For example, the notion of composition is grasped as giving a new construction $h : A \to C$ from $f : A \to B$ and $g : B \to C$, where $ha := g(fa)$. Other constructions can only be understood as being implicit in the finitist understanding of natural numbers: the constructions corresponding to the constant function, successor function, and constructions based on iteration. In particular, the finitist finds the following construction f acceptable, expressed by:

$$f0 := k;$$
$$fn' := g(fn),$$

where $f : N \to A$ is defined from the construction $g : A \to A$, and the object $k : A$. From the outside looking in, we understand this as the implicit definition of a function through primitive recursion. For the finitist, who does not think the notion of function legitimate, this is just a shorthand description for how a particular fn is constructed from $f(n-1)$, where they see that the sequence $n, n-1, \ldots, 0$ must terminate at 0. To put the point more bluntly, the finitist has no understanding of the *general validity* of function definition through primitive recursion, only particular instances are seen to be finitistically acceptable. Similarly, the finitist is satisfied with every derivation given by the Turing machine representing **PRA**, but is unable to accept

$$\text{For all } \varphi \in L_{PRA} : \mathbf{PRA} \vdash \varphi \text{ iff } \mathbf{PRA} \vdash Pr_{PRA}(\ulcorner \varphi \urcorner),$$

since this would require being able to recognize the validity of the machine's operation in general rather than only in the particular.

The case of Isaacson's first-orderism is in my view different from the epistemic situation of Tait's finitist. Whereas the finitist lacked justification for seeing the local reflection principle as acceptable, Isaacson has justification for accepting the local reflection principle, but chooses not to. It is rationally acceptable, but instrumentally not acceptable, since he is interested in **PA** *an sich* rather than as a formal system. Here is what Isaacson has to say on completing **PA** through the addition of true, but underivable sentences:

> We might consider whether, in view of its truth and independence from **PA**, we should adopt the Gödel sentence for **PA**, call it G, as a new axiom of arithmetic. Such a move would be unnatural. An axiom in this context should be an evident truth, in the terms in which it is expressed. But the truth of this statement, as a statement of arithmetic, is not directly perceivable. $\mathbf{PA} + G$ would not constitute, in this way, a purely arithmetical extension of **PA**. [Isa87, p.159]

The reason the truth of G is not perceivable as true of arithmetic is that it hinges on our understanding of coding: the truth of G is dependent on its link with syntactical properties of **PA**. To return to our machine metaphor: Isaacson considers the principles the Turing machine uses to derive theorems of **PA** to be arithmetical principles. But the Gödel sentence's truth depends on a principle (i.e. the local reflection principle) that says something about the machine itself, rather than arithmetic. It is a truth of **PA** as a formal system, rather than as a (non-exhaustive) set of true arithmetical sentences:

> The arithmetic of the natural numbers can mimic quite other situations. If the truths in the language of arithmetic which express these mimic-situations are to be seen as true, that will depend not on the principles which generate our understanding of the natural numbers, but on those which apply to the situation which is mimicked, and which reveal the coded connection between them. [Isa87, p.159]

In my reading, Isaacson has no issue seeing the reflection principle as true, but since it is not *directly* true, he cannot accept it. He is instrumentally interested only in the directly perceivable truths, not the ones that require the kind of reflection we described (although he is perfectly capable of doing so).

