View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Apollo

Axiomatization and Incompleteness in Arithmetic and Set Theory

Wesley Duncan Wrigley
Sidney Sussex College, University of Cambridge

26th April 2019
This thesis is submitted for the degree of Doctor of Philosophy

Preface

Declaration

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text.

It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my thesis has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text

It does not exceed the prescribed word limit for the relevant Degree Committee.

Acknowledgements

Foremost thanks are due to my supervisors, Tim Button and Michael Potter. Their unending patience and support, together with many profound insights and difficult questions, were essential in bringing this thesis to fruition. I've taken the decision to cite them both only when discussing an idea which is wholly their own. If justice were truly done, they would both be cited on every page.

I'm very grateful to Owen Griffiths and Rob Trueman. Their helpful comments and suggestions have greatly improved the clarity of this work, and their advice and encouragement has made producing it all the more enjoyable.

Thanks are also due to my examiners, Dan Isaacson and Alex Oliver, for their helpful suggestions on how to improve the work, and for an enjoyable and thought-provoking viva.

I'd also like to extend my gratitude to my parents and my sister, for enthusias-

tically supporting my education for the last twenty-three years, and to Angie, for always offering love and support. Without them all, I wouldn't have made it this far.

Lastly, I'd like to thank the AHRC, for generously funding my doctoral research.

Abstract

Axiomatization and Incompleteness in Arithmetic and Set Theory
Wesley Duncan Wrigley

I argue that are (at least) two distinct kinds of mathematical incompleteness. Part A of the thesis discusses *Gödelian* incompleteness, while Part B is concerned with *set-theoretic* incompleteness. Both parts are concerned with the philosophical justification of reflection principles and other axiomatic devices which can be used to reduce incompleteness, and in particular with the justification of such devices from the philosophical standpoint of Kurt Gödel.

In **Part A** I consider Gödel's disjunctive argument. In chapter 1, I argue that the non-mechanical mind considered by Gödel is best modelled by a theory constructed using the transfinite iterated application of a soundness reflection principle to **PA**. I argue that Feferman's completeness theorem shows this account of the mind to be incompatible with some elementary assumptions in the epistemology of arithmetic. In chapter 2, these considerations are developed into a positive argument for the existence of absolutely undecidable arithmetical propositions. The consequences for the indefinite extensibility of the concept *natural number* are then discussed. I argue that properly understood, Feferman's theorem refutes Dummett's position in the debate.

I begin part **Part B** in chapter 3, by reconstructing a version of Gödel's platonism, called *conceptual platonism*. I then examine how such a position relates to various means of reducing set-theoretic incompleteness. In chapter 4 I argue that there is some prospect for this position of effecting a limited reduction in incompleteness by means of reflection principles justified by mathematical intuition. However, such priniciples are incompatible with Gödel's commitment to platonism about properties of properties of sets. In chapter 5 I argue that conceptual platonism does not lend support to the view that a substantial reduction in incompleteness can be effected by large cardinal axioms justified using extrinsic methods analogous to the principles of theory choice in natural science. This undercuts the traditional justification for many large cardinal axioms, so I end with a sketch of how conceptual

platonism could be modified to rehabilitate the large cardinals program.

Contents

Introduction **xi**

Kurt Gödel as a Philosopher . . . xi

The Significance of Incompleteness . . . xiii

Part A: Gödelian Incompleteness . . . xiv

Part B: Set-Theoretic Incompleteness . . . xv

A Gödelian Incompleteness 1

1 Minds, Machines, and Reflection Principles **3**

Introduction . . . 3

1.1 Lucas' Dialectical Argument . . . 4

1.2 The Non-Mechanical Mind . . . 6

1.3 Feferman Reflection . . . 9

1.4 Feferman's Completeness Theorem . . . 12

1.5 Enumerating $\mathcal{O}$ . . . 16

1.6 The Failure of Autonomy . . . 21

1.7 Modest Anti-Mechanism . . . 26

Conclusion . . . 28

2 Absolutely Undecidable Arithmetical Propositions **31**

Introduction . . . 31

2.1 The Evidence Argument . . . 32

2.2 Intensionality . . . 34

2.3 Ordinal Notations . . . 38

2.4 The Irrationality Argument . . . 40

2.5 Gödel's Disjunction Revisited . . . 43

2.6 Recursive Ordinal Selection . . . 45

2.7 Which Propositions are Absolutely Undecidable? . . . 51

2.8 Indefinite Extensibility . . . 53

2.9 Infinite Extensibility 56
2.10 Dummett on Feferman's Theorem 58
2.11 Responses on Dummett's Behalf 61
Conclusion 64

B Set-Theoretic Incompleteness 67

3 Conceptual Platonism 69
Introduction 69
3.1 Two Kinds of Platonism 70
3.2 Mathematical Intuition 73
3.3 Gödel, Anselm, and Hilbert 79
3.4 Mathematical Perception 86
Conclusion 89

