Natural Axiomatic Theories and Consistency Strength: A Lakatosian Approach to the Linearity Conjecture

MSc Thesis *(Afstudeerscriptie)*

written by

Lide Grotenhuis
(born October 11th, 1994 in 's Gravenhage, The Netherlands)

under the supervision of **Dr. Luca Incurvati** and **Dr. Giorgio Sbardolini**, and submitted to the Board of Examiners in partial fulfillment of the requirements for the degree of

MSc in Logic

at the *Universiteit van Amsterdam.*

Date of the public defense:	**Members of the Thesis Committee:**
August 29, 2022	Dr. Maria Aloni *(chair)*
	Dr. Luca Incurvati *(supervisor)*
	Dr. Giorgio Sbardolini *(supervisor)*
	Prof. dr. Arianna Betti
	Prof. dr. Benedikt Löwe
	Dr. Levin Hornischer

Institute for Logic, Language and Computation

Abstract

In the literature on relative consistency results, one often encounters the claim that all natural axiomatic theories are linearly ordered in terms of consistency strength. Without a precise definition of a natural theory, it is not clear how to assess the truth of this claim or how to judge whether the known instances of nonlinearity constitute genuine counterexamples to it. The general aim of this thesis is to take a first step towards such a precise definition. To this end, the thesis consists of two parts. First, after arguing that a pursuit of such a definition is worthwhile, I develop a method for working towards such a definition. This method is primarily inspired by Lakatos' approach to mathematical concept-formation, whose main tenet is that mathematical concepts develop in response to the emergence of counterexamples. Second, I apply the method and analyze the known instances of nonlinearity, including those recently suggested by Hamkins. Building on this analysis, I develop the following tentative definition: an axiomatic theory is natural if its axioms do not carry meta-information and if the theory is presented in a surveyable manner.

Acknowledgements

This thesis would not have come to be without the support of many people.

I want to thank my supervisors, Luca Incurvati and Giorgio Sbardolini, for their invaluable guidance and for giving me the tools and opportunity to develop my own ideas. Luca, thank you for introducing me to the philosophy of mathematics and for giving me the courage to write a thesis in philosophy. Giorgio, thank you for the long and fun conversations and for reminding me to 'have fun with it' when I felt stressed or worried.

I would also like to thank the committee members, Maria Aloni, Arianna Betti, Benedikt Löwe and Levin Hornischer, for taking the time to read my work and for their insightful questions and comments during my defense.

I am also grateful to my wonderful and intelligent friends who made the MoL so very enjoyable. Jeremy, thank you for great conversations over many cups coffee. Daniël, thank you for being the best homework and project partner and for reading my thesis - hopefully we get to work together again in the future. A very special thanks must go to Bart, who endured more hours of me rubberducking then I can count and who never fails to help me persevere when I get stuck.

My dear family and friends - my parents, Igor, Romi, Julia, Esther, Anne, Eric - I thank for supporting me in whatever I do. And finally, Floor, for all your love.

Contents

Introduction **4**

I Problem and methodology 7

1 The linearity conjecture **8**
1.1 Consistency strength . . . 8
1.2 The linearity phenomenon . . . 9
1.3 Instances of nonlinearity . . . 10
1.4 How to proceed? . . . 12

2 Concept-formation in mathematics **15**
2.1 On explication . . . 15
2.2 Explication via conceptual analysis: the case of Church's thesis . . . 18
2.2.1 Proving the thesis? . . . 19
2.3 A conceptual analysis of naturalness? . . . 22
2.4 The Lakatosian approach to concept-formation . . . 24
2.4.1 Lakatos' *Proofs and Refutations* . . . 25
2.4.2 Concept-formation in PR . . . 27
2.5 Outline of our method . . . 28

II Application of the Lakatosian approach 30

3 Analyzing the counterexamples **31**
3.1 Mathematical preliminaries . . . 31
3.2 Counterexample 1: the Rosser sentence . . . 36
3.2.1 Proof analysis . . . 38
3.3 Counterexample 2: Representing numbers by computations . . . 39
3.3.1 Proof analysis . . . 42
3.4 Counterexample 3: Cautious enumerations . . . 43
3.4.1 Proof analysis . . . 45

4 Proof-generated concepts of natural theory **47**
4.1 Self-reference . . . 47
4.2 Axioms carrying meta-information: Proof-generated Concept 1 . . . 49
4.3 On the intensionality of PC1 . . . 51

4.4 Surveyable presentation: Proof-generated Concept 2 54
4.5 Taking stock 57

Concluding remarks **58**

Bibliography **59**

Introduction

By identifying significant restrictions on what axiomatic theories can prove, Gödel's famous incompleteness theorems induced a broad investigation into the logical strength of axiomatic theories. One part of this investigation consists of comparing the consistency strength of such theories, leading to the so-called relative consistency results. Over the past few decades, researches in this field have noted an interesting pattern, namely that the consistency strength of our common mathematical theories tends to be *comparable*, that is, given two such theories, the consistency of one of them tends to imply the consistency of the other. There also exist well-known constructions of theories whose consistency strength is not comparable; however, those constructions are generally viewed as 'unnatural' by the mathematical community. In the literature, one therefore often meets the claim that all *natural* axiomatic theories are linearly ordered in terms of consistency strength.[1] However, as we lack a precise definition of what constitutes a 'natural theory', this claim is an informal one.

The labeling of particular objects or constructions in mathematics as natural is an interesting phenomenon. While the common-sense understanding of the word 'natural' appears to be something along the lines of 'arising in nature' and 'not artificial or man-made'[2], this common-sense understanding hardly seems fit to justify the labeling of mathematical constructions such as proofs, axioms or operations as natural. Nevertheless, the use of the label 'natural' is abundant in mathematics, both as an informal adjective or as part of a formal definition. In the context of mathematics, 'natural' seems to mean something like 'of the correct form' or 'appropriate for the task at hand'. Accounts of the informal use of the notion of naturalness in mathematics, such as that given in [60] and [50], suggest that by labeling a construction as natural, the mathematician is making a claim about the correctness and expected fruitfulness of this construction.

One is therefore led to the question: what is meant by the claim that the consistency strengths of 'natural' axiomatic theories form a linear order? In particular, does this claim reflect a mere observation that our common or fruitful theories tend to show this behaviour, or does it entail something stronger? Interestingly, the treatment of this claim in the literature seems to indicate the latter. For example, work by James Walsh and Antonio Montalbán in [45] and [65] seems aimed at *proving* this claim, whereas Joel Hamkins [22] appears to be trying to refute the claim by providing instances of nonlinearity that, he argues, are natural. These treatments indicate that the claim that the consistency strength of natural axiomatic theories forms a linear order is a *conjecture* rather than a mere *observation*. In this thesis, I will therefore refer to this claim as *the linearity conjecture*.[3]

If this claim is indeed to be taken as a conjecture, as I argue is the case, then one would like

[1] In Chapter 1, I will refer the reader to a number of places where this claim is made; these will include [14], [32], [54], [44], [45], [58] and [65].

[2] See for example the Oxford English Dictionary, which defines the adjective 'natural' as: 'existing in or derived from nature; not made or cause by humankind'.

[3] To my knowledge, the only one who as explicitly referred to this claim as a conjecture is John Steel in [58].

to assess its truth or falsity. In particular, one would like to asses whether the known instances of nonlinearity, such as those introduced by Hamkins, form counterexamples to the conjecture. However, without a precise definition of a natural theory, it is not clear how one could go about this.

The general aim of this thesis will be to make some first steps towards such a precise definition. It must be recognized that there is no clear protocol in place for such an endeavor; a secondary aim of this thesis will therefore be to develop a well-motivated method for obtaining a more robust definition of naturalness in the context of the linearity conjecture. This thesis therefore naturally falls into two parts. In the first part, I will introduce the object of study, namely the linearity conjecture, and formulate a method for making the notion of a natural theory in the context of this conjecture more precise; this method will then be applied in the second part.

Outline

This thesis is structured as follows. In Chapter 1, I will explain the content of the linearity conjecture and describe the context in which it emerged. Subsequently, through quotations I will try to shed some light on the attitude of the mathematical community towards the conjecture and the known instances of nonlinearity. It will be concluded that there appears to be a strong conviction that the conjecture carries some truth and that the well-known instances of nonlinearity are unnatural.

In Chapter 2, we work towards finding a method for our aim of making the intuitive concept of natural theory more precise. Using the philosophy of mathematical practise as general framework, we start by reviewing Carnap's account of explication, which is to be understood as the general process of transforming an informal concept (the *explicandum)* into a precise concept (the *explicatum)*. Subsequently, we discuss two different cases of mathematical explication, namely that of the concept of effective computability as referred to in Church's thesis and that of the concept of polyhedron as described by Lakatos' in his *Proofs and Refutations*. According to Lakatos, mathematical concepts tend to develop in response to counterexamples to preliminary conjectures. In particular, Lakatos argues that the most fruitful concepts are those that arise from careful analysis of an (informal) proof; Lakatos refers to such concepts as *proof-generated concepts*. It will be argued that the Lakatosian approach seems particularly suitable for the task at hand, and I will formulate a method based on this approach that consists of two main steps: (1) formulate a proof-generated concept of natural theory based on an analysis of the known instances of nonlinearity, and (2) assess whether the proof-generated concept is similar to the intuitive notion.

We will then move on the second part of this thesis, in which we apply the outlined method. Chapter 3 will consist of a detailed exposition of the known instances of nonlinearity. These include the instance involving the Rosser construction, which seems to be well-known in the mathematical community, but also the two new instances that were recently introduced by Hamkins [22]. By analyzing the proofs of nonlinearity, we will attempt to identify the key characteristics of the theories involved that are exploited in these proofs.

Following the proof analyses of Chapter 3, in Chapter 4 I will formulate proof-generated concepts of natural theory. Based on the first two instances of nonlinearity, I will propose Proof-generated concept 1, which defines a natural theory as one whose axioms do not carry meta-information. Subsequently, based on the third instance of nonlinearity, I propose Proof-generated concept 2, which defines a natural theory as one that is presented in a surveyable manner. It will argued that these proof-generated concepts (1) dismiss the presented instances of nonlinearity as counterexamples to the linearity conjecture, (2) are of a static rather than a

dynamic nature, and (3) are similar (in the Carnapian sense) to the intuitive notion of a natural theory as one that 'arises in practice' and 'has a genuinely mathematical idea to it'.

Part I

Problem and methodology

Chapter 1

The linearity conjecture

In this chapter, I will introduce our main object of study: the linearity conjecture for the consistency hierarchy of natural axiomatic theories. The notion of consistency strength will be defined and the context in which the conjecture has arisen will be discussed. Through quotations, I will try to shed some light on the attitude of the mathematical community towards the conjecture. Subsequently, we will discuss what an investigation of the conjecture could consist of and formulate the particular aim of this thesis.

1.1 Consistency strength

Before we can state the conjecture, we will discuss the notion of consistency strength. This notion became of interest in response to Gödel's famous incompleteness theorems. These theorems state the following:

Theorem 1.1. *(First incompleteness theorem) Any nice axiomatic theory T is incomplete, that is, there exists a sentence ϕ in the language of T such that $T \nvdash \phi$ and $T \nvdash \neg\phi$.*

Theorem 1.2. *(Second incompleteness theorem) Any nice axiomatic theory T cannot prove its own consistency, that is, $T \nvdash \mathrm{Con}(T)$.*

Here a nice theory is one that is consistent, has a recursive axiomatization[1] and in which we can carry out basic arithmetical reasoning, that is, at the level of *Peano arithmetic*, denoted by PA.[2] The third criterion enables one to arithmetize the syntax of the language of T. In particular, it enables one to encode each formula ϕ by a some natural number $\ulcorner\phi\urcorner$ and to construct a provability predicate for T, which is a formula $\mathrm{Prov}_T(x)$ such that $\mathrm{Prov}_T(\ulcorner\phi\urcorner)$ is true precisely when ϕ is provable in T. Using the provability predicate, the sentence $\mathrm{Con}(T)$ can be defined as $\neg\,\mathrm{Prov}_T(\ulcorner\bot\urcorner)$. It then follows that $\mathrm{Con}(T)$ is true if and only if T cannot prove a contradiction, which in turn is equivalent to the consistency of T. In order for the second incompleteness theorem to hold, the formula $\mathrm{Prov}_T(x)$ should satisfy some particular conditions known as the *Hilbert-Bernays-Löb conditions*.[3] For any nice theory, it is possible to construct such a provability predicate.

[1] See chapter 3 for the definition of a recursively axiomatized theory.

[2] In fact, it suffices to be able to carry out the arithmetic reasoning contained in the weaker arithmetical system known as *Robinson arithmetic*.

[3] We will revisit the construction of the provability predicate and the HBL-conditions in detail in chapter 3.

Assuming that we want our mathematical theories to be nice, which seems to be a quite modest request, the incompleteness theorems tell us that the ultimate mathematical theory does not exist: no matter how strong our axioms, there will always be statements that we cannot prove or refute. Moreover, if our theory is consistent, then we will not be able to show this within the theory itself. Therefore, in order to establish the consistency of a mathematical theory, we will need to consult an even stronger one. We thus arrive at a plethora of theories that can be compared in terms of their consistency strength.

Definition 1.3. *Let T and S be nice theories. We say that the consistency strength of T is lower than or equal to the consistency strength of S, denoted by $T \leq S$, if the sentence* $\mathrm{Con}(S) \to \mathrm{Con}(T)$ *is provable over some suitable base theory.*

A proof of $\mathrm{Con}(T) \to \mathrm{Con}(S)$ from some suitable base theory is called a *relative consistency result*. By 'suitable', we generally mean a theory that is weak in comparison to the theories S and T. After all, the weaker the base theory, the more convincing the relative consistency result. For many relative consistency results, it suffices to take the theory PRA of primitive recursive arithmetic as base. The reasoning within PRA is generally considered to be finitist, and thus consistency results from this base theory ought to convince even the finitist mathematician.[4]

As usual, we will write $S < T$ if and only if $S \leq T$ and $T \not\leq S$. In case we have both $S \leq T$ and $T \leq S$, we say that T and S are *equiconsistent* and write $S \equiv_{\mathrm{Con}} T$. By the second incompleteness theorem, note that if T is a nice theory that extends the base theory, then a sufficient condition for $T \not\leq S$ would be $T \vdash \mathrm{Con}(S)$. Indeed, if we were to have $T \vdash \mathrm{Con}(S)$ and $T \leq S$, then the base theory and thereby T would prove $\mathrm{Con}(S) \to \mathrm{Con}(T)$, giving us the impossibility $T \vdash \mathrm{Con}(T)$. In particular, this means that for any nice theory T we have that the theory $T + \mathrm{Con}(T)$, which one obtains by adding $\mathrm{Con}(T)$ as an axiom to T, has strictly higher consistency strength than T.

1.2 The linearity phenomenon

The relation $\leq$ defines an order on the collection of nice axiomatic theories, also referred to as the *hierarchy of consistency strength*. Interest in this hierarchy stems mostly from set theory. In the beginning of the 20th century, it became clear that all present-day mathematics could be carried out within the set theory ZFC, meaning that all mathematical statements can be formalized in its language and all mathematical theorems can be proven from its axioms. This was a remarkable result that gave hope that ZFC might be the *ultimate* mathematical theory in which all mathematical questions could be settled. As we know, in the 1930s Gödel's incompleteness theorems showed this hope to be false. Subsequently, one of the main tasks within set theory became to determine which questions *cannot* be settled within ZFC.

Assuming ZFC to be consistent, one way to show that a sentence ϕ cannot be proven in ZFC is to show that its negation is consistent with the theory. In other words, one needs to show the relative consistency result $\mathrm{Con}(\mathsf{ZFC}) \to \mathrm{Con}(\mathsf{ZFC} + \neg\phi)$, i.e. $\mathsf{ZFC} + \neg\phi \leq \mathsf{ZFC}$. A famous example of a sentence ϕ for which this has been done is the Continuum Hypothesis, denoted by CH. In 1937, Gödel [16] showed $\mathsf{ZFC} + \mathsf{CH} \leq \mathsf{ZFC}$ via his constructible universe L and in 1963 Cohen [9] managed to prove $\mathsf{ZFC} + \neg\mathsf{CH} \leq \mathsf{ZFC}$ via the method of forcing. These results imply that both $\mathsf{ZFC} + \mathsf{CH}$ and $\mathsf{ZFC} + \neg\mathsf{CH}$ are equiconsistent with ZFC, showing that CH is in fact independent of ZFC.

[4] For a defense of the claim that finitist reasoning is captured by PRA, see Tait [59].

Another way of showing that a statement ϕ is independent of ZFC is by showing that $\mathsf{ZFC}+\phi$ has strictly higher consistency strength than ZFC. As explained above, a sufficient condition for this would be $\mathsf{ZFC}+\phi \vdash \mathrm{Con}(\mathsf{ZFC})$. A prime example of statements whose independence of ZFC is shown in this manner are the *large cardinal hypotheses*, which are strengthenings of the Axiom of Infinity such as 'There exists an inaccessible cardinal', 'There exists a measurable cardinal' or 'There exists a Woodin cardinal'. Relative consistency results for large cardinals have been (and still are) a major research topic in set theory.[5]

In the research on large cardinals, a remarkable pattern has emerged concerning the consistency strength of extensions of ZFC. Using the method of forcing, many interesting extensions of ZFC have been shown to be consistent relative to some large cardinal hypothesis. In many of these cases, the extension can even be proven to be equiconsistent to some large cardinal hypothesis. As stressed by John Steel in [58], it appears that we currently know of no exceptions to this pattern, which he describes as follows:

Phenomenon 1.4. *The natural extensions of* ZFC *are all equiconsistent with* ZFC *or with some large cardinal extension of* ZFC*.*

By 'natural', Steel refers to those extensions that are "considered by set theorists, because they had some set-theoretic idea to them" (p. 157). Examples of such natural extensions would be $\mathsf{ZFC}+\mathsf{CH}$, $\mathsf{ZFC}+$'there are no Kurepa trees' or $\mathsf{ZFC}+$'there is a total extension of the Lebesgue measure'. This phenomenon then suggests that the large cardinal extensions form the backbone of the consistency hierarchy for natural set theories.

A second pattern that has emerged is that the large cardinal hypotheses are all comparable in terms of consistency strength. That is, among the large cardinal hypotheses that we currently know of there seems to be no exception to the following phenomenon:

Phenomenon 1.5. *For any two large cardinal extensions* T *and* S *of* ZFC*, we either have* $T \leq S$ *or* $S \leq T$*.*

This comparability of large cardinal hypotheses has come as a surprise to the set theoretic community, as these hypotheses have arisen from very different branches of set theory with diverse motivations.

The two phenomena stated above together imply that the natural extensions of ZFC are in fact linearly ordered by $\leq$. Interestingly, it seems that the linearity phenomenon is not limited to set theory; it also persists in theories of arithmetic, such as PA and its common subsystems and extensions such as PRA, $\mathsf{I\Sigma_n}$, $\mathsf{ACA_0}$, and $\mathsf{Z_2}$. This has led to the following informal conjecture, which I will call *the linearity conjecture*:

Conjecture 1.6. *All natural axiomatic theories are linearly ordered in terms of consistency strength.*

1.3 Instances of nonlinearity

It is important to note that in the linearity conjecture, the condition of naturality is a necessary one. Namely, it is well-known that the hierarchy of consistency strength for *all* nice axiomatic theories is *not* linearly ordered. It is possible to construct incomparable theories using self-referential sentences such as the Rosser sentence. Essentially, the Rosser sentence of a nice

[5] For a comprehensive account of large cardinals and relative consistency results, see Kanamori [30].

theory T is a sentence γ that is true if and only if for any proof of γ in T there is a smaller[6] proof of $\neg\gamma$ in T. The theorem given below is a generalization of theorem 3 in [22].

Theorem 1.7. *Let T be a nice theory and let γ be the Rosser sentence for the theory $T+\mathrm{Con}(T)$. Then the theories $T+\gamma$ and $T+\neg\gamma$ have incomparable consistency strength, i.e. we have $T+\gamma \not\leq T+\neg\gamma$ and $T+\neg\gamma \not\leq T+\gamma$.*

Note that this theorem implies that neither $T+\gamma$ nor $T+\neg\gamma$ is equiconsistent with T and so both must have a strictly higher consistency strength than T; γ is therefore called a *double jump sentence* for T.

We will provide a proof of this theorem in chapter 3 for the specific case that $T = \mathsf{PA}$; it should be easy to see that the proof also works for the general case. For now, the point is that we can find instances of incomparability at *any* level of the consistency hierarchy, be it at the level of basic arithmetic or at the level of very strong cardinal hypotheses.

In order for the linearity conjecture to carry any force, the theories described in Theorem 1.7 must be examples of *unnatural* theories. Indeed, there seems to be a consensus in the mathematical community that this is the case, as indicated by the following quotations:

"One can construct unnatural extensions (using self-referential sentences, for example) that are of incomparable consistency strengths." (Steel [58], p. 157)

"All known instances of non-linearity and ill-foundedness have been discovered by defining theories in an ad-hoc manner using self-reference and other logical tricks. [...] When one restricts one's attention to the natural axiomatic theories, [...] the resulting structure is a pre-well-ordering." (Walsh [65], p. 2)

"These [instances of a double jump] are all metamathematical examples, the kind of example that only a logician would construct." (Koellner [32])

"One can produce counterexamples by variants of Gödel sentences or of Rosser sentences. [...] Everybody agrees that these examples are not natural." (Caicedo [6])

"Of course it is possible to construct pairs of artificial theories which are incomparable under $\leq$. However, this is not the case for the 'natural' or non-artificial theories ..." (Simpson [54], p. 111)

"One can cook up ad hoc theories that are incomparable under consistency strength, but the natural ones are always comparable." (Montalbàn [44], p. 1211)

Note that the quotation by James Walsh indicates an even stronger version of the linearity conjecture, namely that the consistency hierarchy of natural theories forms a pre-well-order.

It is interesting that the judgement of the unnaturality of theories like $T+\gamma$ and $T+\neg\gamma$ is so pertinent in the literature, while 'being natural' is far from a precise property. As mentioned above, Steel describes the natural set theories as those "considered by set theorists, because they had some set-theoretic idea to them", and similarly Friedman, Rathjen and Weiermann [14] refer to the natural theories as those "which have something like an 'idea' to them" (p. 382). In [65], Walsh takes the natural theories to be the ones that "arise in practice" (p. 2), and something similar is stated by Peter Koellner in [32] who describes them as those "that arise in nature". These quotations suggest two key aspects of natural theories: they capture some genuine mathematical idea and they are studied by working mathematicians.

[6] Here 'smaller' means smaller in terms of the natural numbers that encode these proofs.