Finally, we can answer the question of what, if any, reflection principles can mean to a truth deflationist. Remember the initial quandary: **CT** as a theory is truth-theoretically productive in a way that **TB** is not. This is a boon for deflationists, since truth is put forward as an expressive device first and foremost. On the other hand, **CT** is not syntactically conservative over its base theory **PA**, which has been considered a challenge to rhyme with deflationist tenets, under the understanding of deflationism that truth should be linguistically productive, but not *epistemically* productive. Reflection principles offer one response: starting with a conservative truth-theory **S**, we can recover the truth-theoretic power of **CT** by reflecting over **S**. We have seen several examples of this phenomenon. Corollary 6.1.2.1 shows that over $\mathbf{CT}^-$ the reflection principle expressing that "logic is true for sentences in L_{PA}" is sufficient to derive to that "**PA** is true", and hence, the consistency of **PA**. We also saw that "logic is true for sentences in L_T" over $\mathbf{CT}^-$ is sufficient to recover **CT**. Should these results, resting on an analysis like we gave previously, constitute the defense of the truth deflationist to the conservativity challenge? On the one hand the weakness of the reflection principle involved is appealing, the presupposition that logic is true being one that few would deny. On the other hand, the deductive weakness of $\mathbf{CT}^-$ derives from the lack of extended induction. As such, $\mathbf{CT}^-$ does not represent a stable epistemic position: the induction schema should be understood as open-ended. More promising is the analysis of Section 6.2. Although the reflection principles at play are more involved, and require more justification, this justification will *usually* be available for those who have accepted the theory. In particular, we have seen that $\mathbf{TB}^-$, and a fortiori **TB**, will recover **CT** after two iterations of uniform reflection (see Corollary 6.2.4.1). Uniform reflection requires the presupposition of $\omega-$consistency of the theory. This is of course stronger than mere consistency, but not very much so, in the sense that it is a natural requirement to have for a theory one has accepted instrumentally. The base theory **TB** has two advantages as a starting point compared to $\mathbf{CT}^-$. The first advantage is that it is conceptually even simpler than **CT**: it is hard to see what truth could be if it did not at least include the T-schema. The second advantage is that induction is extended to the truth predicate in **TB**, in accordance with the principle that the induction schema is open-ended. We have argued that the acceptance of uniform reflection can be justified without recourse to an a priori concept of truth, nor the belief in the truth of the theory one accepts. This is crucial, since the process of recovering **CT** from **TB** is a process of unfolding one's concept of truth, and so can not rest on justification in terms of truth.

In conclusion, I take it that the truth deflationist has a coherent response to the conservativity challenge. The basic, but incomplete theory of truth is given by **TB**. Similar to how Kreisel (cfr. Section 3.3) saw **PRA** as an approximation to finitism, which required unfolding to obtain a theory which encompassed all finitist theorems, **TB** requires unfolding to approach the full concept of truth. The justification for doing so is inherent in the acceptance of **TB**. We are neither in the situation of Tait, where reflection on our base theory is impossible because we don't have the requisite notion of function, or in the situation of Isaacson, where there is no instrumental justification for accepting additional, independent reflection principles. We are justified in going beyond **TB** through uniform reflection because our acceptance of **TB** *presupposes* that it is consistent, and ω-consistent. Doing so, we recover **CT** after two iterations of uniform reflection. It turns out that compositional truth-reasoning is implicit in the T-schema, given that one is willing to reflect on the notion of provability. The non-conservativity of **CT** over the base theory is on this analysis not a feature of our truth theory, but a feature of our *reasoning*. It is by meta-theoretic reasoning over **TB** that we obtain **CT**, and a host of other results that are independent of the base theory, rather than by the weight of a substantial concept of truth.

CHAPTER 8

Conclusion

This thesis was concerned with evaluating a recent response to the conservativity challenge in axiomatic theories of truth. We have seen that the best (typed) theories on offer, **TB** formalizing the T-schema, and **CT** formalizing the compositionality of our truth-theoretic reasoning, both *prima facie* fail to meet this challenge. While **TB** is a conservative theory of truth, it fails to derive the kind of generalizations that have been touted as part of the linguistic function of truth. These generalizations are derivable within the compositional theory **CT**, but at the expense of being non-conservative over **PA** (e.g. proving the consistency of **PA**). The onus is on the truth deflationist to explain how this rhymes with deflationism. One option is to bite the bullet and admit that non-conservativity is a property of our best truth theories, and that truth deflationism should be understood differently. Recently, a second option has been put forward: explain the non-conservativity of one's preferred truth theory as the result of adjoining reflection principles, with independent justification, to a conservative basic truth theory.