4 Intuition and Reflection Principles 91
Introduction 91
4.1 Intuition in Set Theory 92
4.2 Reflection Principles 96
4.3 The Limits of Reflection 106
4.4 Platonism and Incompleteness 122
Conclusion 133

5 Quasi-Scientific Methods of Justification in Set Theory 137
Introduction 137
5.1 The Material Bodies Analogy 139
5.2 Gödel and the Regressive Method 144
5.3 Mathematical Data 148
5.4 The Laws of Nature Analogy 160
5.5 Theoretical Virtues 168
Conclusion 183

Concluding Remarks 187

Ineliminable Incompleteness 187
The Large Cardinals Programme 191

Bibliography **195**

Introduction

Kurt Gödel as a Philosopher

In 1931, Gödel published the incompleteness theorems, two results which are now widely acknowledged as amongst the most profound in twentieth century mathematical logic. Moreover, the significance of these results has not been confined to mathematics; they have also become the subject of a number of philosophical controversies. A dizzying variety of implications have been attributed to the theorems, not least by Gödel himself. These implications concern the nature of the human mind, mathematical proof, the concept *natural number*, the existence of mathematical objects, and the shape of the set-theoretic universe.

This thesis is about all of these issues, but perhaps most of all it is about reading Kurt Gödel as a philosopher. Gödel was certainly a cautious writer of philosophy, who was keenly aware that his philosophical views were not mainstream during his lifetime. Indeed, he saw the philosophical landscape of his career as largely dominated by empiricists, sceptics, and conventionalists (1961/?, p.375); by contrast his own views are characterized by a deep-seated belief that rational reflection by the human mind can yield genuine knowledge of mathematical objects 'despite their remoteness from sense experience' (1964, p.268).

The unfortunate result is that what has survived of Gödel's philosophical writings at times comes to little more than fascinating snippets of information from somewhat unconventional sources. For instance, it is not unusual in secondary literature about Gödel to encounter citations of his views which are sourced from conversations with Hao Wang, published decades after Gödel's death (the present work is of no exception in this regard). Much of the more formally published material nevertheless consists of drafts or lecture notes, which are brief or rough. Coupled with Gödel's almost unfailing caution, it is easy to despair of the lack of detailed exposition of, or clear argument for, some very interesting views that can be found in his corpus.

That said, clear themes and positions run throughout Gödel's philosophical work, and in this thesis four of those will be explored with particular reference to the ax-

iomatization of arithmetic and set theory, with a view to reducing the incompleteness of those theories. Indeed my central question will be: how can we strengthen our axiomatic mathematical theories so as to reduce the degree to which they are incomplete? In answering this question, the Gödelian themes with which I'll be particularly concerned are *anti-mechanism* about the human mind, *rationalistic optimism* about mathematical proof, *mathematical platonism*, and Gödel's analogy drawn between the methods of mathematics and the natural sciences. These all play a central role in Gödel's account of mathematical objects and mathematical knowledge, though the role they play is often less than transparent.

This thesis, in grappling with these aspects of Gödel's thought, is at its heart a philosophical endeavour. I want to know whether there is a compelling argument for anti-mechanism grounded in the incompleteness theorems; I want to know whether any arithmetical propositions are absolutely undecidable; I want to know whether large cardinal axioms are implied by the concept *set*, or can be inferred from some analogy between mathematics and the natural sciences. Technicalities and mathematical results will play a crucial role in some of my arguments, but no proofs will be found in these pages. Similarly, though much of the thesis involves close engagement with text, it is not primarily a historical endeavour. There is no archival work present, for example, and getting clear on what Gödel himself thought about these issues is a secondary concern for the most part. The reason for this is simply that I do not believe there to be enough evidence to make confident judgements about the nuances of Gödel's thought at several key junctures.

The result is that at times I will not be able to ascribe views directly to Gödel, and will only be able to describe them as 'Gödelian'. For example, Gödel himself never claimed to have an argument which refuted mechanism in the philosophy of mind; but in chapter 1 I'll examine a Gödelian argument which promises exactly this. Such arguments draw out the implications of core commitments in Gödel's thought, but with some of the details filled in and explicit arguments provided. While I do my best to motivate a textual basis for such arguments, the reader is asked to bear in mind that 'the Gödelian' does not always offer what we find in the actual work of Gödel. Indeed, the Gödelian often proposes arguments that I have been sorely tempted by, but cannot quite seem to make work to my satisfaction. It would be gratifying indeed if the reader has also been tempted by such arguments,

though I believe the views discussed are of interest in their own right.