Do theories such as $T + \gamma$ fail to meet these criteria? Let us consider the case where $T = \mathsf{PA}$. On the one hand, the theory $\mathsf{PA} + \gamma$ does not seem to capture a structure that is of genuine mathematical interest, as the sentence γ does not seem to capture a genuine mathematical property of natural numbers. It is hard to think of a motivation for adding this sentence as an axiom to PA, except for wanting to construct a counterexample to the linearity conjecture. Moreover, the theory $\mathsf{PA} + \gamma$ has not received any consideration in mathematical practice, apart from popping up in theorems like the one above. On the other hand, the simple fact that $\mathsf{PA} + \gamma$ *has* popped up in this context does mean that it *has* been considered by mathematicians. Perhaps one would want to distinguish here between mathematicians and logicians, and argue that $\mathsf{PA} + \gamma$ is merely considered by the logician whose subject matter consists of all possible theories, but not by the mathematician whose subject matter consists only of those theories that have some genuine mathematical idea to them. Of course, such an approach only shifts the question: what constitutes a genuinely mathematical idea?

The notion of a natural theory is thus far from a precise, well-defined concept. It is therefore not clear how we are to assess the truth of the linearity conjecture or how to judge the known instances of incomparability. Apart from the well-known instance of nonlinearity provided by the Rosser sentence above, the only other instances of nonlinearity seem to have been provided by Joel Hamkins in a recent draft [22]. In this paper, Hamkins aims to provide counterexamples to the linearity conjecture by constructing instances of nonlinearity that, he argues, are natural. Yet again, without a precise definition of what a natural theory should be, it is unclear how one is to judge whether Hawkins has succeeded in providing actual counterexamples to the conjecture.

1.4 How to proceed?

Despite the vagueness of the notion of 'natural' and despite the known instances of nonlinearity, many mathematicians seem to believe that the linearity conjecture does carry some truth and is therefore worthy of further investigation. Friedman, Rathjen and Weiermann [14] even describe the situation as follows:

"The fact that 'natural' theories [...] are almost always linearly ordered with regard to logical strength has been called one of the great mysteries of the foundation of mathematics."[7] (p. 382)

The question then becomes: what can an investigation of the linearity conjecture consist of?

The most obvious kind of investigation one could request is the search for a *proof* of the conjecture. However, as it stands, it is not clear what a proof of the conjecture could look like. Walsh [65] describes the problem as follows:

"If it is true that natural axiomatic theories are pre-well-ordered by consistency strength, [...] then one would like to prove that it is true. However, the claim that natural axiomatic theories are pre-well-ordered by consistency strength is not a strictly mathematical claim. [...] Without a precise definition of 'natural', it is not clear how to prove this claim, or even how to state it mathematically." (p. 2)

This comment suggests that, before any *mathematical* work can be done on the conjecture, we need a precise definition of a 'natural theory'.

[7] The 'almost always' in this quotation could be interpreted as suggesting that there are *some* instances of incomparable natural theories. However, to our knowledge the authors have not provided or referred to any such particular instance in the literature.

Currently, there seem to be no accounts of an attempt at finding such a definition. What might come close to such an attempt is the work by Walsh and Montalbán found in [45]: their strategy is to mimic an approach to a similar phenomenon in recursion theory, namely the linearity phenomenon for natural Turing degrees.

Given two decision problems $A, B \subseteq \mathbb{N}$, A is *B-computable*, denoted by $A \leq_T B$, if A can be decided using an algorithm that may employ B as an oracle. The relation $\leq_T$ then defines an equivalence relation $\equiv_T$ on the subsets of $\mathbb{N}$, whose equivalence classes are called *Turing degrees*. Despite the fact that the full Turing hierarchy is quite complex, the 'natural' decision problems that computability theorists tend to come up with are quite well-behaved: they seem to form a well-order under $\leq_T$. An explanation of this linearity phenomenon is provided by a conjecture posed by Donald Martin in the 70s, which essentially states that each function $f : \mathcal{P}(\mathbb{N}) \to \mathcal{P}(\mathbb{N})$ that is invariant on Turing degrees, i.e. that satisfies $f(A) \equiv_T f(B)$ if $A \equiv_T B$, is equivalent to a constant function, the identity function or an iterate of the *Turing jump*, which maps any decision problem A to the Halting problem relativised to A. The connection of Martin's conjecture to the linearity phenomenon is that natural Turing degrees presumably induce such Turing invariant functions, and the conjecture then implies that these degrees are indeed well-ordered. The conjecture is still open, but many partial results are known.[8]

Inspired by this, Walsh and Montalbán aim to show that, just as the Turing jump is canonical for Turing invariant functions on subsets of natural numbers, the *consistency operator* of a theory T defined by $\phi \mapsto \mathrm{Con}(T + \phi)$ is canonical for *monotone* functions on sentences in T. The latter are defined as functions f on sentences in T that satisfy $T \vdash f(\phi) \to f(\psi)$ if $T \vdash \phi \to \psi$. Their work is still in progress, and future results will have to determine the success of their approach.[9] However, if they succeed, one could suggest to simply define the natural theories to be those that satisfy Walsh's and Montalbán's sufficient conditions for linearity.

Would such an approach provide us with a satisfactory demonstration of the linearity conjecture? In the case of the linearity phenomenon for Turing degrees, it seems that the extent to which Martin's conjecture provides a satisfactory explanation to this phenomenon crucially depends on the extent to which we are convinced that the natural Turing degrees indeed induce Turing invariant functions. The way in which these functions are induced is by relativisation: for a natural decision problem A, the function induced by A maps any decision problem B to the relativisation of A to B.[10] Relativisation is a key concept in computability theory, and there seems to be a consensus that natural constructions, which include proofs as well as decision problems, indeed relativise. However, in the context of axiomatic theories, it is not at all clear whether, and if so how, natural theories necessarily induce some monontone functions on sentences. Thus, in order to provide a satisfactory demonstration of the linearity conjecture, it seems that Walsh and Montablán's results would need to be supplemented by an argument showing that their formal conditions for linearity indeed somehow capture the intuitive notion of a natural theory.[11]

The point is here that, when finding a formal substitute for an informal notion, not anything goes. In the case of the linearity conjecture, it will not do to provide just any sufficient conditions for linearity; it also needs to be clear that these conditions indeed capture the intuitive notion

[8] See Montalbán [44] for clear review of Martin's conjecture and its connection of the linearity phenomenon for natural Turing degrees.

[9] In fact, in a very recent arXiv submission by Walsh [66], it is shown that some of their conjectures in [45] turn out false.

[10] For example, if A is the decision problem consisting of all (encodings of) programs that halt on a finite number of inputs, then the function induced by A maps any decision problem B to the set of (encodings of) programs with oracle B that halt on a finite number of inputs.

[11] Of course, I do not intend to suggest in any way that Walsh's and Montalbán's approach is uninformed or without merit. My point is merely that *any* approach to the linearity conjecture will ultimately have to address the question whether their proposed solution indeed adequately captures the intuitive notion of a natural theory.

of a natural theory. At the same time, with the intuitive notion being as vague and inchoate as it is, it is not clear what it means for formal conditions to capture this notion.

In the remainder of this thesis, I will carry out an attempt at moving towards a more precise definition of a natural theory. Of course, there are many ways in which one could go about this, and so we are faced with the task of determining a methodology. To this end, in the next chapter we will discuss different approaches to the general process of forming definitions in mathematics, which in the literature of philosophy of mathematical practice is often referred to as *mathematical concept-formation*.

Chapter 2

Concept-formation in mathematics

In this chapter, a philosophical framework will be set out for our attempt to make the concept of natural theory more precise, and a particular method for carrying out this attempt will be outlined.

Taking Carnap's notion of explication as a starting point, we will discuss three different roles that a precise definition of an intuitive concept can play, namely that of sharpening, analyzing or replacing the intuitive concept. Subsequently, we will review the case of Church's thesis, which can be viewed as a prime example of explication by means of conceptual analysis of an intuitive concept. We will consider whether this approach is suitable for the task at hand, and discuss the conceptual analysis of the concept of naturalness by San Mauro and Venturi.

We will then consider a drastically different approach to concept-formation, namely that found in Lakatos' *Proofs and Refutations*. In contrast to the former approach, the Lakatosian approach highlights the replacing role of the explicatum. I will argue that this approach seems fit for our purposes, and use it to outline the particular method that will be applied in the subsequent chapters.

2.1 On explication

In the discussion on concept-formation in mathematics, an often cited work is Rudolf Carnap's 'Logical Foundations of Probability', in which the first chapter is devoted to the general process of making inexact concepts precise. Carnap calls this process *explication*.

> "By the procedure of explication we mean the transformation of an inexact, prescientific concept, the *explicandum*, into a new exact concept, the *explicatum*." (Carnap [7] p. 3)

Carnap's aim is to find an explicatum for the informal concepts of *degree of confirmation*, *empirical induction* and *probability*, but his treatment of explication is meant to apply to any prescientific concept. The explicandum is an informal concept that is used by scientists at some developing stage of a scientific language; even though we do not have clear rules for its use, we can make clear how the explicandum is practically used by means of informal explanations and listing examples and non-examples. The explicatum, on the other hand, must be an exact concept in the sense that its use is governed by explicit rules that connect it to a well-established

system of scientific concepts. Carnap considers mathematics as a part of science: in the context of mathematics, the explicatum must be incorporated into a well-established system of *logico-mathematical* concepts.

Carnap stresses that the process of explication is tricky, as we do not have an exact way of deciding whether a given explicatum is correct. Intuitively, we would want the explicatum to be coextensive with the explicandum. However, as the explicandum is *vague* in the sense that we lack a decision procedure for its application in every case, it is not clear how to make sense of this request. Carnap therefore proposes to judge an explicatum by its similarity to the explicandum, its fruitfulness and its simplicity. The latter is taken to play a lesser role; when faced with two concepts that seem equal in both similarity and fruitfulness, preference will be given to the simplest one. The explicatum is to be similar to the explicandum in the sense that in most of the cases where we know the explicandum to apply or not to apply, the explicatum must do the same; however, in cases where the application of the explicandum is indeterminate, the explicatum is free to be applicable or not. There is an emphasis here on *most of* the cases: an explicatum's (dis)application may disagree with that of the explicandum if this loss of similarity is compensated by a gain in fruitfulness. As an example, Carnap considers the explication of the term 'fish': animals such as whales and seals that were initially included in the prescientific concept of fish, generally understood to apply to 'animals living in water', are now excluded by the scientific concept of fish as known from zoölogy. The reason that this particular explicatum was favored over concepts more similar to the explicandum was its fruitfulness: the animals to which the scientific concept applies share more particular properties than those falling under the prescientific one, and thus the former allows us to formulate more general empirical statements. According to Carnap, the fruitfulness of an explicatum thus lies in our ability to formulate universal laws involving this concept. In the context of mathematics, an explicatum is fruitful if it yields connections with other mathematical theories and enables us to formulate general theorems.

The process of explication is abundant in mathematics. Consider for example the explication of continuity by the epsilon-delta definition, the explication of a function as a set of ordered pairs, or the explication of validity as truth in every model. In each of these cases, there is a strong conviction that the explicatum is a satisfactory one, as we believe the explicatum to be similar to the explicandum in the sense just described. Notice that this conviction is not without merit: any result involving the precise explicata just listed is interesting to the mathematician *just because* we believe this explicata to be similar to the original, informal explicandum. After all, it was the informal notion that we were interested in the first place.

While Carnap provides us with criteria to judge whether a given explicatum is satisfactory – in particular, whether it is more or less satisfactory than some alternative – no indication is given as to how scientists are to arrive at a certain explicatum. A little hint might be found in the first line of the following quote:

> "The explicatum (in my sense) is in many cases the result of an analysis of the explicandum [...] in other cases, however, it deviates deliberately from the explicandum but still takes its place in some way" (Carnap [7], p. 3)

The first line seems to indicate that an explicatum can be found by careful analysis of the explicandum; if so, the explicatum must in some sense already be implicit to the intuitive explicandum. However, the second line suggests that, irrespective of whether an explicatum is implicit to the explicandum, it can be preferable to replace the explicandum by an explicatum that differs from it in significant ways.

It will be clarifying to link our discussion here to a distinction between the *sharpening*, *analyzing* and *replacing* of inexact concepts as made by Luca Incurvati in [26]. Incurvati's book focuses on *conceptions* of the mathematical notion of set, which are defined as follows:

> "A *conception* of C, where C is a concept, is a (possibly partial) answer to the question 'What is it to be something falling under C?' which someone could agree or disagree with without being reasonably deemed not to possess C." (Incurvati [26], p. 13)

The idea is here that some features of a concept C are so central to our understanding of it that anyone who fails to recognize those features would be deemed not the posses the concept C; in the case of the concept of set, these features include that of being a single, unified entity that is completely determined by its members. These features, however, might not provide a full answer to the question 'What is it to be something falling under C?'. A particular conception of C then provides a more elaborate answer to this question, which respects the central features of C.

Incurvati distinguishes three roles that a conception of a concept C can play. First, a conception can be viewed as a *sharpening* of the concept. The idea is here that the criterion of (dis)application for the concept C fails to be determinate in every possible case, and a particular conception of C then settles (some of) these indeterminate cases. Crucially, how these indeterminate cases are settled does not depend on the concept C is any way; one might view the settling of these cases as arbitrary, at least with regard to the concept C.[1]

In contrast to a sharpening, a conception might be seen as providing an *analysis* of the concept C by spelling out features of C that were already implied by that concept, even though we might have failed to identify them. Gödel seems to have held such a view on conceptions:

> "The precise concept meant by the intuitive idea of velocity clearly is ds/dt, and the precise concept meant by "size" ... clearly is equivalent with Peano measure in the cases where either concept is applicable. In these cases the solutions again are unquestionably unique ... they satisfy certain axioms which, on closer inspection, we find to be undeniably implied in the concept we had." (Gödel, quoted in [67], p. 233)

The third role that a particular conception can be thought to play is that of *replacing* the corresponding concept. In this interpretation, a conception might simply be seen as a new concept that is favored over the initial one due to some preferable features. As an example, Incurvati considers the case where the naive concept of set as the extension of a predicate, which famously leads to a contradiction via Russell's paradox, is replaced by a consistent one.

Incurvati's discussion on concepts and their conceptions naturally translates to Carnap's discussion on explicanda and their explicata. In particular, note that Incurvati's three distinct roles of conceptions can all be found in Carnap's description of explication. The criterion of similarity, taken together with the demand that explicata ought to be *exact* concepts, indicates the sharpening role of the explicatum. In contrast, the criterion of fruitfulness and its ability to justify a compromise on similarity, suggests a replacing role.

The analyzing role of the explicatum is hinted at in the quotation of Carnap given above. One could wonder whether, in cases where the explicatum is obtained from an analysis of the explicandum, it makes sense to speak of the *unique* or *correct* explicatum that was somehow implied by the explicandum. Carnap certainly did not think this, as he explicitly states that explication is an inexact science for which "we cannot decide in an exact way whether it is

[1] Note that the settling of the indeterminate cases need not be arbitrary *in general*; the choice of settling these cases in a certain way can be well motivated. The point is here that these motivations do not appeal to the initial concept C.

right or wrong" (p. 4). However, there is an interesting case of an informal mathematical concept for which it has been claimed that its exact definition is indeed the uniquely correct one, and even *provably* so. The concept in question is that of *effective computability*, and the statement that links this concept to the formal definition of recursive computability is known as Church's thesis. Church's thesis is often viewed as an example of explication[2] and has received considerable attention from philosophers and mathematicians alike, in particular with regard to its provability. In particular, it can be viewed as a case of explication in which the explicatum is found via conceptual analysis of the explicandum.

Since the literature on the status of Church's thesis is so extensive, and since it shares with the linearity conjecture the feature of being an informal claim involving precise, mathematical notions that is nevertheless viewed as *true*, it serves as a useful case study in our attempt to find a method for explicating the intuitive notion of a natural theory. In the next section, we will therefore discuss this case in more detail.

2.2 Explication via conceptual analysis: the case of Church's thesis

Church's thesis states that the class of effectively computable functions is exactly the class of recursive computable functions. Here an *effective computable function* should be understood as a function from natural numbers to natural numbers that can be calculated by an algorithm in the informal sense of the word, that is, as a step-by-step procedure that halts in a finite number of steps and could in principle be carried out by a human computer. This informal notion has been around for centuries[3] and the need for a more precise definition of effective computability arose only in the beginning of the 20th century, when mathematicians became interested in proving that certain functions were *not* computable. In the 1930s, Church, Kleene, Gödel and Turing independently developed explicata for effective computability, resulting in the formal definitions of λ-computability, recursive computability and Turing computability. Turing's treatment is often considered the most convincing, as his notion of a Turing machine derives from a conceptual analysis of the informal notion of an algorithm as a finite, step-by-step calculation. However, the three formal notions were readily shown to be equivalent, and so the thesis can equivalently be stated in terms of Turing computability or λ-computability instead of recursiveness.

There is a strong consensus in the literature that the thesis is accepted as true by the mathematical community. In computability theory, it is not unusual to invoke the thesis in proofs, as to avoid the often tedious task of formally proving that a function is computable. Also note that it is Church's thesis that puts the force behind many results on recursive functions. To stay close to the topic at hand, note that the impact of Gödels incompleteness theorems is driven by our conviction that Church's thesis is true; only because we believe that effective computability is captured by recursiveness, do we feel that Gödel has shown that there is no hope of finding a complete theory that can be axiomatized in an effective manner. In fact, Gödel himself had reservations about the impact of his incompleteness results until Turing's analysis convinced him that recursiveness formed the "absolute definition" of computability (Gödel [18], p. 8).

[2]See e.g. Black [5] and Mendelson [43]. For a more critical discussion of whether the thesis can be viewed as such, see Quinon [48].

[3]Dating back to Euclid's algorithm for the greatest common divisor.

2.2.1 Proving the thesis?

Given the fact that Alonzo Church [8] initially introduced the notion of recursiveness as a definition of effective computability, it is interesting that Church's thesis is called a *thesis*. Janet Folina [13] writes:

> "... [Church's thesis] is not considered to be a mere definition, for it has substantial content which seems capable of being true or false." (p. 302)

In standard textbooks covering basic computability theory, one usually finds two arguments in favor of the thesis: (1) of the large number of effective procedures known to humankind, none has turned out to be non-recursive, and (2) all exact notions[4] proposed as explicatum for effective computability have turned out to be equivalent. These arguments are not meant to demonstrate the truth of the thesis, but as evidence that makes the thesis plausible.

The standard conception of Church's thesis seems to be that we cannot do much better than this. Kleene [31] writes:

> "While we cannot prove Church's thesis, since its role is to delimit precisely an hitherto vaguely conceived totality, we require evidence that it cannot conflict with the intuitive notion which it is supposed to complete." (p. 319)

However, more recent accounts of the thesis such as that of Smith [56], Black [5], Shapiro [52], Mendelson [43], Gandy [15] and Sieg [53], have argued that we *can* do better than this: each argues that the thesis is provable, or stronger still, in the case of Gandy, Sieg and Smith, that it already has been proven. Most of their arguments are similar, as their main task is to establish that (1) the alleged 'vagueness' of the notion of effective computability does not block the provability of the thesis and (2) there is a particular proof technique available for the thesis. Instead of considering each account individually, I will therefore set out the main arguments offered in favor of these two claims.

The first argument addresses the vagueness of the notion of effective computability. This argument attacks the claim, which seems to be implied in the quotation of Kleene, that two concepts cannot be demonstrably coextensive if one is vague while the other is precise. Black's, Smith's and Shapiro's response is that even though the concept of effective computability might be vague in *sense*, it need not be vague in extension. To illustrate, the concept of a tall person is vague in extension, as there clearly are borderline cases of which it is unclear whether the concept applies.[5] However, there simply do not seem to be any borderline cases of effective computability in this sense; for any alleged algorithm, the intuitive notion seems to suffice to judge whether it describes a genuine algorithm, i.e. a genuinely effective procedure.

The second argument attacks the claim that a genuine mathematical proof cannot involve informal notions. Mendelson writes:

> "The assumption that a proof connecting intuitive and precise mathematical notions is impossible is patently false. In fact, half of CT (the 'easier half') ... is acknowledged to be obvious to all textbooks in recursion theory." p. 249

[4] Apart from the three notions mentioned above, there are many more, such as Kolmogorov-Uspenskii computability or computability in terms of register machines.

[5] A controversial view on this point is provided by William [68], who argues that terms such as 'red' can have a fixed extension despite the fact that ordinary speakers might feel that there exist borderline cases of which they cannot be certain whether the concept applies to it or not.

By the easier half of the thesis, Mendelson refers to the implication that each recursive function is effectively computable. The recursive functions can be defined as the least class of numerical functions that contains the so-called initial functions (consisting of the zero, successor and projection functions) and is closed under the operations of composition, primitive recursion and μ-recursion. Mendelson argues that we can use a standard induction argument to show that these functions are in fact effectively computable: we can construct effective algorithms to compute the initial functions, and for each of the three operations we can, given some effective computable functions, describe an algorithm that would compute the function that results from applying the operation to those computable functions. Mendelson claims that this "is as clear a proof as I have seen in mathematics, and it is a proof in spite of the fact that it involves the intuitive notion of effective computability" (p. 250). Smith gives a different 'proof' of the easy implication: he employs the equivalence of recursiveness and Turing computability, and subsequently invokes the argument that each Turing-program is effectively computable, as its instructions simply describe an effective algorithm. He also provides another instance of a proof in which the informal concept of effective computability could occur: if someone manages to find an example of an intuitively computable function that is provably not recursive, then this would serve as a *disproof* of the thesis.

What constitutes a mathematical proof is not an undisputed matter. As explained by Avigad [4], the standard conception of a mathematical proof seems to be that of a mathematical argument that could, by some "logically adept and sufficiently motivated mathematicians", be reduced to a formal derivation in some axiomatic system accepted by the mathematical community (p. 7379). Clearly, the notion of a proof employed by Mendelson and Smithis not that of a formalizable proof.

For our purposes, we need not take a position on whether Mendelson's or Smith's arguments deserve the label 'proof'. What is important for us is the observation that the arguments provided by Mendelson and Smith of the easy implication are not mere plausibility arguments; they are genuinely mathematical arguments that are meant to demonstrate the *truth* of this implication, not its mere plausibility. Smith makes the following useful distinction:

> "We might distinguish, then, three levels of mathematical argument - mere plausibility considerations, informal demonstrations, and ideally formalized proofs (or truncated versions thereof)." (Smith [56], p. 353)

We leave open the matter of whether the label 'proof' is to apply to the second notion.

The two arguments discussed above are meant to show that the vagueness or informality of 'effective computability' does not necessarily deem the thesis unprovable. This, of course, does not yet indicate how we are to prove the thesis. In particular, we need a proof strategy for the so-called *hard implication*, i.e. the statement that each effectively computable function is indeed recursive.

Shapiro's suggestion for what a proof of the hard implication could look like, derives from a comment by Gödel, in which he proposes that it might be possible to "state a set of axioms which would embody the generally accepted properties of this notion, and to do something on that basis" ([52], p. 286). Shapiro suggests that this set of axioms could be obtained from a conceptual analysis of the notion of effective computability, and the 'something on that basis' would then consists of a formal proof that these axioms imply the formal concept of recursiveness. The difficult part, of course, will be to establish whether the axioms derived from the conceptual analysis will in fact be evidently true properties of any effective algorithm. One such conceptual analysis is famously provided by Turing [61]. Turing starts from the conception of an algorithm as a step-by-step procedure that can be carried out in finitely many steps by an idealized, human

calculator. This idealization involves that we leave out at any constraints on time, space or resources; the procedure must only be capable of being carried out *in principle*. Turing then identifies conditions that any such algorithm should satisfy, such as requiring a finite alphabet and allowing only limited movement within the workspace at each step. From these conditions, Turing eventually arrives at his definition of a Turing program.