Evaluating this option presented two challenges: showing that the approach is feasible technically, and evaluating the philosophical justification for adjoining independent reflection principles. First off, there are many permutations of conservative truth theories and reflection principles to consider. In Chapter 6 we went over the results in the literature on adjoining reflection principles to $\mathbf{CT}^-$, which is conservative by lacking extended induction, and **TB**. We saw that even a minimal amount of reflection, amounting to "first-order logic is true", is sufficient to obtain a non-conservative theory over $\mathbf{CT}^-$ (see Theorem 6.1.2 and its corollary). We have also seen that adjoining uniform reflection over $\mathbf{CT}^-$, and two iterations of uniform reflection over **TB** is sufficient to obtain **CT** respectively. We have argued that while the results over $\mathbf{CT}^-$ are interesting for their own sake, and were useful in order to evaluate the philosophical justification for reflection (see Section 7.3), $\mathbf{CT}^-$ is not appropriate as a basic truth theory. The induction principle is an open-ended schema, and should be extended to any newly introduced predicate. Hence $\mathbf{CT}^-$ cannot be a proper ground in itself for one's approximate formalisation of deflationist truth. Instead, I take it that **TB** is an appropriate, and minimal, conservative truth theory which any truth deflationist will have to accept. If one is then also willing to accept the uniform reflection principle over an accepted theory, we obtain **CT**.

In chapter 7, we then considered the literature on justification, or the lack thereof, for the implicit commitment thesis (ICT). According to this thesis, by accepting a theory **Th**, we are also implicitly committed to certain reflection principles. I have argued that the usual defense of ICT, based on an implicit or explicit understanding of the *truth* of the accepted theory **Th**, is flawed. Particularly in the case of truth theories, we can not rely on such a prior understanding being available. Instead, I have defended, along lines similar to Horsten [Hor], that we should understand the justification for the local and uniform reflection principle as deriving from our presupposition of a purely syntactical property of the theory. It is sufficient to accept a theory instrumentally for the acceptance of reflection principles to be justified. This kind of acceptance surely is the case for the truth deflationist accepting **TB**.

As such, I think the truth deflationist is vindicated in defending a non-conservative theory of truth. Truth is indeed simple, merely the acceptance of the T-schema, but can be unfolded through reflection to encompass the usual uses of truth. That a full(er) theory of truth is non-conservative is a byproduct of the meta-theoretic reasoning we engage in, rather than something inherent in the concept of truth.

Index

Bibliography

[Aye35] A. J. Ayer. The criterion of truth. *Analysis*, 3(1/2):28–32, 1935.

[Bas80] C. Van Fraassen Bas. *The Scientific Image*. Oxford University Press, 1980.

[BC14] David Bourget and David J Chalmers. What do philosophers believe? *Philosophical studies*, 170(3):465–500, 2014.

[Bek05] Lev D. Beklemishev. Reflection principles and provability algebras in formal arithmetic. *Russian Mathematical Surveys*, 60(2):197–268, April 2005.

[Boo94] George S. Boolos. *The Logic of Provability*. Cambridge University Press, February 1994.

[Cie10] Cezary Cieśliński. Truth, conservativeness, and provability. *Mind*, 119(474):409–422, April 2010.

[Cie17] Cezary Cieśliński. *The Epistemic Lightness of Truth*. Cambridge University Press, November 2017.

[CM11] P. Curd and R.D. McKirahan. *A Presocratics Reader (Second Edition): Selected Fragments and Testimonia*. Hackett Classics Series. Hackett Publishing Company, 2011.

[Dav67] Donald Davidson. Truth and meaning. *Synthese*, 17(3):304–323, 1967.

[Dav69] Donald Davidson. True to the facts. *The Journal of Philosophy*, 66(21):748–764, 1969.

[Dea14] Walter Dean. Arithmetical Reflection and the Provability of Soundness. *Philosophia Mathematica*, 23(1):31–64, 12 2014.

[End77] H.B. Enderton. *Elements of Set Theory*. Elsevier Science, 1977.

[EV15] Ali Enayat and Albert Visser. *New Constructions of Satisfaction Classes*, pages 321–335. Springer Netherlands, Dordrecht, 2015.

[Fef60] Solomon Feferman. Arithmetization of metamathematics in a general setting. *Fundamenta Mathematicae*, 49(1):35–92, 1960.

[Fef62] Solomon Feferman. Transfinite recursive progressions of axiomatic theories. *The Journal of Symbolic Logic*, 27(3):259–316, 1962.

[Fef64] Solomon Feferman. Systems of predicative analysis. *The Journal of Symbolic Logic*, 29(1):1–30, 1964.

[Fef91] Solomon Feferman. Reflecting on incompleteness. *The Journal of Symbolic Logic*, 56(1):1–49, 1991.

[Fef09] Solomon Feferman. *Predicativity*. Oxford University Press, September 2009.