The Significance of Incompleteness

The position that I'll argue for is, in brief, that there are (at least) two distinct kinds of incompleteness, namely *Gödelian* incompleteness and *set-theoretic* incompleteness. In part A, I'll argue that Gödelian incompleteness is *ineliminable* because there are very good reasons to think some arithmetical propositions are absolutely undecidable. The situation is less clear with respect to set-theoretic incompleteness, though I will argue in part B that the two means of reducing the degree to which set theory is incomplete offered by Gödel, namely the use of mathematical intuition and quasi-scientific methods, both fail to effect a significant reduction in incompleteness. A consequence is that much of the large cardinals programme in set theory is left without adequate philosophical support. In the course of arguing for these claims, a motley of other topics will be discussed in varying levels of detail: anti-mechanism, indefinite extensibility, platonism, mathematical intuition, the concept *set*, second-order logic, mathematical data, large cardinal axioms, and theoretical virtue.

What unites these seemingly disparate topics is that they each bear on claims made by Gödel in the extraordinarily rich paper *Some basic theorems on the foundations of mathematics and their implications*, Gödel's 1951 Gibbs Lecture. This paper will not be the primary focus of each chapter (for instance, the study of mathematical intuition draws heavily on the 1964 version of *What is Cantor's Continuum Problem?*); but in the Gibbs lecture Gödel proposes his famous disjunctive argument, argues that the incompleteness theorems support platonism, and that such realism forms the basis of importing quasi-scientific methods into the foundations of mathematics. Each chapter does, therefore, address some aspect of the Gibbs lecture, and attempting a comprehensive evaluation of it on my part was the genesis of the entire thesis.

Part A: Gödelian Incompleteness

Chapter 1: Minds, Machines, and Reflection Principles

The first two chapters of this thesis address Gödel's *disjunctive* argument: the incompleteness theorems imply that no *Turing Machine*, the mathematical model behind the digital computer, can be used to prove all arithmetical truths. Hence those theorems imply that either human beings equally cannot prove all arithmetical truths, or that human means of producing arithmetical proofs cannot be modelled by a Turing machine. Gödel's own view was *anti-mechanism*; although the argument in the Gibbs lecture officially goes no further than the disjunction, he thought that any arithmetical truth was provable in principle, and hence that no Turing machine could model the mathematical capacities of the human mind. Similar views have been expressed both by philosophers such as J.R. Lucas, and by working scientists and mathematicians such as Roger Penrose.

I argue that this anti-mechanist project is doomed to failure. Firstly, I clarify the notion of a non-mechanical mind at work in Gödel's argument, and show that the conception of mind underlying it is best modelled by a theory constructed from the transfinite iterated application of a certain reflection principle to Peano Arithmetic. In the second part of the chapter, I argue that no human being, in respect of their mathematical abilities, could plausibly be regarded as the kind of non-machine that the Gödelian argument employs. This is because the anti-mechanical model of our arithmetical capacities has unacceptable implications for the epistemology of arithmetic. The first attempt at reducing incompleteness therefore achieves only a modest degree of success: our ability to iteratively apply reflection principles to arithmetic is limited, because the unlimited ability implies an untenable form of anti-mechanism about the mind.

Chapter 2: Absolutely Undecidable Arithmetical Propositions

In the second chapter, I develop these anti-anti-mechanist considerations into an argument for the existence of true arithmetical propositions which have no humanly recongnizable proof. Gödel offers a number of cryptic arguments against this idea, which I reconstruct and criticize. The *evidence argument* claims that, since the po-

tentially unprovable propositions code certain information about arithmetical theories, they are exactly as evident as the axioms from which they are constructed. I argue, however, that Gödel's argument is valid only in special cases, not in general.

Secondly, the *irrationality* argument claims that human reason would be, in some sense, irrational or inconsistent if there were absolutely undecidable propositions. I explain why, if such unprovable truths exist, no recognizable example of one can be exhibited. Hence we can never be in the kind of irrational scenario envisaged by Gödel.

I go on to detail exactly what ability we would need to possess in order to vindicate Gödel's view, which I call the *Recursive Ordinal Selection Ability*. I argue that we have no good reason to believe we possess this ability, or anything equivalent to it, and every reason to suppose that we do not.

I finish this chapter by showing how the mathematical apparatus of reflection principles previously examined can be deployed to refute Dummett's claim that the concept *natural number* is vague. In particular, I argue that Dummett's position is not stable even in the context of his constructivism.

Part B: Set-Theoretic Incompleteness

Chapter 3: Conceptual Platonism

One aspect of Gödel's thought which plays little direct role in the Gibbs lecture is *mathematical intuition*. Given the centrality of the notion in Gödel's later philosophy, and in order to better understand Gödel's claims that the incompleteness theorems support platonism, I undertake the interpretative task of clarifying Gödel's view. I argue that there is no one unified position adopted by Gödel across the material available to us, but that a sensible Gödelian position can be reconstructed. I call this view *conceptual platonism*.

According to the conceptual platonist, the existence of mathematical objects is a consequence of facts about the concepts under which they fall. In certain cases, such as that of the concept *set*, reflecting on the features of the concept can give us non-deductive and non-empirical intuitive knowledge of the truth of sentences

which axiomatize that concept.