If one accepts Turing's analysis, the demonstration of the hard implication would be finished by the formal proof that Turing computability implies recursiveness. Gandy, Sieg and Smith argue that Turing's analysis, or at least some more elaborate version of it, is indeed a satisfactory conceptual analysis that provides necessary conditions for computability.[6]

Taking these proof suggestions together, the full demonstration of Church's thesis would take the form of a *squeezing argument*. The idea is due to Kreisel [35], who used it so show that the intuitive concept of first-order validity, understood as 'truth in virtue of form', is coextensive with the model-theoretic notion of validity. The idea of a squeezing argument is relatively simple. Let I be some informal concept, and for any concept X let $|X|$ denote its extension. Now suppose we could find some formal concepts S and N that provide sufficient and necessary conditions for the concept I, i.e. suppose that we could (informally) demonstrate that

$$|S| \subseteq |I| \subseteq |N|.$$

Differently put, the extension of I is 'squeezed' inbetween that of S and N. Now suppose in addition, we could (formally) demonstrate that

$$|N| \subseteq |S|.$$

Then we could conclude $|S| = |I| = |N|$, thereby obtaining our co-extension result. In the demonstration of Church's thesis outlined above, S would be instantiated by recursiveness, whereas N would be instantiated by the conditions or axioms obtained from a conceptual analysis of the notion of effective computability.

There is an important remark of caution concerning squeezing arguments that has been raised by Smith in [55]. Smith stresses that the squeezing arguments for effective computability and logical validity only 'kick in' after the intuitive notion that is to be squeezed has already undergone some conceptual sharpening. To illustrate, the intuitive notion of an algorithm as an effective procedure that can be carried out in principle could still be sharpened in many distinct ways, as the following questions ought to make clear: carried out by who, us or machines? If by us, then as we actually are or as some idealized version of us? Does feasibility play a role? Many years of working with the notion of effective computability led mathematicians to sharpen the idea to that of a finite, step-by-step symbolic computation that could be carried out by an idealized human computer, ignoring constraints of time and space (but not of working memory). And it is *this* notion, still informal yet already sharpened to a considerable degree, that the squeezing argument applies to. To make the point clear, Smith distinguishes three levels of concept. At the *pretheoretic level*, we start with some inchoate ideas of computability, mainly by referring to examples of "common-or-garden real-world computation" (p. 29). At the *prototheoretic level*, we further develop these ideas into one particular direction, resulting in the sharpened idea of effective computability as described above. Then there is the fully *theoretic level*, where we find the precise notions of recursiveness and Turing compatibility and the like. The squeezing argument only applies to the prototheoretic notion of computability, as the pretheoretic notion

[6] Sieg and Smith both feel that Turing's treatment leaves some gaps; Sieg claims to have filled these gaps by his own conceptual analysis of computability, whereas Smith argues that Kolmogorov's and Uspenskii's treatment of computability as given in [33] does the trick.

is simply too unrefined to have a precise extension; when moving from the pretheoretic to the prototheoretic concept of computability, different routes could have been taken. Smith writes:

> "... it would be plainly over-ambitious to claim that in refining our inchoate ideas and homing in on the idea of effective computability, we are just explaining what we were talking about all along." (Smith [55], p. 29)

The force of the squeezing argument is thus in showing that, even though we have to sharpen our intuitive notion to some degree, we might not need to sharpen it completely before the extension of this notion takes on a precise form.

Connecting this to our earlier discussion on the possible roles of explication, it seems that the concept of computability has undergone two phases of explication. The first is a sharpening phase, where the number of indeterminate cases for application of the pretheoretic concept was reduced. Following Smith, the particular sharpening that occurred was *arbitrary* in the sense that it was not implied by the initial concept. This is not to say, of course, that this sharpening was arbitrary in the sense that is was uninformed; considerations of interest and fruitfulness will certainly have played a role in the process. The second phase is an analyzing phase, where the fully theoretical concept is obtained from the prototheoretical one by a careful conceptual analysis of the latter. Church thesis can then be interpreted as stating that the outcome of this analysis was *correct*, in the sense that the prototheoretic and the theoretic concept are exactly coextensive.

2.3 A conceptual analysis of naturalness?

The case of Church's thesis makes it clear that the process of explication is a fundamental part of mathematics and not a mere prelude to the actual mathematical work. In particular, it shows that explication is not equivalent to stating a definition; a definition cannot be true or false, whereas Church's thesis suggests that an explication *can* have a truth value. In the case of Church's thesis, the explication is generally thought to be a true one; this conviction is important, as it puts the force behind many important theorems involving recursive functions. It seems safe to say that this point generalizes to all instances of mathematical explication: our conviction that an explicatum is correct is tantamount to our interest in any formal results involving the explicatum. Moreover, the analysis above shows that, when arguing for the correctness of a given explication, we have more than mere plausibility arguments at our disposal. Irrespective of whether one agrees with Gandy, Sieg and Smith that a satisfactory squeezing argument has been provided by (some version of) Turing's analysis, it is at least clear that such an argument is possible in principle and that the reasoning employed in such an argument is of a genuinely mathematical nature, possibly even deserving of the label 'proof'. Lastly, the case of Church's thesis suggests that a conceptual analysis of an intuitive notion can provide us with precise conditions that are evidently true for any instance of this notion, provided that this informal notion has already been sharpened to some degree.

Can the intuitive concept of natural theory be subjected to such a conceptual analysis? There is one significant way in which the notion of naturalness differs from that of effective computability: whereas the concept of effective computability has been sharpened to a degree that there appears to be no case of an algorithm of which it is unclear whether it does or does not classify as a genuinely effective computation, the application of the concept of natural theory is far from this precise. Using Smith's terminology, the concept of natural theory should be placed at the pretheoretic level: we have some inchoate ideas about what a natural theory ought to be, namely one that 'has a genuine mathematical idea to it' or 'arises in mathematical practice', and

we have some examples of natural theories, e.g. ZFC and PA and their common subsystems and extensions, and some non-examples, namely those theories involving the Rosser sentence. This suggests that the notion of a natural theory needs to undergo some sharpening before it can be subjected to a conceptual analysis such as that provided by Turing.

We should mention, however, that something like a conceptual analysis of the general notion of naturalness in mathematics has been carried out by Luca San Mauro and Giorgio Venturi [50]. The aim of their study is to provide an account of the informal use of this notion, that is, of the use of this notion outside of formal definitions.[7] In particular, they aim to determine whether naturalness is a static or a dynamic property of mathematical objects or constructions. Here a *static* property is to be understood as one that is stable over time and inherent to the object that we assign this property to, as opposed to a *dynamic* property that is contextual in the sense that it pertains to the relations that the object in question as to other objects.

By considering case studies from set theory and computability theory, San Mauro and Venturi find the use of naturalness to be governed by both contextual and normative considerations: a mathematical concept or construction is generally considered to be natural if it has stable connections to other mathematical contexts and, in addition, if the concept or construction in question is one that is exemplary of how we want to do mathematics. Their conclusion is the following:

> "In conclusion, we suggest that naturalness should be considered as a device of self-regulation within mathematical practice, a device that through a dynamic and communitarian process informs us of the ways in which we want this practice to be performed." (San Mauro and Venturi [50], p. 310)

San Mauro and Venturi therefore conclude that 'being natural' is not an inherent property of certain mathematical objects that remains stable over time.

This classification of naturalness as a dynamic rather than a static notion seems problematic for anyone wanting to demonstrate the truth or falsity of the linearity conjecture. If what classifies as a natural theory changes over time, then the linearity conjecture becomes nothing more than a description of the status quo: it would come to mean that the consistency strength of theories that hold a certain position in the *current* mathematical climate forms a linear order. If this is indeed how things are, then any kind of demonstration of the conjecture seems impossible, as the conjecture would link a dynamic concept, namely that of a natural theory, to the static concept of linearity. After all, we cannot predict how mathematical practice is to develop, and thus such a demonstration could not appeal to any intrinsic property of natural theories, since these might change over time. Instead of calling it a *conjecture*, we then better refer to it as a mere *observation* describing our current mathematical practices.

The worry that the concept of natural theory is merely contextual is also expressed by Hamkins in his discussion of the linearity conjecture:

> "We have no coherent robust concept of what counts as natural, and empty naturality talk is too often used merely to reject the unfamiliar. For someone to declare a construction or idea 'unnatural' is often little different from them saying, 'I don't like it' or 'it uses unexpected ideas'. " (Hamkins [22], p. 33)

Hamkins here seems to imply that, in addition to being a contextual notion that is connected to familiarity, the use of the notion of naturalness is often uninformed or even *empty*.

I would like to push back against Hamkins claim that 'naturality talk' is empty. As the work of San Mauro and Venturi shows, the notion of naturalness serves a purpose in mathematical

[7] An example of such a definition would be that of a natural transformation in category theory.

practice, and so referring to a mathematical construction as 'natural' is not a meaningless undertaking. Moreover, if naturality talk in mathematics was indeed empty, then a large number of discussions in mathematics would have to be deemed meaningless. To name a few, apart from the discussion at hand, consider the discussions on the naturalness of axioms of ZFC, the naturalness of ordinal notations in ordinal analysis, or the naturalness of decision problems in computability theory. Even in the case of some formal definitions that involve the notion of naturalness, such as 'natural number' or 'natural transformation', the reference to naturalness seems to be more than a mere syntactic label.

Apart from this, I admit that San Mauro and Venturi as well as Hamkins may be right that the notion of naturalness in mathematics *in general* is governed by contextual considerations. However, this does not necessarily imply that the particular concept of natural theory, as referred to in the context of the linearity phenomenon, does not cut any ice. To illustrate this, let us reconsider the linearity phenomenon for Turing degrees described in the former chapter. As Montalbán exemplifies, there is a conviction that the linear behavior of the natural Turing degrees is not a mere description of the status quo:

> "The contrast between the general behavior in [the Turing hierarchy] and the behavior of the naturally occurring objects is so stark that there must be a deep reason behind it."
> (Montalbán [44], p. 1211)

And indeed, Martin's conjecture suggests that there is such a reason. So the fact that the general notion of naturalness is a dynamic one did not prevent computability theorists from isolating a static, robust property of the natural Turing degrees. Building on this case, it seems that to assume that the concept of natural theory is merely contextual would be to prematurely dismiss the intuition of the mathematical community, as the fact that theories involving the Rosser sentence are widely considered to be unnatural serves as an indicator that these theories must satisfy some robust property explaining this.

Taking this intuition of the mathematical community seriously, the task at hand is now to sharpen the pretheoretic, inchoate concept of natural theory into a prototheoretic one that is of a static rather than a dynamic nature. The difficulty, of course, is that any such particular sharpening is not induced by the pretheoretic concept itself, and thus a wide range of prototheoretic concepts are possible. What we need, then, is a method that enables us to obtain a particular sharpening of the pretheoretic notion. Interestingly, such a method seems to be provided by Imre Lakatos, whose view on mathematical concept-formation will be discussed in the next section.

2.4 The Lakatosian approach to concept-formation

Thus far, our discussion of mathematical explication has treated the formation mathematical concepts and definitions as isolated from the formation of mathematical theorems. A drastically different view on concept-formation in mathematics can be found in Imre Lakatos' *Proofs and Refutations* (PR) [38]. This work is often described as a revolutionary piece in the philosophy of mathematics, which shifted away from the traditional focus on ontological questions concerning the nature of mathematical objects and how one might access them. Instead, Lakatos calls for a historical account of mathematics that focuses on the development and progress of real-life mathematical practice.

2.4.1 Lakatos' *Proofs and Refutations*

In PR, Lakatos aims to show that concept-formation in mathematical practice occurs through a dynamic interplay between conjectures, proofs and counterexamples. He does so by means of a case-study on Euler's formula for polyhedra, through which he both discovers and explains his *method of proofs and refutations* as the general pattern of mathematical discovery. His work takes the form of a fictitious dialogue between a group of students and their teacher; their discussion of Euler's formula reflects the *real* history of the discovery and the proving of the formula, which is pointed out to the reader via the footnotes.

Lakatos introduces his work as a challenge for mathematical formalism with

> the modest aim to elaborate the point that informal, quasi-empirical mathematics does not grow through a monotonous increase of the number of indubitably established theorems but through the incessant improvement of guesses by speculations and criticism". (p. 5)

Formalism is to be understood as "the school of mathematical philosophy which tends to identify mathematics with its formal axiomatic abstraction" (p.1). The idea is here that, for the formalist, mathematical theories are to be identified with an axiomatic system, their theorems with well-formed formulas and their proofs with sequences of such formulas that are governed by a set of fixed inference rules. Moreover, definitions are nothing more than abbreviations for interesting formulas; they enable us to write theorems in a succinct and clear way.[8]

According to Lakatos, formalism fails to account for the process of mathematical discovery that is crucial for mathematical growth. While formalization serves as a very powerful tool to obtain rigour in our mathematical results, a formalised theory leaves very little room for discovery: one can either solve questions that could just as well be solved by a properly programmed machine, or, in case the theory is undecidable, one can try to guess which sentences are theorems. However, Lakatos claims that discovery in informal mathematics is neither mechanical nor irrational; instead, it is a rational process that generally follows his method of proofs and refutations.

The method of proofs and refutations describes the dynamic interplay between theorems, counterexamples and proofs during the process of mathematical discovery. According to the method, this interplay follows a general pattern.

First, a preliminary, informal conjecture is formulated. In PR, the conjecture considered reads "All polyhedra satisfy $V - E + F = 2$", where V, E and F respectively represent the number of vertices, edges and faces of the polyhedron. At the time of Euler's formulation of the conjecture, 1758, a polyhedron was generally understood to be 'any solid bounded by planes or plane faces'.

Subsequently, a proof of the conjecture is suggested. It is important to note here that a 'proof' should not be understood in the formal sense; after all, formal proofs can only deal with formal statements and not with informal conjectures. The proof considered in PR is that proposed by Cauchy in 1813, which employs informal notions such as 'removing surfaces' and 'stretching the polyhedron without tearing it'. For Lakatos, such a proof should be understood

[8] The student of mathematics will most likely recognize the dominance of the formalist view in mathematical literature: textbooks and papers in mathematics follow a deductivist style, in which one is presented a list of formal definitions, followed by a theorem and its proof. The definitions can be involved and their precise formulation often seems *ad hoc* until one reads the proof of the theorem; it is only there that one realises that the definitions were chosen just so to make the proof work. This order of presentation is unlikely to reflect the order of discovery; in general, the definitions are inspired by the proof and the proof is inspired by some preliminary conjecture. In the formalist view, however, there is no room for this dynamic between proofs, definitions and theorems that led to the final result. In *Appendix 2* of PR, Lakatos calls for a revision of this deductivist style in mathematical literature.

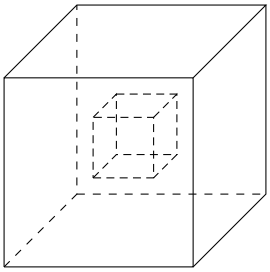

Figure 2.1: A nested cube, that is, a cube with a cube-shaped hole.

as a "thought-experiment [...] which suggests a decomposition of the original conjecture into subconjectures or lemmas" (p. 10). The main purpose of such a proof is not to show that the preliminary conjecture is indubitably *true*. Instead, its main purpose is to reveal more places for criticism on the conjecture, that is, for counterexamples.

In reaction to both the conjecture and its proof, counterexamples may emerge. In the case of Euler's conjecture, these included the nested cube (see Figure 2.1), the cylinder and the starpolyhedra. Lakatos identifies three different ways in which mathematicians responded to these counterexamples. The first response is to simply deny that the counterexample forms a proper counterexample, that is, to reject the proposed counterexample on the basis of not classifying as a proper polyhedron. Lakatos refers to this as the method of *monster-barring*. In the case of Euler's conjecture, monster-barring resulted in a sequence of increasingly restrictive definitions of 'polyhedron', each of which meant to keep out yet another monster.

The second response is referred to as *exception-barring*. The exception-barrer takes counterexamples more seriously than the monster-barrer: she accepts the counterexample as such and contracts the domain of validity of the conjecture as to ensure that it does not contain the counterexample. To illustrate, when faced with the counterexample of a nested cube to Euler's formula, she will add a condition to the conjecture such as 'All polyhedra *without cavities* satisfy $V - E + F = 2$'.

A worry for the exception-barrer is that, when contracting the domain of the conjecture, she might unintentionally be excluding some Eulerian polyhedra[9] as well; perhaps it is not the cavity per se, but some other property of the nested cube that makes it fail to be Eulerian. To avoid this problem, the third method does not only take the counterexample seriously, but in addition takes the proof suggestion into account. It consists of a careful analysis of the proof, with the aim to find the 'guilty lemma' that is refuted by the counterexample. This might be a difficult task, as the guilty lemma need not have been stated explicitly in the proof; in that case, one first needs to give an elaboration of the proof. Once the lemma is found, it can be built into the conjecture as a condition. Lakatos refers to this as *the method of lemma-incorporation*. To illustrate, it turns out that the nested cube does not satisfy an implicit assumption made in Cauchy's proof, namely that each polyhedron can be stretched onto a plane when one face is removed. Thus, as a response to this counterexample, the lemma-incorporator will make this assumption explicit in the conjecture.

Note that the method of lemma-incorporation leads to an improvement of both the conjecture and the proof. As argued by Lakatos, this method forms the most fruitful response to a counterexample, that is, the response that facilitates the largest increase in mathematical knowledge, as it makes hidden assumptions explicit and leads to the most general version of the conjecture.

[9] That is, polydra satisfing Euler's formula.

After the conjecture has been improved, new counterexamples may emerge, and so the method of proofs and refutations continues. In the eyes of Lakatos, this is precisely what improved conjectures and proofs are meant to do: by making hidden assumption explicit, they open up new places for criticism. Eventually, the emergence of counterexamples may come to a halt, and the conjecture and its proof might be translated into a formal theory. Such a formalization may convince us that our proof was valid, and that no subsequent *formalizable* counterexamples will be found. However, Lakatos stresses, the formal translation is no substitute for the informal theory; this point is made explicitly in [37], where Lakatos writes that

> "... we have no guarantee at all that our formal system contains the full empirical or quasi-empirical stuff in which we are really interested and with which we dealt in the informal theory." (p. 67)

2.4.2 Concept-formation in PR

In PR, Lakatos explicitly connects his method of proofs and refutations to concept-formation. Note that the methods of monster-barring, exception-barring and lemma-incorporation can all be viewed as doing some conceptual *contracting*. The monster-barrer can be viewed as contracting the domain of application of the concept 'polyhedron', whereas the exception-barrer and the lemma-incorporator can be seen as contracting the domain of application of the conjecture. However, Lakatos suggests that this view fails to take into account that, as our knowledge grows, the semantics of our language *changes*.

Instead of looking at the monster-barrer as contracting the concept of polyhedron, one could view the refutionist as *stretching* the original concept, by suggesting an example of a polyhedron to which the original concept was never meant to apply. Through the voice of student Pi, Lakatos[10] remarks:

> "The conjecture was true in its *intended interpretation*, it was only false in an *unintended interpretation* smuggled in by the refutationists. Their 'refutation' revealed no *error* in the original conjecture, no *mistake* in the original proof: it revealed the falsehood of a *new* conjecture which nobody had stated or thought of before." (p. 90).

In this light, the monster-barrer does not contract the concept of polyhedron, but simply keeps it fixed.[11]

For Lakatos, then, the role of refutations is *heuristic*: they do not show that the conjecture is false, but they show that more is going on than we initially imagined. By simply dismissing a counterexample, as the monster-barrer does, one refuses to change their interpretation of the problem at hand and by doing so, fails to facilitate any epistemic growth. In contrast, by accepting the counterexample as such, one agrees to change their language by means of concept-stretching. Subsequently, one has to improve the conjecture by means of a new concept. When using the method lemma-incorporation, this new concept will be induced by the guilty lemma of the proof; Lakatos refers to this as the *proof-generated concept*. To come back to our example

[10] As PR is presented in the form of a dialectic between students with opposing views, it is sometimes difficult to make out which claims are to be interpreted as Lakatos' own. I am assuming that in chapter 8 of PR, the student Pi represents Lakatos' own view on concept-formation.

[11] This view of monster-barring might help to understand how many respected mathematicians seem to have adhered to this method, which on the face of it might seem like nothing more than a cheap, linguistic trick. In the light of the counterexamples, it seems to believe that Cauchy could have made the 'mistake' of thinking that he had proven the conjecture for *all* polyhedra. However, from Lakatos's perspective, Cauchy did not make a mistake at all.

of the nested cube, the proof-generated concept is that of a solid bounded by planes that, after one face has been removed, can be stretched onto a plane.

Connecting Lakatos' view on concept-formation to our discussion on explication, one can view Lakatos' proof-generated concept as an explicatum whose main role is that of *replacing* the initial concept. This is made particularly clear in the following quote of the student Pi:

> "*Proof-generated* concepts are neither 'specifications', nor 'generalisations' of the naive concepts. The impact of proofs and refutations on naive concepts is much more revolutionary than that: they *erase* the crucial naive concepts completely and *replace* them by proof-generated concepts." (p. 95)

The motivation for replacing the naive concept by the proof-theoretic one can again be cast in terms of fruitfulness: the proof-theoretic concept enables us to formulate an improved conjecture, which eventually might lead us to the formulation of a theorem.

2.5 Outline of our method

The Lakatosian approach to concept-formation provides a very different view on concept-formation than that suggested by our case study of Church's thesis: rather than arising out of conceptual analysis, the Lakatosian view entails that mathematical concepts develop in response to emerging counterexamples and the analysis of proofs. Lakatos' approach can therefore be viewed as a 'mathematics first' approach, as it takes the driving force behind mathematical concept-formation to be the construction of proofs and counterexamples rather than philosophical considerations.

The fact that Lakatos' method of proofs and refutations does not seem to apply to the explication of the notion of effective computability shows that the method might not be suitable to model every instance of mathematical concept-formation. However, the method does seem applicable to the task at hand. As in Lakatos' case study, we have a preliminary conjecture that contains an informal notion lacking a precise definition. Moreover, this conjecture has received criticism in the form of counterexamples, namely the instances of nonlinearity involving the Rosser sentence and the instances of nonlinearity provided by Hamkins in [22].

Following the Lakatosian approach, I therefore propose the following method for obtaining a more precise definition of the notion of a natural theory:

1. For each counterexample, study the proof of incomparability.
2. Identify the key characteristic(s) of the theories involved that are exploited by the proof.
3. Formulate a proof-generated concept of natural theory based on these characteristics.
4. Assess whether the proof-generated concept is similar the intuitive notion of a natural theory.

Let us make some remarks on the method just outlined.