[Fie99] Hartry Field. Deflating the conservativeness argument. *Journal of Philosophy*, 96(10):533–540, 1999.

[Fis14] Martin Fischer. Truth and speed-up. *Review of Symbolic Logic*, (2):319–340, 2014.

[FNH17] Martin Fischer, Carlo Nicolai, and Leon Horsten. Iterated reflection over full disquotational truth. *Journal of Logic and Computation*, 27(8):2631–2651, 2017.

[FP03] G. Frege and G. Patzig. *Logische Untersuchungen*. Beitrage Zum Siedlungs-Und Wohnungswesen. Vandenhoeck und Ruprecht, 2003.

[Gen35] Gerhard Gentzen. Untersuchungen über das logische schließen. *Mathematische Zeitschrift*, 39(1):176–210, 1935.

[Gir11] J.Y. Girard. *The Blind Spot: Lectures on Logic*. European Mathematical Society, 2011.

[Hal01a] Volker Halbach. Disquotational truth and analyticity. *Journal of Symbolic Logic*, 66(4):1959–1973, December 2001.

[Hal01b] Volker Halbach. How innocent is deflationism? *Synthese*, 126(1/2):167–194, 2001.

[Hal14] Volker Halbach. *Axiomatic Theories of Truth*. Cambridge University Press, 2014.

[Hil26] David Hilbert. Über das unendliche. *Mathematische Annalen*, 95(1):161–190, December 1926.

[HL17] Leon Horsten and Graham E Leigh. Truth is simple. *Mind*, 126(501):195–232, 2017.

[Hor] Leon Horsten. On reflection. `https://www.academia.edu/37686643/On_Reflection`. Accessed: 2020-10-09.

[Hor98] Paul Horwich. *Truth*. Oxford University Press, December 1998.

[Hor11] Leon Horsten. *The Tarskian turn: Deflationism and Axiomatic Truth*. The MIT Press, 2011.

[HP98] Petr Hájek and Pavel Pudlák. *Metamathematics of first-order arithmetic*. Perspectives in mathematical logic. Springer, Berlin, 2nd edition, 1998.

[Iem20] Rosalie Iemhoff. Intuitionism in the Philosophy of Mathematics. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, fall 2020 edition, 2020.

[Isa87] Daniel Isaacson. Arithmetical truth and hidden higher-order concepts. In *Logic Colloquium '85*, volume 122 of *Studies in Logic and the Foundations of Mathematics*, pages 147 – 169. Elsevier, 1987.

[Kay91] Richard Kaye. *Models of Peano Arithmetic (Oxford Logic Guides)*. Clarendon Press, February 1991.

[Ket99] Jeffrey Ketland. Deflationism and tarski's paradise. *Mind*, 108(429):69–94, 1999.

[Kim18] I. Kimhi. *Thinking and Being*. Harvard University Press, 2018.

[KKHL81] Henryk Kotlarski, Stanislaw Krajewski, and Alistair H. Lachlan. Construction of satisfaction classes for nonstandard models. *Canadian Mathematical Bulletin*, 24(3):283–293, 1981.

[KL68] Georg Kreisel and Azriel Lévy. 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, 1968.

[Kle52] Stephen Cole Kleene. *Introduction to Metamathematics*. Wolters-Noordhoff and North-Holland Publishing, seventh edition, 1952.

[Kot86] Henryk Kotlarski. Bounded induction and satisfaction classes. *Zeitschrift für Mathematische Logik und Grundlagen der Mathematik*, 32(31-34):531–544, 1986.

[Kre58] Georg Kreisel. Ordinal logics and the characterization of informal concepts of proof. In *Proceedings of the International Congress of Mathematicians*, volume 14, page 21, 1958.

[Kre67] Georg Kreisel. Informal rigour and completeness proofs. In Imre Lakatos, editor, *Problems in the Philosophy of Mathematics*, volume 47 of *Studies in Logic and the Foundations of Mathematics*, pages 138 – 186. Elsevier, 1967.

[Kre68] Georg Kreisel. A survey of proof theory. *The Journal of Symbolic Logic*, 33(3):321–388, 1968.

[Kre70] Georg Kreisel. Principles of proof and ordinals implicit in given concepts. In *Studies in Logic and the Foundations of Mathematics*, volume 60, pages 489–516. Elsevier, 1970.