I make the textual case that this position is representative of Gödel's views, and defend it against charges of 'mysticism' or 'theology'. However, I highlight one respect in which the account is deficient as regards our mathematical knowledge; namely it is deeply unclear under what conditions a concept has *objective content* (i.e. is such that some objects must fall under it). Although Gödel gives us some hints, there is no remotely precise procedure for distinguishing concepts having this type of content from those that lack it. Despite that, I conclude that conceptual platonism is clear enough that we can address the question of whether reflection on the concept *set* can effect a substantial reduction in incompleteness, according to a platonist of this kind.

Chapter 4: Intuition and Reflection Principles

Gödel proposes two distinct means by which we can reduce the degree to which set theory is incomplete. The first such method is by the use of mathematical intuition. I examine Gödel's (still popular) idea that reflection on the concept *set* allows us to gain non-deductive knowledge of certain set-theoretic principles. Such knowledge might allow us to extend the axioms of set theory, and hence reduce their incompleteness. I argue that the conceptual platonist might be able to regard **ZFC** extended by stronger versions of the *reflection principle* as known intuitively, if they are willing to pay a hefty philosophical price as regards the extent of their platonism. However, using recent mathematical results in the area, I support Koellner's claim that principles of this kind (that might be justified by reflecting on the concept *set*) *cannot* strengthen the axiomatization of set theory in such a way as to significantly reduce its incompleteness, in a sense that can be made quite precise.

I then turn to Gödel's claim in the Gibbs lecture that the incompleteness theorems support platonism, which I understand to be conceptual platonism. I examine three arguments offered by Gödel, plus an additional argument tailored to the specific features of conceptual platonism. I argue that in each case, the incompleteness theorems does lend platonism support, but that the degree to which this holds is extremely limited. In essence, the incompleteness theorems serve to refute anti-platonist positions which involve commitments not shared by platonism's most

prominent rivals, including intuitionism and formalism.

Chapter 5: Quasi-Scientific Methods of Justification in Set Theory

In the final chapter, I examine the justification of large cardinal axioms going beyond what could possibly be regarded as justified intuitively. Axioms positing very large cardinals can be used to substantially reduce the incompleteness of set theory, but the justification of such axioms is a persistent problem. Gödel thought that they were justified by his platonist interpretation of mathematics: if the existence of sets is independent of us, then, according to Gödel, certain methods for the justification of positing unobservable entities in the natural sciences can be imported into set theory and used to justify large cardinal axioms.

Gödel's thought here draws heavily on Russell's *regressive method*, which I present in order to clarify two different analogies offered by Gödel between mathematics and the sciences. According to the first analogy, mathematical objects play a role in making sense of mathematical experience similar to the role played by material bodies in making sense of our perceptual experiences. I argue that such an analogy could not be used to justify large cardinal axioms, by highlighting the peculiar sense in which positing material bodies allows us to make sense of our experiences generally.

According to the second analogy, large cardinal axioms are analogous to laws of nature in physics, in that they receive *a posteriori* verification by accounting for data and enhancing theoretical virtue. I argue that only a very limited class of mathematical propositions could be regarded as data, and demonstrate that on such a conception large cardinal axioms do not permit the deduction of more data than can be obtained by much weaker hypotheses, namely the consistency of the axioms in question relative to **ZFC**.

Lastly, I discuss Gödel's proposal that large cardinal axioms can be justified when they endow set theory with substantial theoretical virtues, using criteria analogous to those employed in science. I argue that the methods of theory choice in the sciences typically involve some *minimizing* ontological principle (e.g. Occam's razor), and a principle of theoretical economy. Since large cardinal axioms automatically bloat a theory's ontology and ideology, they can be expected to perform poorly when

assessed relative to such standards. I conclude that the analogy between science and mathematics has little to offer in terms of justifying particular large cardinal axioms or the large cardinals programme more generally.

Concluding Remarks

To whatever extent the methods of theory choice from the natural sciences can be applied in set theory, the justification for the large cardinals programme is weakened. Hence an entirely different perspective on set theory is required to vindicate the programme. I end the thesis by outlining a direction for future research addressing this topic, and suggest that a version of the multiverse interpretation of set theory, consonant with remarks made by Gödel in the 1930s, might put the programme on firm philosophical footing. I outline how a Gödelian multiverse view might differ from the most prominent current multiverse view on offer, formulated by Hamkins.

Part A

Gödelian Incompleteness

Chapter 1

Minds, Machines, and Reflection Principles

Introduction

In his 1951 Gibbs Lecture delivered to the American Mathematical Society, Gödel claimed that the incompleteness theorems entail the following disjunction: either the human mind is not a machine, at least in respect of its ability to prove mathematical truths, or else there are number-theoretic problems which are in some sense *absolutely* unsolvable (1951, p.310).

In this first chapter, my aim is to present a problem for Gödelian anti-mechanism, which I take to be the view that the mathematical capabilities of the mind cannot be modelled by any Turing machine, *and* that this conclusion is established as an implication of the incompleteness theorems (as opposed to an argument from, e.g., neuroscientific considerations). The argument proceeds as follows:

In §1, I'll review the most well-known argument for Gödelian anti-mechanism, given by Lucas. I'll argue that, despite plenty of potential problems, Lucas' argument is valuable for revealing something of the structure that any argument for Gödelian anti-mechanism should employ.