First, it must be noted that our method diverts from the method of proofs and refutations in one significant way. Whereas the latter teaches us to study a proof of the preliminary conjecture, the former suggests an analysis of the proofs demonstrating that the suggested counterexamples are indeed instances of nonlinearity. The simple reason for this is that we lack a proof suggestion for the linearity conjecture, even in Lakatos' informal sense. As a consequence, our method instructs one to identify properties of *unnatural* theories rather than properties of natural theories. There is a subtle difference here, since identifying a condition that is necessary for proving

nonlinearity is not equivalent to providing a condition that is necessary for proving linearity. Nevertheless, by making the notion of an unnatural theory more precise, we will necessarily also obtain a more precise notion of a natural theory.

Second, recall that the task we set out at the end of section 2.3 was to sharpen the intuitive, pretheoretic notion of a natural theory into a prototheoretic one that is of a static nature. As our method entails that the proof-generated concept is based on inherent characteristics of the theories involved, the proof-generated concept is bound to be static.

Third, step 4 needs to be justified, as this step does not follow from the Lakatosian approach. For Lakatos, the proof-generated concept may fully replace the intuitive notion and need not share any particular features with the initial concept. This is because, for Lakatos, the main purpose of the proof-generated concept is to facilitate an increase in knowledge by providing us with a better understanding of the proof by which it was induced. Our aim, however, is to provide a sharpening of the intuitive notion, and so we need to assess whether the proof-generated concept indeed provides such a sharpening. Moreover, step 4 creates an important balance in our method: whereas step 1 to 3 will necessarily provide us with a concept of natural theory that dismisses the proposed counterexamples as genuine counterexamples to the linearity conjecture, step 4 ensures that this dismissal of the counterexamples may not be trivial.

In the second part of this thesis, we will apply the method just outlined. In Chapter 3, we will carry out step 1 and step 2 using the known instances of nonlinearity, namely those provided by the construction involving the Rosser sentence as presented in Theorem 1.7 and those provided by Hamkins in a recent draft [22]. Building on the proof analyses of Chapter 3, we will carry out step 3 and step 4 in Chapter 4.

Part II

Application of the Lakatosian approach

Chapter 3

Analyzing the counterexamples

In this chapter, I will present three types of potential counterexamples to the linearity conjecture, that is, instances of incomparability in terms of consistency strength. For each of these three types, I will attempt to identify the key characteristics of the theories involved that are exploited in the proof of their incomparability.

The proofs in this chapter are due to Hamkins [22]. I have presented his results in a more elaborate and self-contained manner, highlighting details of his constructions that I deem important as to grasp the key ideas that these proofs employ. In particular, I have made the proof and axiom predicates involved, which Hamkins left implicit, explicit.

3.1 Mathematical preliminaries

In order to present and analyze the (alleged) counterexamples to the conjecture, we need some mathematical preliminaries.

We will be working with classical first-order theories, either in the language of arithmetic $\mathcal{L}_A$ or in the language of set theory $\mathcal{L}_S$. Given such a first-order language $\mathcal{L}$, we define an *$\mathcal{L}$-theory* to be a set T of $\mathcal{L}$-sentences, whose elements will be referred to as the *axioms* of T. For any $\mathcal{L}$-formula ϕ, we will write $T \vdash \phi$ if ϕ is derivable from T in some complete derivation system for first-order logic; a proof with axioms in T will be called a *T-proof*. Let us emphasize here that we will not assume our theories to be closed under consequences, that is, if $T \vdash \phi$ then we need not have $\phi \in T$. In general, the theories under consideration will only consist of non-logical axioms; in particular, when writing PA and ZFC, we refer to the set of non-logical axioms of Peano arithmetic and the set of non-logical axioms of Zermelo-Fraenkel set theory with the axiom of choice, respectively. Given a theory T and a sentence ϕ, we will write $T + \phi$ to refer to the theory obtained when adding ϕ as an axiom to T, i.e. to refer to the theory $T \cup \{\phi\}$.

We will generally need our theories to allow for an recursive axiomatization.

Definition 3.1. *Let T be an $\mathcal{L}$-theory. Then T is recursively axiomatizable if the following are recursively decidable: (1) which words in $\mathcal{L}$ are well-formed formulas/sentences, (2) which formulas are axioms of T, and (3) which finite sequences of formulas constitute a first-order derivation.*

Note that for any recursively axiomatisable theory T, it will be decidable whether a finite sequence of formulas constitutes a T-proof. The languages we will consider always satisfy (1) and (3), so the only interesting requirement is (2). As we view theories as sets of axioms, for our purposes a

recursively axiomatizable theory is simply a recursive set of sentences. Note that, for such a set, different decision procedures are possible; given a recursively axiomatizable theory T, we will refer to a particular decision procedure for T as *a recursive axiomatization of* T.

Let us now restrict ourselves to the language of arithmetic $\mathcal{L}_A$, which we take to consist of the non-logical signature $(0, S, +, \cdot)$. We will say that an arithmetical sentence is *true* if and only if it is true in the standard model $\mathbb{N}$ of the natural numbers.

Recall that the Gödel numbering enables us to encode arithmetical formulas by natural numbers in a primitive recursive manner. That is, there exists a primitive recursive algorithm that given a finite arithmetical expression outputs its unique Gödel code, and there exists a primitive recursive algorithm that given a Gödel code outputs the finite arithmetical expression corresponding to that code. For any formula ϕ, we let $\ulcorner \phi \urcorner$ denote its Gödel code. Moreover, in a similar manner we can encode finite sequences of formulas by natural numbers, and so in particular we can encode proofs.

Now fix some recursively axiomatizable theory T. Gödel's crucial observation for proving his incompleteness theorems was that we can construct an arithmetical formula $\mathrm{Prf}_T(x, y)$ that *captures* the relation 'x is a T-proof of y' in the following sense:

1. if $n \in \mathbb{N}$ is the Gödel code of a T-proof of ϕ then $\mathsf{PA} \vdash \mathrm{Prf}_T\left(\overline{n}, \overline{\ulcorner \phi \urcorner}\right)$;
2. if $n \in \mathbb{N}$ is **not** the Gödel code of a T-proof of ϕ then $\mathsf{PA} \vdash \neg\, \mathrm{Prf}_T\left(\overline{n}, \overline{\ulcorner \phi \urcorner}\right)$.

Here we use the standard notation for numerals, i.e. for any natural number n we let $\overline{n}$ be an abbreviation for applying the successor function n-times to the constant 0. For example, $\overline{2}$ is an abbreviation for $S(S(0))$. Moreover, in the sequel, we will always let n and m denote natural numbers.

The fact that the binary relation 'is a T-proof of' can be captured in PA follows from the more general fact that all recursive functions can be captured in PA.[1] Moreover, they can be captured by a formula of low complexity in terms of the arithmetical hierarchy, namely by a so-called Σ_1 formula.

Definition 3.2. *A partial numerical function $f : \mathbb{N}^k \to \mathbb{N}$ is captured by a formula $\phi(x_1, \ldots, x_k, y)$ in T if for any $m_1, \ldots, m_k, n \in \mathbb{N}$ the following hold:*

1. *if $f(m_1, \ldots, m_k) = n$, then $T \vdash \phi(\overline{m_1}, \ldots, \overline{m_k}, \overline{n})$;*
2. *$T \vdash \forall x \forall y(\phi(\overline{m_1}, \ldots, \overline{m_k}, x) \wedge \phi(\overline{m_1}, \ldots, \overline{m_k}, y) \to x = y)$.*

Here the statement $f(m_1, \ldots, m_k) = n$ implies that $f(m_1, \ldots, m_k)$ is defined.[2]

Definition 3.3. *The sets of arithmetical formulas Σ_n, Π_n and Δ_n for $n \in \mathbb{N}$ are inductively defined as follows:*

(i) both Σ_0 and Π_0 consist of precisely the arithmetical formulas that only contain bounded quantifiers;

(ii) Σ_{n+1} is the smallest set that contains Π_n and is closed under conjunction, disjunction, bounded quantification and (unbounded) existential quantification;

[1] In fact, they can already be captured in the weaker arithmetical system known as Robinson arithmetic.

[2] We are following the terminology used by Smith [56]. Our notion of *capturing* is also referred to as *binumerating* (e.g. by Feferman [11]), *strongly representing* (e.g. by Picollo [47]) or *numeralwise expressing* (e.g. by Auerbach [2]) in the literature.

(iii) Π_{n+1} *is the smallest set that contains* Σ_n *and is closed under conjunction, disjunction, bounded quantification and (unbounded) universal quantification,*

Theorem 3.4. PA *can capture all (partial) recursive functions by a* Σ_1 *formula.*

For a proof of this result, see e.g. the third chapter of Hájek & Pudlák [19]. Let us just make a remark about how one can construct a formula that captures a recursive function f. The language $\mathcal{L}_A$ is expressive enough to fully describe the recursive definition of f. Recall that any recursive function can be defined in terms of the initial functions (the constant zero function, the successor function and the projection functions) subjected to the operations of composition, primitive recursion and minimization. Without going into the details, we can then give an inductive definition of a formula ϕ capturing f as follows.

(1) We can capture the zero function by the formula $\phi_Z(x, y) := (y = 0)$ and the successor function by $\phi_S(x, y) := (S(x) = y)$. The binary projection function projecting onto the first coordinate, for example, can be captured by $\phi_{P_1^2}(x, y, z) := (x = z)$; the other projection functions can be captured in a similar way.

(2) If h is the composition $g_2 \circ g_1$ of unary functions g_1 and g_2, which are captured by $\phi_1(x, y)$ and $\phi_2(x, y)$ respectively, then h is captured by the function $\exists z(\phi_1(x, z) \wedge \phi_2(z, y))$.

(3) Suppose h is defined via primitive recursion on a unary g_1 and a tertiary g_2, i.e. we have $h(x, 0) = g_1(x)$ and $h(x, n + 1) = g_2(x, n, h(n))$, and that these functions are captured by $\phi_1(x, y)$ and $\phi_2(x, y, z, u)$ respectively. We can use a clever way of encoding sequences of natural numbers (by using the so-called *β-function*) to obtain a formula $\phi(x, y, z)$ that captures the statement "There is a sequence of numbers $n_0, \ldots, n_y$ such that $\phi_1(x, n_0)$, and for all $w < y$ we have $\phi_2(x, w, n_w, n_{w+1})$ and $n_y = z$". Then $\phi(x, y, z)$ captures h.

(4) Suppose h is defined via minimization on a binary g, i.e. $h(x) = \mu y(g(x, y) = 0)$, and that g is captured by $\phi_g(x, y, z)$. Then h is captured by the function

$$\phi(x, y) := \phi_g(x, y, 0) \wedge (\forall z \leq y)(z = y \vee \exists u(u \neq 0 \wedge \phi_g(x, z, u))).$$

In steps (2) to (4), the construction generalizes to functions of different arities in a straightforward manner. One can check that formula ϕ obtained in this manner from any recursive function f is indeed a Σ_1 formula.

The point of reviewing this construction is to stress the fact that the formula ϕ capturing f actually fully describes, through some clever encoding, how the function f is built up from the initial functions. In other words, we can view ϕ as describing an effective procedure that calculates f. Smith refers to such formulas as *canonical* in the following sense:

> "A wff that captures a [recursive] function f by being constructed so as to systematically reflect a full [recursive] definition of f ... will be said to *canonically capture* the function."
> (Smith [56], p. 129)

Note that this definition is an informal one. Intuitively, it is capturing in this canonical sense that we would like our formal definition of 'capturing' to, well, capture. However, as Smith makes clear, it is not straightforward how to make this notion precise. Note for example that, for any recursive function f, there will be many different formulas capturing f in our formal sense: if ϕ captures f in T and ψ is any theorem of T, the formula $\phi \wedge \psi$ will also capture f in T. A formal definition of 'canonically capturing' would have to exclude formulas carrying redundant

information like this. Moreover, note there will also be many different formulas that *do* capture a recursive f in a canonical way; for example, the encoding in step (3) could be done in many different ways and there will be many different recursive definitions of f. Nevertheless, it should be clear that for every recursive function f there is at least one formula that canonically captures f, namely the one obtained from the construction above applied to some recursive definition of f. We will therefore assume in what follows that whenever we introduce a formula as capturing some recursive function, it was constructed is this canonical way.

Since any recursive relation has a recursive characteristic function, it follows from Theorem 3.4 that every recursive relation can be captured by a Σ_1 formula in PA in the following sense.

Definition 3.5. *A numerical relation R of arity k is captured by a formula $\phi(x_1, \ldots, x_k)$ in T if for any $m_1, \ldots, m_k \in \mathbb{N}$ the following hold:*

1. *if $R(m_1, \ldots, m_k)$ is true then $T \vdash \phi(\overline{m_1}, \ldots, \overline{m_k})$;*
2. *if $R(m_1, \ldots, m_k)$ is false then $T \vdash \neg\phi(\overline{m_1}, \ldots, \overline{m_k})$.*

It is not hard to see that the relation $R(m, n)$ defined by 'm encodes a T-proof of the formula encoded by n' is indeed recursive: one can decode m and n using a recursive algorithm, and since T is a recursively axiomatized theory there also exists a recursive procedure to check whether the sequence encoded by m forms a first-order derivation whose axioms are in T and whose conclusion is the formula encoded by n. Thus we obtain that PA indeed captures $R(m, n)$ by some formula $\mathrm{Prf}_T(x, y)$. Due to our recursive description of $R(m, n)$, we may assume that $\mathrm{Prf}_T(x, y)$ is built using a Σ_1 formula $\mathrm{Ax}_T(z)$ that (canonically) captures the property 'encodes a sentence in T'. Given any another recursively axiomatized theory S, we will assume that $\mathrm{Prf}_S(x, y)$ is obtained by simply replacing $\mathrm{Ax}_T(z)$ by $\mathrm{Ax}_S(z)$ in $\mathrm{Prf}_T(x, y)$. Moreover, given some sentence ϕ, we will assume $\mathrm{Ax}_{T+\phi}(z)$ is given by $\mathrm{Ax}_T(z) \vee z = \overline{\ulcorner \phi \urcorner}$.

We will refer to $\mathrm{Prf}_T(x, y)$ as the *proof predicate* of T. In general, we will be working with theories T that extend PA; in this case, the formula $\mathrm{Prf}_T(x, y)$ will also capture $R(m, n)$ in T. We will also be working with ZFC and extensions of ZFC; these theories are not an extension of PA in the strict sense, as ZFC is not formulated in the language of arithmetic. However, in the language of set theory we can still define the natural numbers and the arithmetical operations on them in such a way that they satisfy the Peano axioms; the standard way to define the constant 0 and the successor function S is $0 := \emptyset$ and $S(x) := x \cup \{x\}$, and the set of natural numbers is then defined as the least set that contains 0 and is closed under S. Formally, we say that PA is *interpretable* in ZFC, meaning that there exists a translation τ that maps arithmetical formulas to set theoretic formulas such that $\mathsf{PA} \vdash \phi$ implies $\mathsf{ZFC} \vdash \tau(\phi)$ for any arithmetical formula ϕ. It follows that ZFC, and any other theory that interprets PA, can also capture its own provability relation.[3] In the sequel, we will sometimes be sloppy and leave the translation τ implicit; for example, if T interprets PA we will sometimes write $T \vdash \phi$ instead of $T \vdash \tau(\phi)$ for an arithmetical formula ϕ.

With the proof predicate $\mathrm{Prf}_T(x, y)$ at hand, we can define the provability predicate and the consistency sentence of T as follows.

Definition 3.6. *For a recursively axiomatized theory T, the provability predicate is defined as $\mathrm{Prov}_T(x) := \exists y\, \mathrm{Prf}_T(y, x)$. For any sentence ϕ, we will abbreviate $\mathrm{Prov}_T\left(\overline{\ulcorner \phi \urcorner}\right)$ by $\Box_T \phi$.*

Definition 3.7. *For a recursively axiomatized theory T, the consistency sentence is defined as $\mathrm{Con}(T) := \neg\Box_T \bot$.*[4]

[3] See Chapter 6 in Lindström [39] for a detailed treatment of interpretability.

[4] This notation is bit misleading, since T is not a arithmetical term; a more natural choice would perhaps be Con_T. However, this will be unpleasant to read as T becomes more complex, so we will stick with $\mathrm{Con}(T)$.

Note that the provability predicate is a Σ_1 formula. Also note that the provability predicate, and thereby the consistency sentence, depends on a particular choice of for formula $\mathrm{Prf}_T(x, y)$; the results in this chapter will be independent of our choice of $\mathrm{Prf}_T(x, y)$, as long as we construct it in a canonical manner as described above using an axiom predicate $\mathrm{Ax}_T(x)$ that is Σ_1.

We will list some useful properties of provability predicates and consistency sentences, which be found in chapter 1 of [39].

Lemma 3.8. *Let T and S be recursively axiomatized theories. For any sentences ϕ and ψ, the following hold:*

(i) *if $T \vdash \phi$, then $\mathsf{PA} \vdash \Box_T \phi$;*

(ii) $\mathsf{PA} \vdash (\Box_T \phi \wedge \Box_T(\phi \to \psi)) \to \Box_T \psi$;

(iii) $\mathsf{PA} \vdash \Box_T \phi \to \Box_T(\Box_T \phi)$;

(iv) *if $T \vdash \phi \to \psi$, then $\mathsf{PA} \vdash \Box_T \phi \to \Box_T \psi$;*

(v) $\mathsf{PA} \vdash (\Box_T \phi \wedge \Box_T \neg\phi) \to \neg\,\mathrm{Con}(T)$;

(vi) $\mathsf{PA} \vdash \forall x(\mathrm{Ax}_S(x) \to \mathrm{Ax}_T(x)) \to (\mathrm{Con}(T) \to \mathrm{Con}(S))$;

(vii) $\mathsf{PA} \vdash \Box_{T+\phi} \psi \leftrightarrow \Box_T(\phi \to \psi)$;

(viii) $\mathsf{PA} \vdash \Box_T \neg\phi \leftrightarrow \neg\,\mathrm{Con}(T + \phi)$ *and* $\mathsf{PA} \vdash \Box_T \phi \leftrightarrow \neg\,\mathrm{Con}(T + \neg\phi)$.

Properties (i)-(iii) are known as the Hilbert-Bernays-Löb provability conditions; these need to be satisfied in order for the second incompleteness theorem to hold for T. Let us remark here that these conditions need not be satisfied if $\mathrm{Prov}_T(x)$ is not canonical; we will see an example of this in the next chapter.

Properties (i), (v) and (viii) have the following corollary, which will be useful to prove the instances of nonlinearity.

Corollary 3.9. *Let T be a recursively axiomatized theory and ϕ a sentence.*

(i) $\mathsf{PA} \vdash \Box_T \phi \wedge \mathrm{Con}(T) \to \mathrm{Con}(T + \phi)$.

(ii) *If $T \vdash \phi$, then $\mathsf{PA} \vdash \mathrm{Con}(T) \to \mathrm{Con}(T + \phi)$ and $\mathsf{PA} \vdash \mathrm{Con}(T) \to \neg\,\mathrm{Con}(T + \neg\phi)$.*

Proof. (i) By 3.8(viii) we have $\mathsf{PA} \vdash \neg\,\mathrm{Con}(T+\phi) \to \Box_T \neg\phi$ and by 3.8(v) we have $\mathsf{PA} \vdash \Box_T \neg\phi \to \neg\Box_T \phi \vee \neg\,\mathrm{Con}(T)$. This implies that $\mathsf{PA} \vdash \neg\,\mathrm{Con}(T + \phi) \to \neg(\Box_T \phi \wedge \mathrm{Con}(T))$; contraposing gives the desired result.

(ii) If $T \vdash \phi$, then $\mathsf{PA} \vdash \Box_T \phi$ by 3.8(i). By (i), we then obtain $\mathsf{PA} \vdash \mathrm{Con}(T) \to \mathrm{Con}(T + \phi)$. Moreover, from 3.8(viii) it follows that $\mathsf{PA} \vdash \neg\,\mathrm{Con}(T + \neg\phi)$ and so *a fortiori* we have $\mathsf{PA} \vdash \mathrm{Con}(T) \to \neg\,\mathrm{Con}(T + \neg\phi)$. □

The property of proving every true Σ_1 sentence is known as Σ_1-*completeness*.[5] The following lemma states that PA is PA-provably Σ_1-complete and can be found in chapter 1 of [39].

Lemma 3.10. *For any Σ_1 sentence ϕ, if ϕ is true then $\mathsf{PA} \vdash \phi$. Moreover, this result is provable in PA, i.e. we have $\mathsf{PA} \vdash \phi \to \Box_{\mathsf{PA}} \phi$.*

The second key ingredient for Gödel's incompleteness proof, which we will also need in our proofs below, is known as the *diagonalization lemma* or *Gödel's fixed point lemma*.

[5] Recall that by 'true' we mean true in the standard model of arithmetic.

Lemma 3.11. *(Diagonalization Lemma). Let $\phi(x)$ be an arithmetical formula with one free variable x. Then there exists an arithmetical sentence γ in such that* $\mathsf{PA} \vdash \gamma \leftrightarrow \phi\left(\overline{\ulcorner \gamma \urcorner}\right)$.

As the particular construction of the fixed point γ will be of interest to us later, a proof of the lemma will be here.

Proof. Let $\mathrm{D}(x, y)$ be a formula that in PA captures the recursive function d that is defined by $d(\ulcorner \psi(x) \urcorner) = \ulcorner \psi(\overline{\ulcorner \psi(x) \urcorner}) \urcorner$; on natural numbers that do not encode a formula with one free variable, d acts as the identity.[6] Let n be the Gödel code of the formula $\forall y(D(x, y) \to \phi(y))$ and let γ be the formula $\forall y(D(\overline{n}, y) \to \phi(y))$. Then $\ulcorner \gamma \urcorner = d(n)$, and so since $D(x, y)$ captures d in PA we have $\mathsf{PA} \vdash \forall y(D(\overline{n}, y) \leftrightarrow y = \overline{\ulcorner \gamma \urcorner})$. So we obtain

$$\mathsf{PA} \vdash \gamma \leftrightarrow \forall y(y = \overline{\ulcorner \gamma \urcorner} \to \phi(y))$$

and thus indeed $\mathsf{PA} \vdash \gamma \leftrightarrow \phi\left(\overline{\ulcorner \gamma \urcorner}\right)$. □

Lastly, we will need the formalized version of the second incompleteness theorem for nice theories T, which can be found in Smith [56].

Definition 3.12. *A theory T is nice if it is consistent, recursively axiomatizable and interprets* PA.

Theorem 3.13. *Let T be a nice theory. Then $T \vdash \mathrm{Con}(T) \to \neg \Box_T \mathrm{Con}(T)$.*

3.2 Counterexample 1: the Rosser sentence

Before considering Hamkins' new instances of nonlinearity, let us revisit the instance given in Theorem 1.7 in the case that $T = \mathsf{PA}$. In order to analyze this counterexample properly, we first need to provide a proof of the following theorem.

Theorem 3.14. *Let γ be the Rosser sentence for the theory* $\mathsf{PA} + \mathrm{Con}(\mathsf{PA})$. *Then the theories* $\mathsf{PA} + \gamma$ *and* $\mathsf{PA} + \neg\gamma$ *have incomparable consistency strength over the base theory* PA.