[Kre87] Georg Kreisel. Church's thesis and the ideal of informal rigour. *Notre Dame Journal of Formal Logic*, 28(4):499–519, 10 1987.

[Kri75] Saul Kripke. Outline of a theory of truth. *Journal of Philosophy*, 72(19):690–716, 1975.

[Lei15] Graham E. Leigh. Conservativity for theories of compositional truth via cut elimination. *The Journal of Symbolic Logic*, 80(3):845–865, 2015.

[McG92] Vann McGee. Maximal consistent sets of instances of tarski's schema. *Journal of Philosophical Logic*, 21(3), August 1992.

[McG06] Van McGee. In praise of the free lunch. In Thomas Bolander, Vincent F. Hendricks, and Stig Andur Pedersen, editors, *Self-Reference*. Center for the Study of Language and Information, 2006.

[NP19] Carlo Nicolai and Mario Piazza. The implicit commitment of arithmetical theories and its semantic core. *Erkenntnis*, 84(4):913–937, 2019.

[Qui86] Willard Van Orman Quine. *Philosophy of Logic*. Harvard University Press, Cambridge, Massachusetts, 1986.

[Ram27] Frank Plumpton Ramsey. Facts and propositions. *Aristotelian Society Supplementary Volume*, 7(1):153–170, 1927.

[Rat99] Michael Rathjen. The realm of ordinal analysis. *London Mathematical Society Lecture Note Series*, pages 219–280, 1999.

[Rog87] Hartley Rogers. *Theory of recursive functions and effective computability*. MIT Press, Cambridge, Mass, 1987.

[RS20] Michael Rathjen and Wilfried Sieg. Proof Theory. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, fall 2020 edition, 2020.

[Rus99] Bertrand Russell. *The Problems of Philosophy*. (Hackett classics). Dover Publications, 1999.

[Sch65] Kurt Schütte. Predicative well-orderings. In *Studies in Logic and the Foundations of Mathematics*, volume 40, pages 280–303. Elsevier, 1965.

[Sha98] Stewart Shapiro. Proof and Truth: Through Thick and Thin. *The Journal of Philosophy*, 95(10):493, October 1998.

[SKH52] Heinrich Scholz, Georg Kreisel, and Leon Henkin. Problems. *The Journal of Symbolic Logic*, 17(2):160–160, 1952.

[Smi13] P. Smith. *An Introduction to Gödel's Theorems*. Cambridge Introductions to Philosophy. Cambridge University Press, 2013.

[Smo77] C. Smorynski. The incompleteness theorems. In Jon Barwise, editor, *Handbook of Mathematical Logic*, volume 90 of *Studies in Logic and the Foundations of Mathematics*, pages 821 – 865. Elsevier, 1977.

[Str13] Andrea Strollo. Deflationism and the invisible power of truth. *Dialectica*, 67(4):521–543, December 2013.

[Sza18] Jan Szaif. Plato and aristotle on truth and falsehood. In *The Oxford Handbook of Truth*, page 9. Oxford University Press, 2018.

[Tai81] W. W. Tait. Finitism. *The Journal of Philosophy*, 78(9):524, September 1981.

[Tar33] Alfred Tarski. The concept of truth in the languages of the deductive sciences. (polish). 1933.

[Ten02] Neil Tennant. Deflationism and the gödel phenomena. *Mind*, 111(443):551–582, 2002.

[Tur36] Alan Mathison Turing. On computable numbers, with an application to the entscheidungsproblem. *J. of Math*, 58:345–363, 1936.

[Wey18] Hermann Weyl. *Das kontinuum*. Veit, 1918.

[WŁ17a] Bartosz Wcisło and Mateusz Łełyk. Notes on bounded induction for the compositional truth predicate. *The Review of Symbolic Logic*, 10(3):455–480, March 2017.

[WŁ17b] Bartosz Wcisło and Mateusz Łełyk. Strong and weak truth principles. *Studia Semiotyczne—English Supplement Volume XXIX*, page 107, 2017.

[Wri04] Crispin Wright. Warrant for nothing (and foundations for free)? *Aristotelian Society Supplementary Volume*, 78(1):167–212, 2004.

[Zac07] Richard Zach. Hilbert's program then and now. *Philosophy of Logic*, pages 411–447, 2007.