In §2, I'll argue that this Gödelian conception of the non-mechanical mind requires a certain kind of mathematical model, built from a transfinite reflection sequence of a certain kind based on **PA** (the first-order Peano axioms). The salient technical details are presented in §3.

In §4, I'll introduce Feferman's completeness theorem, which serves two purposes. Firstly, it vindicates modelling the anti-mechanist's view by a transfinite reflection sequence. Secondly, as I'll argue in §5, it highlights exactly which abilities the Gödelian anti-mechanist must suppose are possessed by the idealised mathematician. I'll argue that the need to posit such abilities puts a substantial explanatory burden on the Gödelian, which at present we have no reason to believe can be offloaded.

In §6, I'll present some further mathematical properties of reflection sequences based on **PA** and use them to argue that the anti-mechanist must suppose that the idealised mathematician has inexplicable access to certain arithmetical truths, making Gödelian anti-mechanism an unworkable model of our arithmetical capacities.

Finally, in §7, I'll deploy Turing's completeness theorem to show that the arguments given apply also to substantially weaker forms of anti-mechanism than that to which Gödel was inclined.

1.1 Lucas' Dialectical Argument

In the Gibbs lecture, Gödel argues for his famous disjunction, that either the mathematical capabilities of the human mind do not correspond to any Turing machine, or that some mathematical problems are unsolvable in an absolute sense. He does not offer an argument for a particular disjunct. It is well known, however, that the disjunct which Gödel was inclined to accept was the anti-mechanical one (Wang 1974, pp.324-326). My central aim in this chapter is to reconstruct what I think is the most plausible version of a Gödelian argument for this view, and show that it fails. Given Gödel's characteristic caution on the matter, the reconstruction will have to start outside his own writings. The most well-known argument for the anti-mechanist disjunct of Gödel's disjunction is the Lucas–Penrose argument. In this section, I'll take a closer look at Lucas' anti-mechanist argument.[1] Lucas' argument is notable for being *dialectical* in form; rather than present a knock-down argument that minds are not machines, Lucas offers an argument schema to refute the mech-

[1] I'll set aside Penrose's version, largely because the aims of his argument, insofar as it differs from Lucas', are orthogonal to our present concern. My aim is to address the question of whether the idealised arithmetical output of human mathematicians can be shown to be distinct from the output of any Turing machine. On the other hand, Penrose aims to establish that '[h]uman mathematicians are not using a knowably sound algorithm in order to ascertain mathematical truth', and to demonstrate 'something very significant about the mental quality of *understanding*' (Penrose 1994, p.76). The present discussion is somewhat removed from concerns about what *actual* human mathematicians are, or are not, doing. This is because the total output of all past and present human mathematicians is finite, and hence there is certainly a Turing machine which enumerates the Gödel numbers (under some suitable coding) of all the arithmetical truths that have been proved by us so far. Additionally, my concerns are purely extensional, do not speak to any kind of mental quality.

anist (1968, p.156).

The schema is as follows: the mechanist comes along, and puts forward a thesis of the form 'the human mind can be modelled by machine **M**', where a machine models a mind if and only if the arithmetical output of the two is coextensive under suitable idealisation. Given that some human beings are arithmetically proficient, we assume that **M** enumerates the Gödel numbers of the theorems of **PA**, and perhaps other things too. The output of any Turing machine codes the theorems of a recursively axiomatized theory, so we'll say that **M** can prove a sentence if and only if it is a theorem of the corresponding theory $\mathbf{T_M}$. Thanks to Gödel's theorem, there is some sentence $G_\mathbf{M}$ which, provided that $\mathbf{T_M}$ is consistent, is true and which **M** cannot prove. Lucas then takes up the potentially very tedious task of constructing $G_\mathbf{M}$, and proves it (or at least claims to). Lucas and the machine can both prove $Con_{\mathbf{T_M}} \rightarrow G_\mathbf{M}$, where $Con_{\mathbf{T_M}}$ is some canonical consistency sentence for $\mathbf{T_M}$ under the presentation of that theory coded by the output of **M**. Assuming that (the theory corresponding to) the machine is consistent, then by Gödel's second theorem, the machine can't prove $Con_{\mathbf{T_M}}$, and can get no further. But Lucas, 'standing outside the system', as he puts it, can prove that the Gödel sentence $G_\mathbf{M}$ is true, since it 'says' that it is not a theorem of $\mathbf{T_M}$, which indeed it isn't, by the assumption of the machine's consistency (Lucas 1961, p.117). Since Lucas can prove something that the machine cannot, the latter cannot model the mathematical capabilities of the former. The same goes for any other suitable machine, including the one corresponding to $\mathbf{T_M} + G_\mathbf{M}$. Whatever thesis the mechanist offers, Lucas claims that he can disprove it via the same technique (1961, p.117).