Note that by showing that two theories have incomparable consistency strength over some base theory B, we also show that these theories are incomparable over any weaker base theory. So the fact that PA might be stronger than the preferred base theory does not make this instance of nonlinearity less convincing; if anything, it makes it more convincing.

Proof. [7] We will construct a sentence γ that is true if and only if for any proof of γ from $\mathsf{PA} + \mathrm{Con}(\mathsf{PA})$ there is a smaller proof of $\neg\gamma$ from $\mathsf{PA} + \mathrm{Con}(\mathsf{PA})$; by 'smaller', we mean smaller in terms of Gödel codes. This sentence γ is known as the *Rosser sentence* for $\mathsf{PA} + \mathrm{Con}(\mathsf{PA})$. For brevity, let us write $T := \mathsf{PA} + \mathrm{Con}(\mathsf{PA})$.

Note that the relation $R(m, n)$ defined by 'm encodes a T-proof of the negation of the formula encoded by n' is recursive; we will let $\overline{\mathrm{Prf}}_T(x, y)$ denote a formula capturing this relation in PA. We can then define the Rosser provability predicate RProv_T as follows:

$$\mathrm{RProv}_T(x) := \exists y \left(\mathrm{Prf}_T(y, x) \wedge (\forall z \leq y) \neg \overline{\mathrm{Prf}}_T(z, x)\right). \tag{3.1}$$

[6] This formula is often referred to as the *diagonal function*.

[7] This proof is a worked-out version of Hamkins' proof of Theorem 3 in [22]. Part of the proof is based on the proof of Rosser's theorem in [56].

Note that for any sentence ϕ, $\mathrm{RProv}_T\left(\overline{\ulcorner\phi\urcorner}\right)$ is true if and only if there is a T-proof of ϕ such that there is no smaller T-proof of $\neg\phi$. By the Diagonalization Lemma 3.11, there now exists a sentence γ that is a fixed point of $\neg\,\mathrm{RProv}_T(x)$, so we have

$$\mathsf{PA} \vdash \gamma \leftrightarrow \neg\,\mathrm{RProv}_T\left(\overline{\ulcorner\gamma\urcorner}\right). \tag{3.2}$$

As desired, we obtain that γ is indeed true if and only if for any T-proof of γ there is a smaller T-proof of $\neg\gamma$.

Assuming T to be consistent, one can show that $T \nvdash \gamma$ and $T \nvdash \neg\gamma$.[8] Suppose, for contradiction, that $T \vdash \gamma$. Then there is a T-proof of γ, so since our proof predicate captures the proof relation, it follows that there is some Gödel code (i.e. some natural number) n such that $T \vdash \mathrm{Prf}_T(\overline{n}, \overline{\ulcorner\gamma\urcorner})$. Moreover, consistency of T implies that $\neg\gamma$ is *not* provable from T, so using that $\overline{\mathrm{Prf}}_T$ captures its corresponding relation gives us $T \vdash \neg\overline{\mathrm{Prf}}_T(\overline{m}, \overline{\ulcorner\gamma\urcorner})$ for all $m \leq n$. Since T extends PA, one can show that T then proves $(\forall z \leq \overline{n})\neg\overline{\mathrm{Prf}}_T(z, \overline{\ulcorner\gamma\urcorner})$. Thus T proves $\mathrm{RProv}_T(\ulcorner\gamma\urcorner)$, which contradicts (3.2).

The argument that T cannot prove $\neg\gamma$ is similar. If $T \vdash \neg\gamma$, then we have $T \vdash \overline{\mathrm{Prf}}_T(\overline{n}, \overline{\ulcorner\gamma\urcorner})$ for some n. Moreover, since consistency of T implies $T \nvdash \gamma$, a similar argument as above gives us $T \vdash (\forall y \leq \overline{n})\neg\mathrm{Prf}_T(y, \overline{\ulcorner\gamma\urcorner})$, which implies $T \vdash \forall y(\mathrm{Prf}_T(y, \overline{\ulcorner\gamma\urcorner}) \to \overline{n} \leq y)$. But then $T \vdash \forall y(\mathrm{Prf}_T(y, \overline{\ulcorner\gamma\urcorner}) \to \overline{n} \leq y \wedge \overline{\mathrm{Prf}}_T(\overline{n}, \overline{\ulcorner\gamma\urcorner}))$ and thus

$$T \vdash \forall y(\mathrm{Prf}_T(y, \overline{\ulcorner\gamma\urcorner}) \to (\exists z \leq y)\overline{\mathrm{Prf}}_T(z, \overline{\ulcorner\gamma\urcorner})).$$

This shows that $T \vdash \neg\,\mathrm{RProv}_T(\overline{\ulcorner\gamma\urcorner})$, which again contradicts (3.2).

Having established that γ is independent of T, we can now show that $\mathsf{PA}+\gamma$ and $\mathsf{PA}+\neg\gamma$ are incomparable. We will first show that $\mathsf{PA}+\gamma \not\leq \mathsf{PA}+\neg\gamma$ and subsequently that $\mathsf{PA}+\neg\gamma \not\leq \mathsf{PA}+\gamma$.

Since T cannot prove γ, it follows that the theory $T+\neg\gamma$ is consistent and thus has a model; let's call it M. Then $M \models \neg\gamma$, so M believes there is a T-proof of γ such that there is no smaller T-proof of $\neg\gamma$. The key observation is now that this situation is described by a Σ_1 sentence and thus provable in PA, that is, M thinks that PA proves that such a T-proof of γ exists. Formally, from $M \models \mathsf{PA}$ and $M \models \neg\gamma$ together with (3.2) we obtain

$$M \models \mathrm{RProv}_T\left(\overline{\ulcorner\gamma\urcorner}\right).$$

Since the Rosser provability predicate is a Σ_1 formula and M is a model of PA, it follows from Lemma 3.10 that

$$M \models \Box_{\mathsf{PA}}\,\mathrm{RProv}_T\left(\overline{\ulcorner\gamma\urcorner}\right). \tag{3.3}$$

From (3.2) and Lemma 3.8(iv), we then find

$$M \models \Box_{\mathsf{PA}}\neg\gamma, \tag{3.4}$$

so M thinks that PA proves $\neg\gamma$. Since M is a model of $\mathrm{Con}(\mathsf{PA})$, it follows from Corollary 3.9(ii) that M is a model of both $\mathrm{Con}(\mathsf{PA}+\neg\gamma)$ and $\neg\,\mathrm{Con}(\mathsf{PA}+\gamma)$. As M is also a model of PA, this implies that $\mathsf{PA} \nvdash \mathrm{Con}(\mathsf{PA}+\neg\gamma) \to \mathrm{Con}(\mathsf{PA}+\gamma)$ and thus we obtain $\mathsf{PA}+\gamma \not\leq \mathsf{PA}+\neg\gamma$.

Conversely, since T cannot prove $\neg\gamma$, it follows that the theory $T+\gamma$ is consistent. By the second incompleteness theorem, this theory cannot prove its own consistency and so there exists a model M' of $T+\gamma+\neg\,\mathrm{Con}(T+\gamma)$. By Lemma 3.8(viii), M' then believes that there is a

[8] In fact, this is how one proves Gödel's First Incompleteness Theorem.

T-proof of $\neg\gamma$. In particular, as M' is a model of PA, M' believes there must be a smallest such proof, let's call it P. Since $M' \models \gamma$, M' thinks that there can be no T-proof of γ that is smaller than P. In other words, M' believes that there exists a T-proof P of $\neg\gamma$ such that everything smaller than P is not a T-proof of γ. Note that this situation can be described by a Σ_1 sentence, so M' believes that PA proves that such a proof P exists. Since the existence of such a P implies $\neg\,\mathrm{RProv}_T(\overline{\ulcorner\gamma\urcorner})$ and thereby γ, it follows that M' believes that PA proves γ. Since $M' \models \mathrm{Con}(\mathsf{PA})$, it follows from Corollary 3.9(ii) that M' is a model of $\mathrm{Con}(\mathsf{PA}+\gamma)$ and $\neg\,\mathrm{Con}(\mathsf{PA}+\neg\gamma)$. Thus we obtain $\mathsf{PA}+\neg\gamma \not\leq \mathsf{PA}+\gamma$. □

Note that the proof can easily be adapted to show that there are incomparable theories $T+\gamma$ and $T+\neg\gamma$ over the base theory PA, where T is any nice theory and γ is the Rosser sentence for $T+\mathrm{Con}(T)$.

3.2.1 Proof analysis

Despite the fact that we lack a clear definition of naturalness, there seems to be consensus in the literature that theories like $\mathsf{PA}+\gamma$ and $\mathsf{PA}+\neg\gamma$ are in fact unnatural. After introducing these counterexamples, Hamkins writes

> "Nobody likes [these] examples of nonlinearity [...] Those sentences are viewed as unnatural—weird self-referential logic-game trickery."[9] (Hamkins [22], p. 8).

He then describes the main goal of his paper as providing examples of nonlinearity that *are* natural. Before considering those, let us pause and reflect at the counterexample at hand. In particular, let us try to answer the question: where does the trickery occur?

First note that, as a syntactic object, there is nothing fishy about the sentence γ. It is a well-defined sentence in the language of arithmetic that we could write down explicitly, without using any abbreviations, if we were willing to do the work. After all, γ is obtained via the Diagonal Lemma, the proof of which is constructive in the sense that it gives us an effective procedure to construct γ from the proof predicates Prf and $\overline{\mathrm{Prf}}$, and these proof predicates can in turn also be written down explicitly in the language of arithmetic. Thus, albeit long and rather complex, the sentence γ is syntactically unproblematic.

The potential problem, then, must lie in the semantics of γ. We have constructed γ so that it is PA-provably equivalent to $\mathrm{RProv}_T\left(\overline{\ulcorner\gamma\urcorner}\right)$. This means that in any model of PA, the sentence γ is true precisely when for any T-proof of γ, there is smaller T-proof of $\neg\gamma$. We thus see that the truth of γ corresponds to the truth of a statement about what the theory T can prove about γ. Studying the proof of Theorem 3.14, it is exactly this property of γ that is exploited in order to obtain that $\mathsf{PA}+\gamma$ and $\mathsf{PA}+\neg\gamma$ have incomparable consistency strength.

In the proof of Theorem 3.14, we prove two incomparability statements. The first one, $\mathsf{PA}+\gamma \not\leq \mathsf{PA}+\neg\gamma$, is the easier one: we take a model M of $T+\neg\gamma$ and use the fact that $\neg\gamma$ is a Σ_1 sentence to conclude that M thinks that PA can already prove $\neg\gamma$ and thereby that M thinks that $\mathsf{PA}+\neg\gamma$ is consistent while $\mathsf{PA}+\gamma$ is not. For the other incomparability statement, $\mathsf{PA}+\neg\gamma \not\leq \mathsf{PA}+\gamma$, we need to do a bit more work as γ is not Σ_1.[10] To prove this statement, we take a model M' of $T+\gamma+\neg\,\mathrm{Con}(T+\gamma)$ and aim to show that M believes that PA proves γ.

[9] Let us stress here that Hamkins just describes the math community here. He himself does not seem to think this, as he believes naturality talk to be thoroughly unsatisfactory.

[10] Note that, in order to obtain incomparable theories $\mathsf{PA}+\gamma$ and $\mathsf{PA}+\neg\gamma$, it is necessary that γ and $\neg\gamma$ are not both Σ_1 sentences. If they were, then Σ_1-completeness would imply that PA proves one of them (namely the one that is true in the standard model) and so the other would be inconsistent with PA. However, inconsistent theories are trivially comparable to any theory in terms of consistency strength.

The reason we succeed is that M' believes that PA proves that there is a T-proof of $\neg\gamma$ with no smaller proof of γ, and due to the self-referencing semantics of γ this suffices to show that M' believes that PA proves γ.

Let us abstract away from the details and consider the general structure of the proof just described. In order to show that two theories T and S are incomparable in terms of consistency strength, we need to construct a model M in which $\mathrm{Con}(T)$ is true and $\mathrm{Con}(S)$ is false and a second model M' in which $\mathrm{Con}(S)$ is true and is $\mathrm{Con}(T)$ false. In order to show that $\mathrm{Con}(S)$ or $\mathrm{Con}(T)$ is false in a model, we need to show that this model believes that there exists a proof of a contradiction from S or T, respectively. Note that, in general, we assume the theories S and T to be consistent, and thus we expect that such a proof does not exist in the standard model of arithmetic. Actually finding such a proof should therefore be a hopeless task. Nevertheless, we want to show that such a proof does exist inside some nonstandard model.

In the proof of Theorem 3.14, this is exactly where the self-referencing semantics of γ helps us out. First, in order to obtain the models M and M', we need the sentences γ and $\neg\gamma$ to be independent from $\mathsf{PA} + \mathrm{Con}(\mathsf{PA})$. Of course, the Rosser sentence was originally constructed for just this purpose: being able to prove the Rosser sentence implies being able to prove that there is a smaller proof of its negation, and thus no consistent theory can prove its Rosser sentence; similarly, no consistent theory can prove the negation of its Rosser sentence, as being able to prove this negation implies being able to prove that there is a proof of the Rosser sentence. Second, we need to show that M is a model of $\mathrm{Con}(\mathsf{PA} + \neg\gamma) + \neg\,\mathrm{Con}(\mathsf{PA} + \gamma)$ and that M' is model of $\mathrm{Con}(\mathsf{PA} + \gamma) + \neg\,\mathrm{Con}(\mathsf{PA} + \neg\gamma)$; as explained above, the first statement follows from Σ_1-completeness, while for the second statement the proof relies again on the fact that γ is equivalent to a statement about its own provability.

3.3 Counterexample 2: Representing numbers by computations

Hamkins' first construction of an instance of nonlinearity that he considers to be natural involves a representation of natural numbers in terms of what Hamkins calls a *universal computable function*. This is an algorithm that can compute *any* function, as long as it is run in the right model.

Theorem 3.15. *Let T be a nice theory. We can construct a computable function U_T such that for any partial function $f : \mathbb{N} \to \mathbb{N}$ there is a model M of T such that inside M, U_T computes f. Formally, for any arithmetical formula $\phi_{U_T}(x, y)$ that canonically captures U_T, the following holds: for each $n \in \mathbb{N}$, if $f(n)$ is defined then*

$$M \models \phi_{U_T}(\overline{n}, \overline{f(n)}),$$

and otherwise

$$M \models \neg\exists y \phi_{U_T}(\overline{n}, y).$$

Proof. We assume the reader to be familiar with the indexation of recursive functions and with Kleene's Recursion Theorem.[11] We will define a recursive function U_T that, informally, can be described as follows: on any input, U_T searches for a T-proof that U_T itself does *not* compute some finite function h, and if such a proof is found then it *does* behave exactly like this function.

[11] For details, see e.g. [57].

We will first define a recursive function g. Given a natural number p, we let $g(p)$ be an index of the recursive function that is computed by the following program:

> On input n, the program generates an arithmetical formula $\phi_p(x, y)$ that canonically captures the function indexed by p and then searches systematically for a finite list of pairs of natural numbers $((n_1, m_1), \ldots, (n_k, m_k))$ and for a T-proof of the statement
>
> "$\phi_p(x, y)$ does not compute the partial function h given by the input-output pairs $(n_1, m_1), \ldots, (n_k, m_k)$".[12]
>
> More precisely, the program employs some encoding of pairs (P, h) of T-proofs P and finite functions h, and checks these pairs in the order of increasing code. If a pair (P, h) is found such that P is T-proof of "$\phi_p(x, y)$ does not compute h", then the program outputs $h(n)$ whenever it exists, and otherwise it loops endlessly.

Note that g is then a total computable function, and thus Kleene's recursion theorem implies that there exists an index e such that the functions indexed by e and $g(e)$ are identical. We let U_T be the recursive function indexed by e. Let $\phi_{U_T}(x, y)$ be an arithmetical formula that canonically captures U_T in PA.

Since T is consistent, it cannot prove the statement "$\phi_{U_T}(x, y)$ does not compute h" for any finite function h. If it could, then there would be a finite function h^* and a T-proof P of "$\phi_{U_T}(x, y)$ does not compute h^*" such that the pair (P, h^*) is least in terms of the encoding mentioned above. However, by construction of U_T, that would mean that U_T actually *does* compute the function h^*. Moreover, by Σ_1-completeness, the existence of such a pair (P, h^*) would be provable in PA. Since $\phi_{U_T}(x, y)$ *canonically* captures U_T, the particular syntactic construction of $\phi_{U_T}(x, y)$ must reflect that the existence of such a pair (P, h^*) implies that $\phi_{U_T}(x, y)$ computes h^*. Thus it follows that it would be provable in PA that $\phi_{U_T}(x, y)$ computes h^*. In particular, T would prove that $\phi_{U_T}(x, y)$ computes h^*, but then T would prove a contradiction as we also have a T-proof P showing that ϕ_{U_T} does not compute h^*.

We obtain that for each finite function h, the statement "$\phi_{U_T}(x, y)$ computes h" is consistent with T. In particular, for any partial function $f : \mathbb{N} \to \mathbb{N}$ it follows that any finite subset of the set of sentences

$$\{\phi_{U_T}(\overline{n}, \overline{f(n)}) : f(n) \text{ is defined}\} \cup \{\neg\exists y \phi_{U_T}(\overline{n}, y) : f(n) \text{ is not defined}\}$$

is consistent with T. By the compactness theorem, we obtain that the set of all these sentences is consistent with T. Thus there is a model M of T in which they all hold. □

Note that the partial function f can even be a non-computable function. There is nothing paradoxical about this, as the model in which f will be computed by U_T will necessarily be a nonstandard model of arithmetic. In the standard model, the function U_T is not defined on any input, as consistency of T implies that there can be no T-proof of "$\phi_{U_T}(x, y)$ does not compute h" for any finite function h.

Thinking of U_T as an algorithm rather than a partial recursive function, in the sequel we will often identify U_T with its arithmetical description $\phi_{U_T}(x, y)$ and write the informal statement "U_T halts on n" to abbreviate the sentence $\exists y \phi_{U_T}(\overline{n}, y)$.

[12] This statement can be spelled out as follows:

$$\neg\forall x \forall y \Big(\phi_p(x, y) \leftrightarrow \bigvee_{1 \leq i \leq k} x = \overline{n_i} \wedge y = \overline{m_i} \Big)$$

Theorem 3.15 shows that a description of a computable process need not be enough to pin down a particular natural number. Hamkins uses this to formulate statements that seem to be large cardinal hypotheses while having incomparable consistency strength over ZFC. For a proof of the following theorem, we refer to reader to Theorem 8 in [22].

Theorem 3.16. *Consider the theory*

$$T := \mathsf{ZFC} + \text{“}\textit{There exist infinitely many inaccessible cardinals}\text{”}.$$

Then the theories

$$\mathsf{ZFC} + \text{“}\textit{There are } U_T(n) \textit{ many inaccessible cardinals}''$$

for $n \in \mathbb{N}$ are pairwise incomparable in terms of consistency strength over ZFC.

Note that $U_T(n)$ is not a numeral here, but the result of ‘running’ the algorithm U_T on input n. In the statement “There are $U_T(n)$ many inaccessible cardinals$''$ we are implicitly assuming that U_T halts on n. Thus, in the standard model, this statement is trivially satisfied as U_T does not halt on any input in this model.

As pointed out by Hamkins, the proof of Theorem 3.16 is not really dependent on inaccessible cardinals; it also works for other large cardinal notions. In fact, the key ideas already come to light when restricting ourselves to the simpler case of arithmetic, which is what we will do.

Theorem 3.17. *Let T be the theory* $\mathsf{PA} + \mathrm{Con}(\mathsf{PA})$. *Then the theories*

$$\mathsf{PA} + \text{“}U_T \textit{ halts on input } n''$$

for $n \in \mathbb{N}$ are pairwise incomparable in terms of consistency strength over PA.

Proof. Let $n, m \in \mathbb{N}$ such that $n \neq m$. By Theorem 3.15, there exists a model M of $\mathsf{PA}+\mathrm{Con}(\mathsf{PA})$ in which U_T halts on n and not on m. By the construction of U_T, the fact that U_T halts on n in M implies that M thinks that there is a least pair (P, h) of a finite function h and a T-proof P that U_T does **not** compute h, and that this function h is defined on n but not on m. Due to Σ_1-completeness, the existence of this pair (P, h) can be proven in PA, and so M thinks that PA can prove that U_T halts on a specific finite set of natural numbers that includes n but not m. In particular, M thinks that PA proves that U_T halts on n and not on m. Since M is a model of $\mathrm{Con}(\mathsf{PA})$, we find by Corollary 3.9(ii) that M is a model of both $\mathrm{Con}(\mathsf{PA}+\text{“}U_T \text{ halts on input } \overline{n}'')$ and $\neg\,\mathrm{Con}(\mathsf{PA} + \text{“}U_T \text{ halts on input } \overline{m}'')$. This shows that

$$\mathsf{PA} \nvdash \mathrm{Con}(\mathsf{PA} + \text{“}U_T \text{ halts on input } \overline{n}'') \to \mathrm{Con}(\mathsf{PA} + \text{“}U_T \text{ halts on input } \overline{m}'').$$

□

Using the theorem above, one can show that for any decision problem A such that there is a many-one reduction from the halting to A, there are instances of incomparability among the statements of the form “This is an element of A” and “This is not an element of A”. Hamkins concludes from this that there are instances of incomparability among the seemingly natural statements of the form

“This specific set of tiles admits a tiling”,

or

“This specific finite group presentation is the trivial group”.

It is important to note, however, that for the incomparable statements one obtains in this way the particular arithmetical description of the 'specific' set of tile or group presentation will be given in terms of the universal computable function.

3.3.1 Proof analysis

Let us analyze the proof of the instances of nonlinearity given in Theorem 3.17. Let us abbreviate "U_T halts on input n" by ϕ_n for each $n \in \mathbb{N}$. In order to show that $\mathsf{PA} + \phi_n$ and $\mathsf{PA} + \phi_m$ are incomparable for $n \neq m$, we need to argue that there exists a model M in which $\mathrm{Con}(\mathsf{PA} + \phi_n)$ is true while $\mathrm{Con}(\mathsf{PA} + \phi_m)$ is false. Notice that the general structure of the argument is quite similar to that given in the proof of Theorem 3.14, as it consists of the following two steps:

1. It is shown that the statements ϕ_n and ϕ_m are strongly independent[13] over $\mathsf{PA} + \mathrm{Con}(\mathsf{PA})$; this is taken care of in the proof of Theorem 3.15 with T instantiated by $\mathsf{PA} + \mathrm{Con}(\mathsf{PA})$. Thus there exists a model M of $\mathsf{PA} + \mathrm{Con}(\mathsf{PA}) + \phi_n + \neg\phi_m$.

2. It is shown that M thinks that PA proves both ϕ_n and $\neg\phi_m$, which implies that M thinks that $\mathsf{PA} + \phi_n$ is consistent while $\mathsf{PA} + \phi_m$ is not.