In the decades since it first appeared, Lucas' argument has gained a remarkable degree of notoriety, and attracted criticism from all quarters in the philosophy of mathematics. I won't go into details here, because for our purposes, we need not concern ourselves with the particulars, or even validity, of Lucas' argument. Rather, we are concerned with its *structure*.

Although I don't want to endorse Lucas' argument, I think it embodies a structure that should be common to Gödelian anti-mechanism of any type. A key feature of Gödel's anti-mechanism is that any arithmetical truth is provable in-principle; no incompleteness is ineliminable. Although his support for the view is only cautiously hinted at in the Gibbs lecture, it was later confirmed as Gödel's own view in (Wang

1974, p.324-326). This *rationalistic optimism*, narrowly speaking, is the view that any well-posed mathematical proposition can be proved or can be refuted. More broadly, it is the view that 'for clear questions posed by reason, reason can find clear answers' (Gödel 1961/?, p.381).

Hence the Gödelian must think that (subject to sufficient idealisation) the human mind has the ability to overcome any incompleteness it may encounter. Lucas thinks, rightly or wrongly, that he can do this by proving a Gödel sentence. Gödel himself might have said that rational intuition enables us to overcome the problem. But my point is that the particular anti-mechanist account doesn't matter for our purposes; what is important is that the anti-mechanist motivated by Gödel's theorem must think that we can iteratively overcome incompleteness where it is encountered, be that in conversation with a mechanist, or in the course of ordinary mathematics, or elsewhere entirely. We encounter a well-defined proposition that we are currently unable to prove or refute, and then find some means of overcoming it (assuming Gödel's rationalistic thesis). It is a little unclear *how much* idealisation is allowed in this debate; unless otherwise stated, I'll be operating under the standard assumption that the idealised mathematician has an arbitrarily large (though finite) stock of materials, time, and brain-power available for their reasoning.

The only alternative to an iterative view would be to think that the idealised mind can overcome all incompleteness in one fell swoop, but it is difficult to imagine how such a theory, whatever its merits or defects, could be motivated by the incompleteness theorems specifically.

So the 'dialectical' nature of the argument offered by Lucas should be a common element of any Gödelian account of the non-mechanical mind, whatever we might think of the details. In the next section, we'll investigate what iterative abilities a Gödelian anti-mechanist might ascribe to us to ensure that incompleteness can always be overcome, given a sufficient degree of idealisation.

1.2 The Non-Mechanical Mind

If we are to entertain the thought that the mind is some kind of non-machine in respect of its ability to prove theorems (under a high degree of idealisation), it is in-

cumbent upon us to ask: what kind of non-machine might the mind be? It would be too much for us to demand from the anti-mechanist a complete account of the workings of the human mind, but it is perfectly fair to ask for a sketch of how idealised human theorem-proving works, if how it works is rather unfamiliar. Of course, we do not simply want an extensionally adequate account of what is provable by such a mind; according to Gödel at least that would simply be $\{\phi \mid \phi \in L_{\mathbf{Q}} \wedge \phi \text{ is true}\}$, where $L_{\mathbf{Q}}$ is the first-order language of arithmetic. Rather, we need an outline of by what procedures such things might be provable, since they cannot be the familiar procedures of a Turing machine or a formal proof.

Gödel does not go into detail in the Gibbs lecture, but the 'dialectical' picture of a mind able to successively overcome incompleteness wherever it is found suggests an interesting model of idealised human theorem-proving. A *reflection principle* is a statement which can be iteratively added to a theory, and the validity of which follows from the soundness of the base theory to which the principle is added. For example, Lucas' argument starts with **PA** as a base theory, and at every step of the dialectic, Lucas is (allegedly) justified in appending a Gödel sentence to the theory of present concern. More generally, a Gödelian anti-mechanist can take the mind to be modelled by a theory as strong as **PA** to which some reflection principle has been iteratively applied.

A *reflection sequence* based upon a theory **T** is the result of the iterated addition of some reflection principle to **T**. In the case envisaged by Lucas, the sequence is based on **PA**, and is formed by the iterated addition of a canonical Gödel sentence to the theory at each ordinal successor stage in the sequence. We seem, therefore to have made progress in giving an account of what the proposed non-mechanical mind looks like with respect to its arithmetical abilities: the human mind can, under suitable idealisation, iteratively apply a reflection principle and prove things in the resultant theories.[2] A virtue of this account is that reflection sequences are mathematically tractable, and many results exist concerning them, beginning with Turing's paper on 'ordinal logics' (1939) and continuing in the work of Feferman

[2]As far as constructing a Gödelian account of anti-mechanism goes, there are perhaps other options that could be explored. But the application of a reflection principle fits well with the dialectical picture under considerations, and has other features to be explored below that make it suitable from a Gödelian perspective.

and others.