How are the claims in these two steps established? For the first claim, the proof exploits that the program U_T is constructed as to ensure that being able to prove that U_T does not compute some finite function h is tantamount to being able to prove that there is a T-proof that U_T *does* compute h. Therefore, no consistent theory T can prove that U_T computes some finite function h; in particular, T cannot prove that U_T does not compute a finite function h that halts on n but not on m. So it is consistent with T that U_T halts on n and not on m.

For the second claim, in order to show that M believes that PA proves both ϕ_n and $\neg\phi_m$, we again use that U_T halting on n and not on m already implies that there exists a T-proof that U_T does not compute some least function h that halts on n and not on m. Due to Σ_1-completeness, the existence of this proof is provable in PA, and due to the construction of U_T the existence of this proof is equivalent to U_T computing h. Thus PA can prove that U_T halts on n and not on m. We thus find that Theorem 3.17 is a special instance of the following, more general result:

Theorem 3.18. *For T a nice theory, let ϕ_1 and ϕ_2 be strongly independent Σ_1 sentences over $T + \mathrm{Con}(T)$. If both $\phi_1 \wedge \neg\phi_2$ and $\neg\phi_1 \wedge \phi_2$ are equivalent to a Σ_1 sentence, then $T + \phi_1$ and $T + \phi_2$ have incomparable consistency strength over PA.*

Let us emphasize that, in addition to the general proof structure, the construction of the particular sentences involved in Theorems 3.17 and 3.14 is based on a very similar idea. In both proofs, we construct a sentence η that is equivalent to one that canonically captures a statement about the provability of η. This equivalence takes the form of a sophisticated version of the liar paradox, namely it is of the form

η is true if and only if there is a T-proof of $\neg\eta$ of some particular form.

In the proof of Theorem 3.14 and Theorem 3.17, the role of η is played by the sentences $\neg\gamma$ and "U_T computes h", respectively. This 'liar paradox' then ensures that η is both independent of T and equivalent to a Σ_1 sentence.

[13] Two sentences ϕ and ψ are strongly independent over a theory T if every Boolean combination of these sentences is independent over T.

3.4 Counterexample 3: Cautious enumerations

Hamkins' second construction of an instance of nonlinearity is quite different from the two instances considered so far. Instead of adding a particular kind of axiom to a nice theory T, this construction leaves the set of axioms of T unchanged; what is altered is the way in which the axioms of T are represented. Recall that, when constructing the provability predicate of a nice theory T, we have some particular recursive algorithm in mind that checks whether a sentence is an element of T; this algorithm is then canonically captured by the formula $\mathrm{Ax}_T(x)$. For theories such has PA and ZFC, there is a standard way in which this recursive algorithm can proceed. In the case of PA, given a sentence ϕ the algorithm can check, one by one, if ϕ is equal to one of the finitely many axioms of Robinson arithmetic, and lastly whether ϕ is an instance of the Induction Schema. Similarly, for ZFC, the algorithm can simply check whether ϕ is an instance of the Comprehension Schema, an instance of the Replacement Schema or equal to one the (finitely many) other axioms of ZFC. However, there are different ways of representing the theories PA and ZFC and thereby different ways of building their axiom predicate. Hamkins' idea is to build an axiom predicate $\mathrm{Ax}_T(x)$ that captures a so-called *cautious enumaration* of the theory T.

A cautious enumeration of a theory T is a recursive enumeration of T that, while listing the axioms of T, searches for some indication that the theory T might be unsound; if such an indication is found, then the enumeration is halted. Hamkins considers cautious enumerations of ZFC to be

> "... both sensible and realistic – in this sense it is a natural theory – for if we were actually enumerating ZFC and a proof was pointed out to us along the way that the theory we have already committed to proves the full ZFC theory to be inconsistent, then we would have ample reason to pause and reflect on whether we should continue with the enumeration. ... The cautious enumeration is what we would actually do – so it is a natural theory."
> (Hamkins [22], p.16)

The first cautious enumeration of ZFC that Hamkins considers is the following.

Definition 3.19. *Let* ZFC° *denote the theory we obtain if we cautiously enumerate* ZFC *as follows. While enumerating the axioms of* ZFC*, search for a* ZFC*-proof of* $\neg\,\mathrm{Con}(\mathsf{ZFC})$*; if such a proof is found, the enumeration is halted.*

Note that, if ZFC is indeed consistent, then ZFC° will have exactly the same axioms as ZFC. We thus take these theories to determine the same *set* of axioms; the difference lies in the way we enumerate them, that is, in the representation of these axioms. Crucially, the provability predicates of ZFC and ZFC° will be different.

Following Hamkins' description, the the axiom predicate of ZFC° could be defined as follows:

$$\mathrm{Ax}_{\mathsf{ZFC}^\circ}(x) := \mathrm{Ax}_{\mathsf{ZFC}}(x) \wedge (\forall y \leq x)\, \mathrm{NPrf}_{\mathsf{ZFC}}\left(y, \overline{\ulcorner \neg\, \mathrm{Con}(\mathsf{ZFC}) \urcorner}\right), \tag{3.5}$$

where $\mathrm{NPrf}_{\mathsf{ZFC}}(y, z)$ canonically captures the recursive relation 'y does **not** encode a ZFC-proof of the formula encoded by z'.[14] The idea is here that a natural number n encodes an axiom of ZFC° if and only if it encodes an axiom of ZFC and if the first n proofs from ZFC do not prove the sentence $\neg\,\mathrm{Con}(\mathsf{ZFC})$. This suggestion fits well with Definition 3.19. However, one might feel that it does not fit well with Hamkins description of cautious enumerations quoted above, which suggests that the enumeration must only search for a proof of $\neg\,\mathrm{Con}(\mathsf{ZFC})$ from "the theory we

[14] We need to bother with $\mathrm{NPrf}_{\mathsf{ZFC}}(y, z)$ instead of simply writing $\neg\,\mathrm{Prf}_{\mathsf{ZFC}}(y, z)$ to ensure that our axiom predicate remains Σ_1.

have already committed to". In order to account for this, one could define a new proof predicate $\mathrm{NPrf}'_{\mathsf{ZFC}}(y,z)$ that is obtained by replacing $\mathrm{Ax}_{\mathsf{ZFC}}(x)$ by $\mathrm{Ax}_{\mathsf{ZFC}}(x) \wedge x \leq y$ in $\mathrm{NPrf}_{\mathsf{ZFC}}(y,z)$.[15] The intuition is here that $\mathrm{NPrf}'_{\mathsf{ZFC}}(y,z)$ then captures the relation 'y does not encode a proof of the formula encoded by z employing at most the first y axioms of ZFC'. We can then define the axiom predicate for ZFC° as follows:

$$\mathrm{Ax}'_{\mathsf{ZFC}^\circ}(x) := \mathrm{Ax}_{\mathsf{ZFC}}(x) \wedge (\forall y \leq x)\, \mathrm{NPrf}'_{\mathsf{ZFC}}\left(y, \overline{\ulcorner \neg\, \mathrm{Con}(\mathsf{ZFC})\urcorner}\right). \tag{3.6}$$

The following result will be independent of our choice between (3.5) and (3.6).

Theorem 3.20. *The theory* ZFC° *has strictly lower consistency strength than* ZFC *over the base theory* ZFC*.*

Proof. Since every axiom of ZFC° is an axiom of ZFC, and the axiom predicates reflect this, it follows from Lemma 3.8(vi) that $\mathsf{ZFC} \vdash \mathrm{Con}(\mathsf{ZFC}) \to \mathrm{Con}(\mathsf{ZFC}^\circ)$.

Now suppose, for contradiction, that $\mathsf{ZFC} \vdash \mathrm{Con}(\mathsf{ZFC}^\circ) \to \mathrm{Con}(\mathsf{ZFC})$. Then

$$\mathsf{PA} \vdash \Box_{\mathsf{ZFC}}(\mathrm{Con}(\mathsf{ZFC}^\circ) \to \mathrm{Con}(\mathsf{ZFC})),$$

and so it follows from (vii) and (viii) of Lemma 3.8 that

$$\mathsf{PA} \vdash \neg\, \mathrm{Con}(\mathsf{ZFC} + \mathrm{Con}(\mathsf{ZFC}^\circ) + \neg\, \mathrm{Con}(\mathsf{ZFC})). \tag{3.7}$$

Assuming that ZFC is consistent, there is a model M of $\mathsf{ZFC} + \mathrm{Con}(\mathsf{ZFC})$ in which $\neg\, \mathrm{Con}(\mathsf{ZFC} + \mathrm{Con}(\mathsf{ZFC}))$ holds. By Lemma 3.8(viii), it follows that M believes that there exists a ZFC-proof of $\neg\, \mathrm{Con}(\mathsf{ZFC})$. By construction of $\mathrm{Ax}_{\mathsf{ZFC}^\circ}(x)$, M then thinks that ZFC° consists of only finitely many axioms. It is a well-known result that ZFC proves the consistency of each of its finite fragments.[16] As this result is capable of being formalized in ZFC, and since M is a model of ZFC, it follows that M believes that ZFC proves $\mathrm{Con}(\mathsf{ZFC}^\circ)$. By the formalized second incompleteness theorem, that is Theorem 3.13, and the fact that $M \models \mathrm{Con}(\mathsf{ZFC})$, it follows that $M \models \neg\Box_{\mathsf{ZFC}}\, \mathrm{Con}(\mathsf{ZFC})$. By Lemma 3.8(viii), M then thinks that $\mathsf{ZFC} + \neg\, \mathrm{Con}(\mathsf{ZFC})$ is consistent. However, as M thinks that ZFC proves $\mathrm{Con}(\mathsf{ZFC}^\circ)$, it follows from Corollary 3.9(i) that M then believes that $\mathsf{ZFC} + \mathrm{Con}(\mathsf{ZFC}^\circ) + \neg\, \mathrm{Con}(\mathsf{ZFC})$ is consistent. This contradicts (3.7). □

As the proof of Theorem 3.20 shows, any cautious enumeration of ZFC will still have a consistency strength that is comparable to the standard representation of ZFC, as its set of axioms will necessarily be a subset of the axioms of ZFC.[17] However, Hamkins has showed that among the cautious enumerations of ZFC, we can find instances of incomparability. These instances are given by cautious enumerations based on the universal computable function.

Definition 3.21. *Let* U_T *denote the universal computable function for* $T := \mathsf{ZFC} + \mathrm{Con}(\mathsf{ZFC})$*. For each* $n \in \mathbb{N}$*, let* $\mathsf{ZFC}^{(n)}$ *denote the cautious enumeration of* ZFC *that, while enumerating the axioms of* ZFC*, runs* U_T *on* n*; if* U_T *halts on* n*, then the enumeration halts.*

Note that U_T only halts on an input if there is a proof from T that U_T does not behave the way it does, which would imply that T has false consequences. The halting of U_T therefore indeed serves as an indication that T is unsound.

[15] As with $\mathrm{Prf}_{\mathsf{ZFC}}(x)$, we are assuming that $\mathrm{NPrf}_{\mathsf{ZFC}}(x)$ as been constructed in a canonical manner from $\mathrm{Ax}_{\mathsf{ZFC}}(x)$.

[16] This follows from the reflection theorem; see e.g. Corollary II.5.4 in Kunen [36].

[17] In fact, one can iterate the construction of ZFC° to obtain an infinitely descending chain in the hierarchy of consistency strength of the form $\mathsf{ZFC} > \mathsf{ZFC}^\circ > \mathsf{ZFC}^{\circ\circ} > \ldots$. This provides an (alleged) counterexample to the stronger version of the linearity conjecture stating that the consistency hierarchy of natural theories is a well-order.

For any $n \in \mathbb{N}$, the axiom predicate of $\mathsf{ZFC}^{(n)}$ can be defined as follows. Let $\psi_n(x)$ be a formula that canonically captures the recursive predicate $P_n(m)$, which is true if and only if at least one of the following holds:

1. m does **not** encode a pair (P, h) of a finite function h that halts on n and a T-proof P of "U_T computes h";
2. there exists a $k < m$ such that k encodes a pair (P, h) of a finite function h that does **not** halt on n and a T-proof P of "U_T computes h".

Then we can define

$$\mathrm{Ax}_{\mathsf{ZFC}^{(n)}}(x) := \mathrm{Ax}_{\mathsf{ZFC}}(x) \wedge (\forall y \leq x)\psi_n(y) \tag{3.8}$$

Intuitively, this definition states that a natural number m encodes an axiom of $\mathsf{ZFC}^{(n)}$ if and only if m encodes an axiom of ZFC and for the first m pairs (P, h) of T-proofs P and finite functions h, if P proves "U_T computes h" and h halts on n then there is a smaller pair (P', h') such that P' proves "U_T computes h'" and h' does not halt on n.

Theorem 3.22. *The theories* $\mathsf{ZFC}^{(n)}$ *for* $n \in \mathbb{N}$ *have pairwise incomparable consistency strength over* ZFC.

Proof. Let $n, m \in \mathbb{N}$ such that $n \neq m$. Suppose, for contradiction, that $\mathsf{ZFC} \vdash \mathrm{Con}\left(\mathsf{ZFC}^{(n)}\right) \rightarrow \mathrm{Con}\left(\mathsf{ZFC}^{(m)}\right)$. Following the same reasoning as in the proof of Theorem 3.20, we then obtain

$$\mathsf{PA} \vdash \neg\,\mathrm{Con}\left(\mathsf{ZFC} + \mathrm{Con}\left(\mathsf{ZFC}^{(\mathsf{n})}\right) + \neg\,\mathrm{Con}\left(\mathsf{ZFC}^{(\mathsf{m})}\right)\right). \tag{3.9}$$

As above, let T denote $\mathsf{ZFC} + \mathrm{Con}(\mathsf{ZFC})$. By Theorem 3.15, there exists a model M of $\mathsf{ZFC} + \mathrm{Con}(\mathsf{ZFC})$ in which U_T halts on n but not on m. This means that M thinks that there is a least pair (P, h) such that P proves "U_T computes h", and that this h halts on n but not on m. By construction of the axiom predicates for $\mathsf{ZFC}^{(n)}$ and $\mathsf{ZFC}^{(m)}$, it then follows that M believes that $\mathsf{ZFC}^{(n)}$ consists of finitely many axioms while the axioms of $\mathsf{ZFC}^{(m)}$ are exactly the axioms of ZFC. By the same reasoning as in the proof of Theorem 3.20, it follows that M believes that ZFC proves $\mathrm{Con}\left(\mathsf{ZFC}^{(n)}\right)$. Moreover, since M is a model of $\mathrm{Con}(\mathsf{ZFC})$, by the formalized second incompleteness theorem we obtain that M thinks that ZFC cannot prove $\mathrm{Con}(\mathsf{ZFC})$ and thereby that it cannot prove $\mathrm{Con}\left(\mathsf{ZFC}^{(m)}\right)$. So M believes $\mathsf{ZFC} + \neg\,\mathrm{Con}\left(\mathsf{ZFC}^{(m)}\right)$ to be consistent. As ZFC proves $\mathrm{Con}\left(\mathsf{ZFC}^{(n)}\right)$ in M, it follows that M believes $\mathsf{ZFC} + \mathrm{Con}\left(\mathsf{ZFC}^{(n)}\right) + \neg\,\mathrm{Con}\left(\mathsf{ZFC}^{(m)}\right)$ to be consistent. This contradicts (3.9).

□

3.4.1 Proof analysis

As before, let ϕ_n be the sentence "U_T halts on input n" for $n \in \mathbb{N}$. The proof of Theorem 3.22 consists of two main steps:

1. It is shown that the conditions for halting of the cautious enumerations $\mathsf{ZFC}^{(n)}$ and $\mathsf{ZFC}^{(m)}$ are strongly independent over $\mathsf{ZFC} + \mathrm{Con}(\mathsf{ZFC})$. That is, the sentences ϕ_n and ϕ_m are shown to be strongly independent over $\mathsf{ZFC} + \mathrm{Con}(\mathsf{ZFC})$; this is taken care of in the proof of Theorem 3.15. Thus there exists a model M of $\mathsf{ZFC} + \mathrm{Con}(\mathsf{ZFC}) + \phi_n + \neg\phi_m$.

2. It is shown that M believes that $\mathsf{ZFC}^{(n)}$ is a finite fragment of ZFC and that $\mathsf{ZFC}^{(m)}$ has exactly the same axioms as ZFC. It then follows, by arguments that do not rely on the particular definition of $\mathsf{ZFC}^{(n)}$ and $\mathsf{ZFC}^{(m)}$ anymore, that M believes that $\mathsf{ZFC} + \mathrm{Con}(\mathsf{ZFC}^{(n)}) + \neg\,\mathrm{Con}(\mathsf{ZFC}^{(m)})$ is consistent, which suffices to show that $\mathsf{ZFC}^{(m)} \not\leq \mathsf{ZFC}^{(n)}$.

We have already discussed the first step in the former proof analysis. The second step is quite straightforward: to show that M believes that $\mathsf{ZFC}^{(n)}$ has finitely many axioms, the proof exploits the fact that the axiom predicate of $\mathsf{ZFC}^{(n)}$ captures that there will be no axioms larger than the pair (P, h) witnessing that U_T halts on n. Similarly, to show that M believes that $\mathsf{ZFC}^{(m)}$ has the same axioms as ZFC, the proof exploits the fact that the axiom predicate of $\mathsf{ZFC}^{(m)}$ captures that all axioms of ZFC will be included if there is a pair (P, h) witnessing that U_T does not halt on m.

The key characteristics of the cautious enumerations $\mathsf{ZFC}^{(n)}$ that enables us to prove the incomparability of their consistency strength thus appears to be the following: how many axioms of ZFC they employ depends on a particular condition, and these conditions have been constructed as to be strongly independent over $\mathsf{ZFC} + \mathrm{Con}(\mathsf{ZFC})$.

Chapter 4

Proof-generated concepts of natural theory

In this chapter, I will formulate two proof-generated concepts of natural theory based on the proof analyses provided in Chapter 3. There we saw that the key ideas behind the proofs of the counterexamples of the first and second type are very similar: I will therefore treat these counterexamples as of one type, and propose one proof-generated concept of natural theory that characterises the theories occurring in both counterexamples as unnatural. Subsequently, I will turn to the third type of counterexample, the cautious enumerations, and propose a second proof-generated concept of natural theory that will deem Hamkins' cautiously enumerated theories unnatural.

It will be argued that the proposed proof-generated concepts have the following properties: (1) they dismiss the presented instances of nonlinearity as counterexamples to the linearity conjecture, (2) they are of a static rather than a dynamic nature, and (3) they are similar (in the Carnapian sense) to the intuitive notion of a natural theory as one that 'arises in practice' and 'has a genuinely mathematical idea to it'.

4.1 Self-reference

In the first chapter, we saw that objections raised to the naturalness of incomparable theories involving the Rosser sentence often refer to the self-referential nature of this sentence. A first suggestion for a sharpening of naturalness that disqualifies the first counterexample could therefore be this: a natural theory is one that, inter alia, does not contain self-referencing axioms. Before attempting to make this idea more precise, let us note that our proof analyses in Chapter 3 suggest that self-reference is also a crucial ingredient of the second and third counterexample: both Hamkins' instances of nonlinearity rely on the universal computable function, the definition of which involves an obvious element of self-reference. Lack of self-reference therefore seems to be a promising candidate for a proof-generated concept of natural theory.

Anticipating that his counterexamples might be dismissed with the charge of self-reference, Hamkins provides two arguments against the claim that self-reference is disqualifying for naturalness. The first argument starts from the premise that self-reference is the central feature in Cantor's diagonal argument for uncountability of the reals and in Russell's refutation of the naive comprehension scheme in set theory. He writes:

"These arguments are surely amongst the founding central ideas of [set theory], and the diagonalization idea is woven deeply throughout it. Furthermore, these diagonalizations are fundamentally the same as used to prove the fixed-point lemmas that lead to the Gödel and Rosser sentences. What can be the coherent philosophy of 'natural' that counts the constructions of Cantor and Russell as natural, but not the fundamentally similar construction of the Gödel and Rosser sentences?" (p. 28)

The implied assumptions here appear to be that (1) Cantor's and Russell's arguments form natural mathematical constructions, and (2) the classification of theories as unnatural on the basis of self-referencing axioms such as the Rosser sentence would imply the classification of Cantor's and Russell's argument as unnatural. The first claim depends on one's conception of a natural mathematical argument; Hamkins seems to take it as a given that any argument that plays a fundamental role in mathematical practice, is a natural one. As we are not concerned with a particular conception of a natural argument, we might simply grant Hamkins the first claim. I argue that the second claim, however, is too strong. We are looking for a sharpening of the notion of a natural theory, not of the notion of naturalness in general. Therefore, while agreeing that self-reference features in natural mathematical arguments and constructions, one can still hold that self-reference is not an acceptable property of any natural axiomatic theory. In particular, while agreeing that the *construction* of the Rosser sentence γ for $\mathsf{PA} + \mathrm{Con}(\mathsf{PA})$ is natural, one might still argue that adding γ as an axiom to PA yields an unnatural theory.

Hamkins' second argument seems to form a more serious obstacle for taking self-reference as a property of unnatural theories. He claims that among the common extensions of ZFC, one finds axioms that seem to partake in self-reference. Some large cardinal hypotheses are equivalent to the existence of a certain elementary embedding; for example, the existence of a measurable cardinal is equivalent to the existence of a nontrivial elementary embedding $j : V \to M$ of the set-theoretic universe V into some class model M. Hamkins argues that the existence of such an embedding involves a notion of self-reference, since it essentially states that for every object x there exists an object $j(x)$ with the same first-order properties in M as x has in V. He writes:

"... the axiom at bottom posits a system of duplicates $j(x)$, whose properties are stated by (self)-reference back to x. In this light, nearly every large cardinal axiom partakes in self-reference." (p. 28)

There are two ways to respond to Hamkins' second argument. First, Hamkins' charge of self-reference in this context is not very convincing. It is not obvious how, when stating that $j(x)$ has the same properties as x, one is describing $j(x)$ by somehow referring to $j(x)$ itself. Second, and more importantly, the kind of self-reference occurring here (if any) is of a very different kind than the kind of self-reference occurring in the case of the Rosser sentence and the universal function. Let us write γ_T for the Rosser sentence of the theory T. The sentences $\neg\gamma_T$ and $\exists x \phi_{U_T}(\overline{n}, x)$ are constructed so that they are true precisely when the theory T can prove their negation in some particular way. So, not only do these sentences involve a more obvious element of self-reference, the property they seem to ascribe to themselves is a *metatheoretic* property with respect to the theory T. Thus, rather than describing a property of numbers (or of sets, when PA is interpreted in ZFC), it seems that $\neg\gamma_T$ and $\exists x \phi_{U_T}(\overline{n}, x)$ are somehow stating *meta-information* about the theory T.

4.2 Axioms carrying meta-information: Proof-generated Concept 1

By meta-information, I refer to any information about the theory as whole, as opposed to *object information* that is about the mathematical objects that the theory is meant to describe. These notions are closely related to the notions of *metalanguage* and *object language*, where the latter is used to refer to the formal language of the theory at hand and the former to the natural language which we use the talk about this formal theory. A naive classification of object information and meta-information would be that object information can be formulated in the object language, whereas meta-information is necessarily posited in the metalanguage. Gödel has showed us, however, that the situation is not that simple. Through coding, we *can* express meta-information inside the object language in an indirect manner: while the statement $\mathrm{RProv}_T(\overline{\ulcorner \gamma_T \urcorner})$ is directly talking about a property of natural numbers, in the sense that someone unaware of the underlying encoding would not be able to see it as anything more, we know that the truth of this particular statement on natural numbers coincides with the truth of the metastatement "for every proof of γ_T from T there is a smaller proof of $\neg\gamma$ from T". Therefore, requesting $\mathrm{RProv}_T(\overline{\ulcorner \gamma_T \urcorner})$ or $\neg\,\mathrm{RProv}_T(\overline{\ulcorner \gamma_T \urcorner})$ holds is tantamount to making a request about what T can or cannot prove.

While the notion of carrying meta-information seems intuitively clear, it is difficult to make it precise. Following our discussion above, one might be inclined to define an object statement as *carrying meta-information* if its truth coincides with the truth of a metastatement. However, this would imply that *every* object statement trivially carries meta-information, as every object statement is either true or false, and thus every object statement would be materially equivalent to either a true or a false metastatement (both of which certainly exist). What needs to be taken into account is an intensional aspect of the object statement, in the sense that it is constructed as to fully reflect the content of the metastatement.

Consider, for example, the metastatement:

(S) For every proof of γ_T from T there is a smaller proof of $\neg\gamma_T$ from T.

Assuming that the theory T is consistent, S is trivially true because there does not exist any proof of γ_T from T (as shown in the proof of Theorem 3.14). The statement S is therefore materially equivalent to any true object statement about natural numbers, e.g. the sentence $\forall x(x+0=x)$. However, the sentence $\forall x(x+0=x)$ does not reflect the content of S in any way. Compare this to the sentence $\mathrm{RProv}_T(\overline{\ulcorner \gamma_T \urcorner})$, which is also materially equivalent to S: assuming that one is aware of the underlying encoding of formulas and proofs involved, one could simply 'see' that this sentence is equivalent to the metastatement S by construction, without even considering the truth values of these two statements. The reason for this is that the proof predicates used to construct $\mathrm{RProv}_T(\overline{\ulcorner \gamma_T \urcorner})$ all *canonically* capture their corresponding proof relation, and so by merely studying their syntactic construction one can see that to satisfy these predicates is to encode a proof from T.

Recall that Smith gave the following definition of 'canonically capturing' in the context of recursive relations:

> "A wff that captures a [recursive] function f by being constructed so as to systematically reflect a full [recursive] definition of f ... will be said to *canonically capture* the function."
> (Smith [56], p. 129)

Extending this notion of canonically capturing beyond recursive relations, we could say that the sentence $\mathrm{RProv}_T(\overline{\ulcorner \gamma_T \urcorner})$ canonically captures the metastatement S, while the sentence $\forall x(x+0=x)$ does not. This leads us to the following informal definition:

A metastatement S is *canonically captured* by an object statement ϕ if ϕ has been constructed as to systematically reflect the content of S.

We can then say that an object statement *carries meta-information* if it canonically captures a metastatement.

In this new terminology, it seems clear that the sentences γ_T and $\exists x \phi_{U_T}(x)$ carry meta-information about the theory T. I therefore propose the following proof-generated concept of natural axiomatic theory:

Proof-generated concept 1: *A natural axiomatic theory is a theory whose axioms do not carry meta-information.*

I will refer to this concept as PC1.

Let us emphasize here that, to save this concept of triviality, it is important that one distinguishes between a theory's axioms and its theorems. In every theory that interprets PA, one can construct sentences that carry meta-information, and in every such theory, one can also *prove* sentences that carry meta-information. For example, we know that for any finite fragment F of PA, PA proves the sentence $\mathrm{Con}(\mathsf{F})$, and this sentence certainly carries meta-information.[1] Thus every theory that interprets PA has consequences that carry meta-information in the sense just described. However, according to our proof-generated concept, what disqualifies $\mathsf{PA} + \gamma$ as a natural theory, and what qualifies PA as a natural one, is that we add a statement carrying meta-information as an *axiom* to the axioms of PA, whereas the axioms PA itself do not seem to carry such information.

The application of PC1 crucially depends on the application of the concept of 'systematically reflecting the content of a metastatement', which is certainly not straightforward; I will address this issue in the next section. However, assuming for now that this application is at least intuitively clear, let us make three observations about PC1:

(1) PC1 disqualifies theories of the form $T + \gamma_{T+\mathrm{Con}(T)}$, $T + \neg\gamma_{T+\mathrm{Con}(T)}$ and $T +$ "$U_{T+\mathrm{Con}(T)}$ *halts on* $\overline{n}$" as natural theories and thereby dismisses counterexamples that are of the first or second type as genuine counterexamples to the linearity conjecture.[2]

(2) PC1 is a *static* rather than a *dynamic* concept. Whether or not the axioms of a theory carry meta-information does not depend on contextual properties of the theory, that is, on its position in the body of mathematical knowledge and practice of a particular time.

(3) PC1 is *similar* to the intuitive, inchoate notion of a natural theory as one that 'has a genuinely mathematical idea to it' and 'arises in practice'.

Let us elaborate on the third point. I take it as a relatively uncontroversial view that the non-logical axioms of theories considered in mathematical practice are stating properties of a particular mathematical object or structure under consideration. The student of mathematics' first encounter with axiomatic systems is likely to consist of axiomatic treatments of natural numbers, reals, sets, groups, vector spaces, and the like. Axioms in such contexts are meant to state properties of these objects or structures, which can then be used to obtain further results in a rigorous way. One can distinguish here between a *describing* or *prescribing* role: the axioms can be viewed as merely describing an independently existing mathematical entity, or one can view these entities as being 'defined into existence' by virtue of the axioms. The reader is free

[1] This result is due to Mostowski [46].

[2] I will address counterexamples of the third type in section 4.4.

to adopt either viewpoint; the point is that, according to both conceptions, the axioms are to be taken as necessarily true properties of the mathematical entities in question and their main purpose is to derive theorems involving these entities.[3]

In my opinion, PC1 can be viewed as a natural continuation of this view: if the axioms of an axiomatic theory ought to state properties of some mathematical entity, and if theorems involving these entities are meant to be derived from these properties, then these axioms should not make a direct request about what the theory as a whole should be able to prove or disprove. In other words, the metatheoretic properties of a theory should *follow from* and not be *stipulated by* its axioms.

4.3 On the intensionality of PC1

The *intension* of a referring statement is commonly explained as the *mode* of referring; this is to be contrasted with the statement's *extension*, which is the thing that is being referred to, i.e. the *referent*. In the context of arithmetization, intensional aspects of an arithmetical predicate are usually taken as those relating to the particular form of the sentence, rather than the set of numbers that satisfy it. Our notion of carrying meta-information clearly has such an *intensional* aspect: whether or not an object statement carries meta-information depends on its particular syntactic construction and is not determined by its mere logical properties. In particular, it is possible that two object statements are provably equivalent while only one carries meta-information; for example, the statements $\neg(0 = 1)$ and $\Box_{\mathsf{PA}}\neg(0 = 1)$ are provably equivalent in PA, yet only the latter appears to be carrying meta-information according to our definition. Such an intensional concept invites some obvious concerns: is such a concept even coherent, and if so, can its application be made precise in concrete logico-mathematical terms?

Gödel's proofs of his two incompleteness theorems sparked a lively debate concerning the intensional aspect of the independent sentences he constructed. The Gödel sentence G_T of a nice theory T, constructed via diagonalisation of the formula $\neg\mathrm{Prov}_T(x)$, is often described as a sentence that 'says of itself that it is unprovable'. Informally, one direction of the proof of the first incompleteness theorem can then be given as follows: if T is consistent and T could prove G_T, then T would prove $\Box_T G_T$ because the provability predicate captures provability, but T would also prove $\neg\Box_T G_T$ because of 'what G_T says'. However, careful inspection of Gödel's proof reveals that no intensional properties of G_T are *necessary* to prove the first incompleteness theorem: his argument can be made in a purely syntactic manner, and whether or not one agrees that G_T indeed says something of itself does not have an effect on G_T's unprovability in T.[4] All

[3]The view on axioms described here is sometimes referred to as the *old* or *traditional* view (see e.g. [63] and [51]). Modern perspectives on axiomatics stress that, apart from viewing axioms as true properties of some mathematical entity, axioms can also be adopted for purely external mathematical reasons. An often mentioned distinction is that between *intrinsic* and *extrinsic* reasons for adopting axioms in set theory, as introduced by Gödel in [17] and further discussed by Maddy [41]: intrinsic reasons appeal to true properties of the concept of set, whereas extrinsic reasons appeal to the fruitfulness of the axiom in question. Examples of axioms of which it has been argued that they are adopted due to their success as a set-theoretic axiom are the Axiom of Choice and the large cardinal hypotheses. I do not deny that extrinsic considerations play a role in the formation and adoption of axioms. However, I would like to stress that they rarely play an exclusive role: even in the case of the Axioms of Choice and the large cardinal hypotheses, accounts can be found providing intrinsic motivations for these axioms. For example, see Ferreirós [12] for an intrinsic justification of the Axiom of Choice based on a conception of 'arbitrary sets' and see Incurvati [26] for a discussion on how the iterative conception of set can serve as an intrinsic justification of the large cardinal hypotheses.

[4]For independence of G_T, one also needs T to be ω-consistent. However, this condition is not necessary to show that T is incomplete: as our proof of Theorem 3.14 shows, the Rosser sentence γ_T can already be shown to be independent of T if T is consistent.

that is needed is that $\mathrm{Prov}_T(x)$ captures provability in the sense of Definition 3.5 and that G_T is a fixed point of $\neg\,\mathrm{Prov}_T(x)$; whether $\mathrm{Prov}_T(x)$ canonically captures the provability relation or how the fixed point G_T is obtained is irrelevant.

The intensional aspect of G_T in the context of the first incompleteness theorem is therefore often viewed as merely heuristic: whereas it guided Gödel in constructing an independent sentence, it does not play a further role in his eventual result. The only significance that is sometimes still bestowed on the intensional aspect of G_T is that it enables us to see that G_T is *true*: because G_T is not provable in T, as established by Gödel's proof, and because G_T is true if and only if G_T itself is unprovable, we obtain that G_T is indeed true.[5]

The situation is different for the second incompleteness theorem. This theorem is widely accepted as a refutation of Hilbert's program that, in a nutshell, aimed at establishing the consistency of all classical, infinitary mathematical reasoning by purely finitary means. What exactly constitute the demands of the program and whether the second theorem really shows them to be unattainable is not an uncontroversial matter. Nevertheless, there seems to be consensus that if such infinitary reasoning is to be captured by a formal, effectively axiomatizable theory T that extends basic arithmetical reasoning, and if these finitary means should be capable of being formalised in a system no stronger than PA, then Gödel's second theorem shows that such a consistency proof is impossible. Thus, the force of the second incompleteness theorem does not lie in the mere fact that it shows that the particular formula $\mathrm{Con}(T)$, whose construction is described in the previous chapter, is underivable in PA; rather, its force lies in establishing that *any* arithmetical sentence that properly formalizes the statement 'T is consistent' is underivable.

In order to show that this is indeed what the second theorem achieves, one must give an account of what it means for a formula to constitute a proper formalization of the metastatement 'T is consistent'. This point is made particularly clear by the fact that, if one is not careful when constructing $\mathrm{Con}(T)$, one can end up with a consistency sentence that *is* derivable in T. Consider, for example, the predicate

$$\mathrm{Prf}'_T(x,y) := \mathrm{Prf}_T(x,y) \wedge \neg\,\mathrm{Prf}_T\left(x, \overline{\ulcorner \bot \urcorner}\right),$$

where $\mathrm{Prf}_T(x,y)$ denotes the canonical proof predicate for T. Assuming T to be consistent, it easy to see that $\mathrm{Prf}'_T(x,y)$ captures the proof relation in the sense of Definition 3.5. Thus, viewed as relations on natural numbers, $\mathrm{Prf}_T(x,y)$ and $\mathrm{Prf}'_T(x,y)$ are perfectly coextensive. However, if we define $\mathrm{Con}'(T) := \neg\exists x\,\mathrm{Prf}'_T\left(x, \overline{\ulcorner \bot \urcorner}\right)$, then $\mathrm{Con}'(T)$ is derivable in T, simply because $\forall x \neg \left(\mathrm{Prf}_T(x, \overline{\ulcorner \bot \urcorner}) \wedge \neg\,\mathrm{Prf}_T\left(x, \overline{\ulcorner \bot \urcorner}\right)\right)$ is a first-order theorem.

Provable consistency sentences like $\mathrm{Con}'(T)$ do not seem to pose a threat to the claim that no proof of consistency of T could be formalized in T. Note that any convincing consistency proof of T should not rest on an assumption that to be a T-proof means, inter alia, not to be a proof of a contradiction. However, this is precisely what the predicate $\mathrm{Prf}'_T(x,y)$ appears to do: its construction suggests that to be a T-proof according to $\mathrm{Prf}'_T(x,y)$ is to be a T-proof (in the informal sense) that does not prove a contradiction. Therefore, $\mathrm{Prf}'_T(x,y)$ fails in genuinely reflecting what it means to be a T-proof.

Work by Hilbert and Bernays [25] that was further refined by Löb [40] showed that for the underivability of $\mathrm{Con}(T)$ in T it suffices that the provability predicate on which $\mathrm{Con}(T)$ is based satisfies the three *HBL-conditions*, also referred to as the *derivability conditions*, which were

[5] Isaacson [27] uses this observation to argue for his thesis that PA is complete when one restricts to *genuinely arithmetical statements*. The latter are defined as statements whose truth can be perceived on the basis of a grasp of the natural numbers alone. To see the truth of G_T, Isaacson argues, one needs to recognize that the arithmetic of natural numbers can be used to encode proof-theoretic notions, and such a recognition requires more than a grasp of arithmetical properties of the natural numbers alone.

mentioned in the previous chapter. I repeat them here:[6]

1. If $T \vdash \phi$ then $T \vdash \Box_T \phi$;
2. $T \vdash (\Box_T \phi \wedge \Box_T(\phi \to \psi)) \to \Box_T \psi$;
3. $T \vdash \Box_T \phi \to \Box_T(\Box_T \phi)$.

Indeed, $\mathrm{Prf}'_T(x, y)$ fails in meeting these conditions. To see this, let $\phi := \neg(0 = 0)$ and let the operator $\Box_T$ be defined from the canonical proof predicate for T and let $\Box'_T$ be defined using $\mathrm{Prf}'_T(x, y)$. Since T proves $\phi \to \bot$ and since $\ulcorner \phi \to \bot \urcorner$ is not equal to $\ulcorner \bot \urcorner$, we have $T \vdash \Box'_T(\phi \to \bot)$. Moreover, since T cannot prove of itself that it does not prove a contradiction, in particular T cannot prove $\neg \Box_T \phi$. So there is a model M of T such that $M \models \Box_T \phi$, and again since $\ulcorner \phi \urcorner$ is not equal to $\ulcorner \bot \urcorner$ we obtain $M \models \Box'_T \phi$. So $M \models \Box'_T \phi \wedge \Box'_T(\phi \to \bot)$, while $M \not\models \Box'_T \bot$ since due to the definition of $\mathrm{Prf}'_T(x, y)$ we have $T \vdash \neg \Box'_T \bot$.

Checking whether a given provability predicate actually satisfies the HBL-conditions is a tedious task.[7] In [11], Feferman provided a simple criterion by showing that any provability predicate $\mathrm{Prov}_T(x)$ that is build in a straightforward way (which is explicitly written out in his paper) by using any Σ_1 formula as an axiom predicate $\mathrm{Ax}_T(x)$ will satisfy the three conditions. Indeed, it is Feferman's construction that we have implicitly been using in the previous chapter.

It appears that the HBL-conditions are generally accepted as necessary conditions for any proof predicate that properly formalizes the proof relation. In [3], Auerbach argues that the HBL-conditions are implied by the semantics of the language of proof theory and thus, since any formalization of this language ought to reflect this intended semantics, the formalization must satisfy the conditions as well. For example, if we can prove, in the informal language of proof theory, that $\mathsf{PA} \vdash \phi$, and if PA itself is to serve as a formalization of our informal reasoning in which $\Box_{\mathsf{PA}}$ acts as the translation of the provability property, then we must have $\mathsf{PA} \vdash \Box_{\mathsf{PA}} \phi$. This establishes the first condition. Similarly, if we can prove the first condition of a given provability predicate, i.e. if we can prove 'if $T \vdash \phi$ then $T \vdash \Box_T \phi$' in the informal language of proof theory, then our formalization should reflect this again, and thus we must have $T \vdash \Box_T \phi \to \Box_T(\Box_T \phi)$.[8]

Gödel's second theorem thus serves as an example of an intensional result that can be formulated in concrete logico-mathematical terms. A second example of such a result is Löb's theorem, which provided an answer to Henkin's question 'Are formulas expressing their own provability provable or independent?' posed in [23].[9] A first reply to Henkin was given by Kreisel [34], who constructed two such formulas, one of which is provable and one refutable. However, Henkin [24] rejected the provability predicates used by Kreisel, since he felt that neither predicate genuinely expressed the notion of provability. In the end, a satisfactory answer was provided by Löb [40], who showed that any fixed point of a provability predicate that satisfies the HBL-conditions *is* provable.

[6] As before, we write $\Box_T \phi$ for $\mathrm{Prov}_T(\overline{\ulcorner \phi \urcorner})$.

[7] See for example the proof that the canonical provability predicate for PA satisfies the HLB-conditions in [56].

[8] Auerbach does not mention the second condition in [3]. This might be due to Jeroslow's result in [29], which shows that the second condition is redundant for proving the second incompleteness theorem if one extends the language of arithmetic with function symbols for each primitive recursive function. However, it is not hard to see how Auerbach's reasoning can be used to justify the second condition as well: in the informal langauge of proof theory, we can prove 'if ϕ and $\phi \to \psi$ are provable in T then ψ is provable T', and thus this must be reflected by the formalization.

[9] This example is taken from a two-part paper by Halbach & Visser ([20], [21]), who give a detailed account of Henkin's question and Löb's answer that focuses on the intensional aspects. They also introduce other intensional problems, which still lack a precise solution.

These cases show that logicians are generally quite capable of recognizing whether an arithmetical formula reflects a certain metamathematical statement, and that in some cases the necessary intensional aspects can be translated into precise, formal terms. This raises the question whether such a formal translation can also be found for PC1.

An attempt at finding such a translation lies beyond the scope of this thesis. Whereas the intensional results just described may stem one hopeful, let me just comment on why finding such a translation is likely to be quite difficult. Consider the sentence $\exists x \phi_{U_T}(\overline{n}, x)$, which we have taken to mean "the algorithm U_T halts on input n", where U_T is defined as the algorithm that searches for a T-proof that this algorithm itself does not act in a certain way. Let us refer to this as the *intended interpretation* of $\exists x \phi_{U_T}(\overline{n}, x)$. This intended interpretation is *corrrect*, in the sense that $\exists x \phi_{U_T}(\overline{n}, x)$ is indeed true if and only if the algorithm U_T halts on input n. However, this is not the only correct interpretation of $\exists x \phi_{U_T}(\overline{n}, x)$; once fully written out in the language of arithmetic, this Σ_1 sentence can be interpreted using the standard interpretation of this language, i.e. it can be interpreted as being true if and only if there exist natural numbers with certain properties and relations; let us call this the *arithmetical interpretation* of $\exists x \phi_{U_T}(\overline{n}, x)$. Of course, apart from the intended and the arithmetical interpretation of $\exists x \phi_{U_T}(\overline{n}, x)$, there will also be other correct interpretations.

Now suppose that, rather than working with the intended interpretation of $\exists x \phi_{U_T}(\overline{n}, x)$, Hamkins had used a different interpretation of $\exists x \phi_{U_T}(\overline{n}, x)$ that did not appeal to meta-information in any way. Suppose moreover, that he had only referred to this interpretation in his proof of incomparability of $\mathsf{PA} + \exists x \phi_{U_T}(\overline{n}, x)$ and $\mathsf{PA} + \exists x \phi_{U_T}(\overline{m}, x)$. If Hamkins had constructed his instance of nonlinearity in this way, it seems that PC1 would not dismiss $\mathsf{PA} + \exists x \phi_{U_T}(\overline{n}, x)$ as an unnatural theory.

The subtle point is here that whether an arithmetical statement carries meta-information crucially depends on our interpretation of it. Of course, this is not just a problem for PC1, but also for the other intensional notions just discussed: the only way to judge whether an arithmetical formula properly reflects the statement 'T is consistent' or 'This sentence is provable' is by inspecting some particular interpretation of it given some choice of Gödel encoding. In the case of the second incompleteness theorem and Löb's theorem, we have managed to find formal conditions that capture the necessary intentensional aspects, namely the HBL-conditions. This does not mean, however, that we found a formal definition of what it means for an arithmetical formula to properly reflect the statement 'T is consistent' or 'This sentence is provable'. After all, the HBL-conditions only constitute necessary conditions for properly reflecting the provability relation. Such a full formalization might not even be possible, simply because the intensional notion of 'properly reflecting' depends on a particular interpretation of the object statement in question and formally we cannot distinguish between different correct interpretations.

These comments aim to show that a (partial) formal treatment of the notion of carrying meta-information will be difficult to obtain. However, the cases of the second incompleteness theorem and Löb's theorem exemplify that such a treatment is not necessarily out of reach and that intentional considerations can at least serve as a heuristic in finding new formal results.

4.4 Surveyable presentation: Proof-generated Concept 2

As noted in the previous chapter, the instances of nonlinearity provided by the cautiously enumerated theories are of a very different nature than those provided by the counterexamples of the first or second kind. The axiomatization of the theories $\mathsf{ZFC}^{(n)}$ for $n \in \mathbb{N}$ is defined using the halting condition $\exists x \phi_{U_T}(\overline{n}, x)$. In our terminology, this halting condition carries meta-information.

However, although $\exists x \phi_{U_T}(\overline{n}, x)$ is used to define the axiomatization of $\mathsf{ZFC}^{(n)}$, it is not an *axiom* of $\mathsf{ZFC}^{(n)}$. Thus the theories $\mathsf{ZFC}^{(n)}$ are not characterised as unnatural by PC1.

The cautiously enumerated theories make clear that consistency strength, as defined in Definition 1.3, is actually not a property of a given set of axioms; rather, it is a property of a certain *presentation* of a set of axioms. As a result of this, when stating what it means for a theory to be natural in the context of the linearity conjecture, one also needs to give an account of how such a theory is to be presented. In particular, this means that PC1 cannot suffice as a full description of what it means to be a natural theory.