Unfortunately, the reflection principle implicit in Lucas' argument that governs the procedure by which the human arithmetically one-ups the machine is not strong enough to do its job. Lucas' reflection principle (for successor ordinals) is as follows:

Lucas Reflection: $\mathbf{T}_{\alpha+1} = \mathbf{T}_\alpha \cup \{G_{\mathbf{T}_\alpha}\}$

where $G_{\mathbf{T}_\alpha}$ is the standard Gödel sentence for $\mathbf{T}_\alpha$ under some standard axiomatization.[3] Lucas' proposed justification for his reflection principle is the pure confusion that the 'essence of the Gödelian formula is that it is self-referring' (1961, p.124). The thought is that Lucas, unlike the machine, can reflect on the nature of the constructed sentence, and hence deduce its truth.

However, as Putnam (1960, p.366) argues, the important point is that the theory to which the machine corresponds is *consistent*. If it isn't, then the machine can prove $G_{\mathbf{T}_\alpha}$ just as Lucas can. So if Lucas can claim to prove something that the machine can't, what he needs is in fact a means of determining $\mathbf{T}_\alpha$'s consistency. If the theory is consistent, then Lucas, but not the machine, can prove its Gödel sentence. So it is Lucas' (alleged) knowledge of the machine's consistency which does the work in distinguishing their arithmetical capabilities.

Rightly or wrongly, Lucas insists that the assumption of the consistency of the theory corresponding to the machine is in play throughout the course of the dialectic. So we might try the following reflection principle (for successor ordinals):

Consistency Reflection: $\mathbf{T}_{\alpha+1} = \mathbf{T}_\alpha \cup \{Con_{\mathbf{T}_\alpha}\}$

where $Con_{\mathbf{T}_\alpha}$ is a canonical consistency sentence for $\mathbf{T}_\alpha$. But again, this won't do, since even if Lucas can use a reflection principle of this kind, there is no guarantee that some machine offered by the mechanist in the course of the back-and-forth will be out-performed by him at all. This is because there is, in general, no reason to expect of a given consistent theory that the union of the theory together with its canonical consistency statement is consistent. The key reason for this is that certain consistent but unsound theories prove their own inconsistency. Suppose for instance,

[3] There's no need to rehearse the details here, but it should be remembered that these notions only make sense relative to a specified set of arithmetical axioms with a canonical means for constructing Gödel sentences. Your favourite textbook treatment will do the job.

that $\mathbf{T} = \mathbf{PA} + \neg Con_{\mathbf{PA}}$. The extension of $\mathbf{T}$ by its consistency sentence will prove that $\mathbf{T}$, and hence also $\mathbf{PA}$, is consistent. Yet it will also prove $\neg Con_{\mathbf{PA}}$, since that sentence is an axiom.

Clearly, if Lucas finds himself competing with a Turing machine the corresponding theory of which is inconsistent, he will at most be able to match the machine, and will never out-perform it. However, Lucas' remarks suggest that the assumption he intends to put in play is stronger than the consistency of the relevant machines, since he writes that an essential part of the game is that 'a rational being, standing outside the system can see that it [the Gödel sentence for the system in question] is true' (1961, p.117). While the consistency of a theory is of course sufficient for the truth of its Gödel sentence, the crucial assumption is that the machines offered by the mechanist only prove truths, not merely that they are consistent. Once again, I think this is at least structurally representative of the Gödelian position: at the heart of anti-mechanism is the conviction that we can *know* more than a machine could ever prove, not that we can simply write down more sentences. Given that knowledge implies truth, the model of the anti-mechanical mind should, I think, be taken from a reflection sequence constructed by a *soundness* principle.

1.3 Feferman Reflection

Roughly speaking, our target reflection principle asserts some soundness property of the members of the sequence preceding it. Franzén favours the formulation of a single sentence expressing the soundness of a theory in order to achieve this (2004, §14.1). For present purposes, this presents a quite unnecessary complication. Instead, we can take as our canonical reflection principle *Feferman reflection* (also known as Feferman's 'principle II'), which says that, for any formula ϕ, if, of every number, a theory proves that it is ϕ, then every number is ϕ. More formally:

Feferman Reflection: $\forall x\ Pr_{\mathbf{T}_\alpha}(\overline{\phi}(\dot{\overline{x}})) \rightarrow \forall x\ \phi(x)$

where $Pr_{\mathbf{T}_\alpha}$ is a provability predicate for $\mathbf{T}_\alpha$ coded in the standard Gödelian fashion; and where $\overline{\phi}(\dot{\overline{x}})$ denotes the Gödel number of the result of substituting the numeral denoting x for the first variable appearing in ϕ. Feferman reflection should

be understood without loss of generality as restricted to formulae of $L_{\mathbf{Q}}$ with a single variable (Feferman 1962, p.274). It should be noted that the arguments to follow carry over with a variety of alternative reflection principles (see (Feferman 1962, p.274) for details), but we'll stick to using Feferman reflection for the sake of simplicity, since the formal results that will later be of some importance are all stated rather elegantly with respect to it. The essential point is that what goes for the soundness principle discussed in this chapter goes for a variety of other alternatives. The present selection is merely for convenience.