It is notable that, in mathematical literature on relative consistency proofs, the presentation of the theories in question is standardly left implicit. In standard textbooks in set theory treating relative consistency proofs, such as those by Jech [28] and Kanamori [30], theories are defined as sets and no mention is made of the particular presentation one is to have in mind when carrying out the relative consistency proofs. After defining consistency strength for his discussion on the consistency hierarchy in [58], Steel writes:

> "There is an intensional aspect here, in that the order really is on presentations of theories, rather than theories, but we shall ignore that detail here." (p. 155)

The reason for this widespread negligence concerning the presentation of theories is, I presume, that the intended presentation is generally very clear. When talking about the set of axioms of ZFC, we think of this set as defined by the following presentation of ZFC: a sentence is an axiom of ZFC if and only if it is equal to an instance of the Comprehension Schema, or to an instance of the Replacement Schema, or to one of the other finitely many axioms. Moreover, when talking about a certain extension $\mathsf{ZFC} + \phi$, we think of its axioms as being presented as follows: a sentence is an axiom of this extension if and only if it equal to an axiom of ZFC, where ZFC is presented in the manner just described, or it is equal to ϕ.

When taking the presentation into account, we see that the theories ZFC and $\mathsf{ZFC}^{(n)}$ are simply different theories. Let us therefore focus on the presentation of the theories $\mathsf{ZFC}^{(n)}$, which comes down to the following: a sentence ϕ is an axiom of $\mathsf{ZFC}^{(n)}$ if and only if it is an axiom of ZFC (presented in the standard way) and the algorithm $U_{\mathsf{ZFC}+\mathrm{Con}(\mathsf{ZFC})}$ does not halt on n after $\ulcorner \phi \urcorner$ steps.

The presentation of $\mathsf{ZFC}^{(n)}$ reminds of what Detlefsen [10] calls *consistency-minded theories*. For Detlefsen, a theory is consistency-minded if it incorporates a consistency constraint into its notion of proof. Starting from a nice theory T with some intended presentation, a consistency-minded theory T' can generally be obtained in two ways: either by directly changing the notion of proof, or indirectly by changing the notion of an axiom. An example of the first way would be to define T'-provability following the Rosser provability predicate $\mathrm{RProv}_T(x)$, that is, to define being a T'-proof as being a T-proof of some formula ϕ such that no smaller T-proof, according to some ω-ordering on T-proofs, proves $\neg\phi$. An example of a second way, first suggested by Feferman [11], would be to define the axioms of T' following the predicate

$$\mathrm{Ax}_{T'}(x) := \mathrm{Ax}_T(x) \wedge \neg\, \mathrm{Prov}_{T\restriction x}(\ulcorner \bot \urcorner), \tag{4.1}$$

where $\mathrm{Prov}_{T\restriction x}(y)$ is defined using the axiom predicate $\mathrm{Ax}_T(z) \wedge z \leq x$. The presentation of T' can then be given as follows: a sentence ϕ is an axiom of T' if and only if ϕ is an axiom of T and ϕ is consistent with the set of T-axioms smaller than ϕ, according to some ω-ordering on formulas.

Both strategies outlined above result in a theory T' that can prove its own consistency sentence $\mathrm{Con}(T')$, as shown in [64]. This means, of course, that the provability predicates

obtained fail to satisfy the HBL-conditions and thus that $\mathrm{Ax}_{T'}(x)$ given in (4.1) cannot be Σ_1.[10] Nevertheless, Detlefsen argues that the notion of proof employed in these consistency-minded theories may be acceptable to the Hilbertian, and thereby aims to establish that the second incompleteness theorem need not demarcate the end of Hilbert's program after all.

Detlefsen's account has found few supporters; a common reply is that the notion of provability employed in these theories strays too far from the intuitive one.[11] In [3], Auerbach provides a second critique on the consistency-minded theories:

> "The recommendation that we reform our mathematical practice and replace the canonical notion of derivability with a Rosser-style one will indeed assure us, quite easily, of consistency. But that epistemic gain is offset by the epistemic loss occasioned by *not knowing what it is that is consistent*." (Auerbach [3], p. 87, emphasis added)

Auerbach's critique highlights an important point: by employing a notion of axiom that is defined in terms of a consistency constraint, our understanding of the theory T' becomes dependent on our understanding of this constraint; as a result, whenever we do not have a clear understanding of when this constraint is satisfied, we also do not have a clear understanding of what the axioms of T' *are*.

Hamkins' cautious enumerations can be viewed as a generalization of Detlefsen's consistency-minded theories. Indeed, Hamkins describes the halting condition of a cautious enumeration as "a certain kind of contrary indicator, a reason to doubt the truth of the theory" (p. 16); Detlefsen's consistency constraints can then be viewed as particular type of such indicators. There is, however, a significant difference between the cautious enumerations discussed in Chapter 3 and Feferman's consistency-minded theory T': whereas it is not decidable which sentences are the axioms of T', it *is* decidable which sentences are the axioms of ZFC° and $\mathsf{ZFC}^{(n)}$. Indeed, the axiom predicates given in (3.5) and (3.8) are Σ_1, whereas the axiom predicate given in (4.1) is not. Thus, in contrast to the consistency-minded theories, the provability predicates of Hamkins' cautious enumerations do satisfy the HBL-conditions.

I would like to argue, however, that Auerbach's critique still stands against the cautious enumerations. Even though we can decide, for any *particular* axiom of ZFC, whether it is an axiom of $\mathsf{ZFC}^{(n)}$, there is no algorithm that can decide whether *every* axiom of ZFC is also an axiom $\mathsf{ZFC}^{(n)}$. A crucial difference between the presentations of ZFC and $\mathsf{ZFC}^{(n)}$ seems to be that, whereas the presentation of the former enables one to fully grasp what the axioms of the theory in question are, the latter does not. Even though ZFC has infinitely many axioms, the fact that they can be stated using finitely many axioms and two axiom schemas provides us with a clear and concise overview of what its axioms are. However, since we cannot decide whether the cautious enumeration underlying $\mathsf{ZFC}^{(n)}$ will halt or not, the presentation of $\mathsf{ZFC}^{(n)}$ does not seem to provide one with such a clear grasp of what the axioms of $\mathsf{ZFC}^{(n)}$ are.

The kind of distinction within recursive axiomatizations I have in mind here is between those that are *surveyable* and those that are not. In the philosophy of mathematics, the notion of surveyability usually pops up the context of proofs. There are two main properties that are generally ascribed to a *surveyable proof*: (1) it can be written down and verified in practice, that is, by a human mathematician, and (2) it is a proof that one can have an intuitive understanding of.[12] Mawby [42] describes the second property as having the ability to "reach intuitive

[10] Indeed, it is generally not decidable whether a finite axiomatic theory is consistent, and so $\mathrm{Ax}_{T'}(x)$ does not define a recursive property on natural numbers.

[11] See e.g. [49] and [69].

[12] The notion of a surveyable proof is often credited to Tymoczko [62]. It should be noted that this notion is a controversial one and that different interpretations of it can be found in the literature.

understanding of why the proof must come as it does, and thus assert that the proof will always produce this result if correctly carried out" (p. 48).

Following these characterisations of a surveyable proof, a *surveyable presentation* of a theory T can be defined as one that enables us to acertain what the axioms of T are, without appealing to conditions that are not decidable nor determined by the presentation itself. We can then formulate the following proof-generated concept of natural theory based on the third type of counterexamples:

Proof-generated concept 2: *A natural axiomatic theory is a theory that is presented in a surveyable manner.*

The proof-generated concept PC2 characterises the cautiously enumerated theories as unnatural and thus dismisses them as counterexamples to the linearity conjecture. It particular, it will dismiss any consistency-minded theory as unnatural, assuming that the consistency condition employed is not trivial. Moreover, like PC1, PC2 is of a static rather than a dynamic nature: whether a theory is presented in a surveyable manner does not depend on its contextual properties.

Whether PC2 is similar to the intuitive notion of a natural theory is bit of a subtle matter; as the intuitive notion of a natural theory seems to be based on a purely extensional view of axiomatic theories, it does not seem to involve an idea of how such a theory is to be presented. However, I would agrue that the mere fact that theories tend to be treated as extensional entities can be viewed as justifying a restriction to surveyable presentations. By leaving the presentation of theories out of consideration when dealing with relative consistency proofs, mathematicians seem to be assuming that we have such a clear understanding of what the axioms of the theories in question are, that there is no need to make the intended presentation explicit. They seem to be assuming that our understanding of the theory simply 'gives' us its full extension, just like a finite list of axioms can give us the full extension of a finite theory, and it is precisely such a complete grasp of the theory's extension that a surveyable presentation of a theory is meant to provide.

4.5 Taking stock

Together, these proof-generated concepts provide us with the following tentative definition of a natural theory:

An axiomatic theory is *natural* if its axioms do not carry meta-information and if the theory is presented in a surveyable manner.

This definition employs the informal notions of 'carrying meta-information' and 'surveyability', which I have attempted to make at least intuitively clear in this chapter. How to make these notions more precise is not staightforward; in particular, the intensional character of the notion of carrying meta-information seems to complicate a translation of this informal notion into formal terms. Nevertheless, this definition can be viewed as an improvement of the status quo, as it provides us with more robust characteristics of natural theories than those found in the literature so far.

Concluding remarks

The starting point of this thesis has been the often made claim in the mathematical literature that the natural axiomatic theories are linearly ordered in terms of consistency strength. Despite the fact that we lack a precise definition of a natural theory, we have seen that the general attitude of the mathematical community towards this claim suggests that it is *true*, or at least capable of being true or false. In particular, our literature review suggests that there is a strong conviction that the well-known instances of nonlinearity, namely those obtained using the Rosser construction, involve theories that are *unnatural*. Nevertheless, current explanations of what is meant by a natural theory in this context go no further than 'a theory that arises in practice' or 'a theory that has a genuinely mathematical idea to it'.

In this thesis, I have tried to find a sharpened definition of the notion of a natural theory as referred to in the linearity conjecture. As the intuitive notion of a natural theory is still in a very premature stage, also referred to as the *pretheoretic* stage by Smith [56], this notion does not seem suitable for a conceptual analysis leading to the formulation of axioms that this intuitive notion is to satisfy. I therefore chose to formulate a method based on the Lakatosian view on mathematical concept-formation, which entails that mathematical concepts in practice develop through a dynamic interplay between conjectures, proofs and counterexamples. The method consists of two main steps: (1) formulate a proof-generated concept of natural theory based on an analysis of the known instances of nonlinearity, and (2) assess whether the proof-generated concept is similar to the intuitive notion. My approach can therefore be viewed as a 'mathematics first' approach: rather than sharpening the notion of a natural theory by appealing to philosophical considerations, I attempted to 'let the mathematics speak for itself' and to take philosophical considerations into account only after the formulation of the proof-generated concept.

By applying the method, I have obtained the following proof-generated concept of natural theory: an axiomatic theory is *natural* if its axioms do not carry meta-information and if it is presented in a surveyable manner. I have argued that this concept is similar to the intuitive notion of a natural theory. The traditional or standard view on axiomatic theories seems to be that axioms of a mathematical theory state properties of some particular object or structure, which can then be used to derive further results. As such, the metatheoretic properties of a mathematical theory should follow from, but not be stipulated by, its axioms. With respect to the surveyable presentation, I have argued that this property harmonizes well with the fact that we commonly treat our theories as extensional entities.

In the Lakatosian view, the main thrust of this sharpened definition of a natural theory is to open up new places for criticism on the linearity conjecture. By providing more robust characteristics, this definition makes demands that new counterexamples to the linearity conjecture should satisfy. When such a new counterexample is found, the method of proofs and refutations continues and a new proof-generated concept is to be formulated. The emergence of such a new

counterexample is possible, and perhaps even to be expected. Recall that the generalized result formulated in Theorem 3.18 shows that Hamkins has discovered a large class of incomparable theories: one finds instances of incomparability whenever one manages to find strongly independent sentences of a certain low arithmetical complexity. A promising candidate for a new counterexample that would satisfy our definition of naturalness would therefore be a construction of such sentences that does not use the method of arithmetization to encode meta-information. The situation reminds of that right after Gödel published his first incompleteness theorem: after the recognition that any nice theory could not prove or refute its own Gödel sentence, the question arose whether it was possible to obtain undecidable sentences without using the arithmetization of syntax, in particular for Peano arithmetic. Eventually, such sentences were found, in the form of the Paris-Harrington sentence and the Kirby-Paris sentence.[13]

Apart from opening up new places for counterexamples, my analysis provides an answer against scepticism towards the notion of a natural theory as expressed by Hamkins [22]. When calling a theory unnatural, one can appeal to more robust properties of the theory in question than one's mere dislike for or unfamiliarity with it. In particular, one can appeal to *inherent* properties rather than mere contextual ones. As such, I hope to have shown that the concept of natural theory in the context of the linearity conjecture need not be empty or futile.

Lastly, my analysis suggests a strong link between naturalness and intensionality. According to my tentative definition, what deems a theory natural or unnatural depends on the way we interpret and present its axioms. Such intensional aspects are difficult to reconcile with the common formal treatment of theories as extensional entities whose axioms are mere syntactic objects. Nevertheless, as the instances of nonlinearity in Chapter 3 exemplify, intensional aspects play a crucial role in our construction of theories. Therefore, in order to understand the linearity phenomenon, and in particular the common dismissal of instances of nonlinearity as unnatural, it seems that intensional aspects should not be ignored. To use the words of Halbach and Visser:[14]

As so often, philosophical notions defying a full formal analysis function as an engine driving progress in logic and, more generally, in mathematics and the sciences. Therefore, they shouldn't be dismissed, even if they prove somewhat elusive.

[13] See [1] for a proof of the independence of these sentences from PA.
[14] Halbach and Visser [21], p. 705.

Bibliography

[1] Rod J. L. Adams and Roman Murawski. *Recursive functions and metamathematics: problems of completeness and decidability, Gödel's theorems*, volume 286. Springer Science & Business Media, 1999.

[2] David D. Auerbach. Intensionality and the Gödel theorems. *Philosophical studies*, 48(3):337–351, 1985.

[3] David D. Auerbach. How to say things with formalisms. In *Proof, Logic and Formalization*, pages 86–102. Routledge, 1992.

[4] Jeremy Avigad. Reliability of mathematical inference. *Synthese*, 198(8):7377–7399, 2021.

[5] Robert Black. Proving church's thesis. *Philosophia Mathematica*, 8(3):244–258, 2000.

[6] Andrés E. Caicedo. (non?)-linearity of the consistency strength ordering in zf. `https://mathoverflow.net/questions/59717/non-linearity-of-the-consistency-strength-ordering-in-zf`, 2011.

[7] Rudolf Carnap. *Logical foundations of probability*. University of Chicago Press, Chicago, 2nd ed. edition, 1963.

[8] Alonzo Church. An unsolvable problem of elementary number theory. *American journal of mathematics*, 58(2):345–363, 1936.

[9] Paul J. Cohen. The independence of the continuum hypothesis. *Proceedings of the National Academy of Sciences*, 50(6):1143–1148, 1963.

[10] Michael Detlefsen. On an alleged refutation of Hilbert's program using Gödel's first incompleteness theorem. *Journal of Philosophical Logic*, pages 343–377, 1990.

[11] Solomon Feferman. Arithmetization of metamathematics in a general setting. *Fundamenta mathematicae*, 49(1):35–92, 1960.

[12] José Ferreirós. On arbitrary sets and ZFC. *Bulletin of Symbolic Logic*, 17(3):361–393, 2011.

[13] Janet Folina. Church's thesis: Prelude to a proof. *Philosophia Mathematica*, 6(3):302–323, 1998.

[14] Sy-David Friedman, Michael Rathjen, and Andreas Weiermann. Slow consistency. *Annals of Pure and Applied Logic*, 164(3):382–393, 2013.

[15] Robin Gandy. The confluence of ideas in 1936. In *A half-century survey on The Universal Turing Machine*, pages 55–111. Oxford University Press, Inc., 1988.

[16] Kurt Gödel. The consistency of the axiom of choice and of the generalized continuum-hypothesis. *Proceedings of the National Academy of Sciences*, 24(12):556–557, 1938.

[17] Kurt Gödel. What is Cantor's continuum problem? *The American Mathematical Monthly*, 54(9):515–525, 1947.

[18] Kurt Gödel. Remarks before the princeton bicentennial conference on problems in mathematics. In S. Feferman et al., editor, *Kurt Gödel, Collected Works*. Oxford University Press, 1990.

[19] Petr Hájek and Pavel Pudlák. *Metamathematics of first-order arithmetic*, volume 3. Cambridge University Press, 2017.

[20] Volker Halbach and Albert Visser. Self-reference in arithmetic i. *The Review of Symbolic Logic*, 7(4):671–691, 2014.

[21] Volker Halbach and Albert Visser. Self-reference in arithmetic ii. *The Review of Symbolic Logic*, 7(4):692–712, 2014.

[22] Joel D. Hamkins. Nonlinearity in the hierarchy of large cardinal consistency strength, 2021. Retrieved from `http://jdh.hamkins.org/wp-content/uploads/linearity-3.pdf`.

[23] Leon Henkin. A problem concerning provability. *Journal of Symbolic Logic*, 17(2):160, 1952.

[24] Leon Henkin. Review: G. kreisel, on a problem of henkin's. *The Journal of symbolic logic*, 19(iss. 3):219–220, 1954.

[25] David Hilbert, Paul Bernays, and Hermann Weyl. *Grundlagen der mathematik*, volume 2. Springer Berlin, 1939.

[26] Luca Incurvati. *Conceptions of Set and the Foundations of Mathematics*. Cambridge University Press, 2020.

[27] Daniel Isaacson. Arithmetical truth and hidden higher-order concepts. In *Studies in Logic and the Foundations of Mathematics*, volume 122, pages 147–169. Elsevier, 1987.

[28] Thomas J. Jech. *Set theory*. Springer monographs in mathematics. Springer, the 3rd millennium ed., rev. and expanded edition, 2002.

[29] Robert G. Jeroslow. Redundancies in the Hilbert-Bernays derivability conditions for Gödel's second incompleteness theorem. *The Journal of Symbolic Logic*, 38(3):359–367, 1973.

[30] Akihiro Kanamori. *The higher infinite: large cardinals in set theory from their beginnings*. Springer Science & Business Media, 2008.

[31] Stephen C. Kleene. *Introduction to metamathematics*. Princeton, NJ, USA: North Holland, 1952.

[32] Peter Koellner. Independence and Large Cardinals. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2011 edition, 2011.

[33] Andrei N. Kolmogorov and Vladimir A. Uspenskii. On the definition of an algorithm. *Uspekhi Matematicheskikh Nauk*, 13(4):3–28, 1958.

[34] G. Kreisel. On a problem of Henkin's. *Indagationes Mathematicae (Proceedings)*, 56:405–406, 1953.

[35] Georg Kreisel. Informal rigour and completeness proofs. In *Studies in Logic and the Foundations of Mathematics*, volume 47, pages 138–186. Elsevier, 1967.

[36] Kenneth Kunen. Set theory, volume 34 of studies in logic, 2011.

[37] Imre Lakatos. *Mathematics, Science and Epistemology: Volume 2, Philosophical Papers*, volume 2. Cambridge University Press, 1980.

[38] Imre Lakatos. *Proofs and refutations: The logic of mathematical discovery*. Cambridge university press, 2015.

[39] Per Lindström. *Aspects of incompleteness*, volume 10. Cambridge University Press, 2017.

[40] Martin H. Löb. Solution of a problem of Leon Henkin. *The Journal of Symbolic Logic*, 20(2):115–118, 1955.

[41] Penelope Maddy. *Naturalism in mathematics*. Clarendon Press, 1997.

[42] Jim Mawby. *Strict finitism as a foundation for mathematics*. PhD thesis, University of Glasgow, 2005.

[43] Elliott Mendelson. Second thoughts about church's thesis and mathematical proofs. *The Journal of Philosophy*, 87(5):225–233, 1990.

[44] Antonio Montalbán. Martin's conjecture: a classification of the naturally occurring turing degrees. *Notices of the American Mathematical Society*, 66(8):1209–1215, 2019.

[45] Antonio Montalbán and James Walsh. On the inevitability of the consistency operator. *The Journal of Symbolic Logic*, 84(1):205–225, 2019.

[46] Andrzej Mostowski. Models of axiomatic systems. *Fundamenta mathematicae*, 1(39):133–158, 1952.

[47] Lavinia Picollo. Reference in arithmetic. *The Review of Symbolic Logic*, 11(3):573–603, 2018.

[48] Paula Quinon. Can church's thesis be viewed as a Carnapian explication? *Synthese*, 198(5):1047–1074, 2021.

[49] Panu Raatikainen. Hilbert's program revisited. *Synthese*, 137(1):157–177, 2003.

[50] Luca San Mauro and Giorgio Venturi. Naturalness in mathematics. In *From Logic to Practice*, pages 277–313. Springer, 2015.

[51] Dirk Schlimm. Axioms in mathematical practice. *Philosophia Mathematica*, 21(1):37–92, 2013.

[52] Stewart Shapiro. Proving things about the informal. In *Turing's Revolution*, pages 283–296. Springer, 2015.

[53] Wilfried Sieg. Calculations by man and machine: Conceptual analysis. *Reflections on the Foundations of Mathematics (Essays in Honor of Solomon Feferman)*, 15:387–406, 2001.

[54] Stephen G. Simpson. The Gödel hierarchy and reverse mathematics. In Solomon Feferman, Charles Parsons, and Stephen G. Simpson, editors, *Kurt Gödel, Essays for his Centennial*. Cambridge University Press, 2010.

[55] Peter Smith. Squeezing arguments. *Analysis*, 71(1):22–30, 2011.

[56] Peter Smith. *An introduction to Gödel's theorems*. Cambridge University Press, 2013.

[57] Robert I. Soare. *Recursively enumerable sets and degrees: A study of computable functions and computably generated sets*. Springer Science & Business Media, 1999.

[58] John R. Steel. Gödel's program. In Juliette Kennedy, editor, *Interpreting Gödel: Critical Essays*, chapter 8. Cambridge University Press, 2014.

[59] William W. Tait. Finitism. *The Journal of Philosophy*, 78(9):524–546, 1981.

[60] Jamie Tappenden. Mathematical concepts: Fruitfulness and naturalness. In Paolo Mancosu, editor, *The Philosophy of Mathematical Practice*, pages 276–301. Oxford University Press, 2008.

[61] Alan Turing. On computable numbers, with an application to the entscheidungsproblem. *Proceedings of the London Mathematical Society*, 42(1):230–265, 1936.

[62] Thomas Tymoczko. The four-color problem and its philosophical significance. *The journal of philosophy*, 76(2):57–83, 1979.

[63] Giorgio Venturi. On the naturalness of new axioms in set theory. *preprint*, 2016.

[64] Albert Visser. Peano's smart children: a provability logical study of systems with built-in consistency. *Notre Dame Journal of Formal Logic*, 30(2):161–196, 1989.

[65] James Walsh. On the hierarchy of natural theories. *arXiv preprint arXiv:2106.05794*, 2021.

[66] James Walsh. A robust proof-theoretic well-ordering. *arXiv preprint arXiv:2201.05284*, 2022.

[67] Hao Wang. *A logical journey: From Gödel to philosophy*. MIT Press, 1997.

[68] Timothy Williamson. *Vagueness*. The problems of philosophy. Routledge, London, 1994.

[69] Richard Zach. Hilbert's program then and now. In *Philosophy of logic*, pages 411–447. Elsevier, 2007.