Feferman reflection is a principle that we should accept of theories we believe to be sound with respect to domain $\mathbb{N}$. For if a theory is sound, its proof predicate actually represents the proof relation. So, if it proves ϕ of each number, it follows by minimal semantic reflection that $\forall x\ \phi(x)$. Hence, if we accept $\mathbf{T}_\alpha$ as sound, we should similarly accept as sound:

$$\mathbf{T}_{\alpha+1} = \mathbf{T}_\alpha \cup \{\forall x\ Pr_{\mathbf{T}_\alpha}(\overline{\phi}(\dot{\overline{x}})) \rightarrow \forall x\ \phi(x) |\ \phi \in L_{\mathbf{Q}}\}$$

This is because $\mathbf{T}_{\alpha+1}$ is an extension of $\mathbf{T}_\alpha$ by Feferman reflection for each ϕ, which is soundness-preserving with respect to the domain. To sum up, according to the present interpretation of Gödelian anti-mechanism, the idealised arithmetical output of the non-mechanical mind can be modelled by a theory constructed from a Feferman reflection sequence based on **PA**.

There remains a question as to the *length* of this reflection sequence: how many times can the non-mechanical mind apply Feferman reflection to **PA**? As Good emphasises (1967, p.146), it is perfectly possible to produce a Turing machine corresponding to $\mathbf{PA}_\omega$, which extends $\mathbf{PA}_n$ by a reflection principle for all $n \in \mathbb{N}$. Even where that principle is Feferman reflection, $\mathbf{PA}_\omega$ is recursively axiomatized and hence corresponds to some Turing machine. So the Gödelian anti-mechanist must accept that the mind is modelled at the very least by a theory more powerful than this. And so the dialectic continues at least to transfinite successor stages. Note that this does not imply that the anti-mechanist must think that the idealised mind can perform a transfinite number of reasoning steps in the course of the dialectic, since it is not obvious that constructing the theory $\mathbf{PA}_\omega$ requires doing infinitely many things. All I'm arguing is that, according to the anti-mechanist, our abilities must

be more powerful than the theory resulting from the ωth application of Feferman reflection to **PA**.

At limit ordinals, we cannot apply Feferman reflection directly to a previous theory, but instead must formulate a means of asserting the Feferman reflection principle of all theories earlier in the sequence. But extending some sound theories by the assertion of their soundness will result in a sound theory, so the use of Feferman reflection at limits stages of the sequence must be acceptable to the Gödelian as well. This argument will only carry weight if we already accept that certain semantic properties (in particular, the property of being an index of a sound theory in a reflection sequence based on Feferman reflection) are suitable for transfinite induction over the ordinals. But given the intimate link between the ordinals and induction, this assumption can be readily granted.[4]

Since the theory $\mathbf{PA}_\omega$ is recursively axiomatized, it has a finite presentation despite an infinite number of axioms, just like **PA**. So thinking that we can reflect at least this much is well within the bounds of ordinary idealisation where we do away with finite limits on time and paper. However, the exact nature of the idealisation involved in the anti-mechanism debate is somewhat unclear. It is therefore possible that applying the reflection principle even further than this may constitute an infinitary idealisation going beyond what is ordinarily permitted in the mechanism debate.[5] If so, I think this makes my account no less Gödelian. Gödel himself was quite happy to entertain the idea that a finite mind is capable of an infinite number of distinguishable mental states (Wang 1996, p.196) and can store an infinite amount of information (Wang 1996, p.193). Hence whatever the true nature of the idealisation here, I think my proposal still represents a reasonable reconstruction of Gödelian anti-mechanism from the meagre textual evidence that is available.

So, our Gödelian thinks that our arithmetical capacities are at least as powerful as the αth member of a Feferman reflection sequence on **PA** for some transfinite α. Hence we need to define a transfinite Feferman reflection sequence up to some limit ordinal λ greater than the index of any member of the sequence:

[4]For those who may have moderate scepticism on the matter, I should point out that (as we will see later) we only require these properties to be suitable for induction in a small initial segment of the recursive ordinals.

[5]My thanks to an anonymous referee from *The Review of Symbolic Logic* for pointing this out.

> $\mathbf{T}_0$ is the theory **PA**, which we recognize to be sound.
>
> For each $\alpha < \lambda$, where α is an ordinal successor, $\mathbf{T}_{\alpha+1}$ is $\mathbf{T}_\alpha$ extended by the Feferman reflection schema.
>
> For each limit ordinal $\beta < \lambda$, the set of axioms of $\mathbf{T}_\beta$ is chosen such that $\mathbf{T}_\beta$ extends $\mathbf{T}_\alpha$ by Feferman reflection for each $\alpha < \beta$ (including 0).

Though I've drawn heavily on Lucas' own presentation of the anti-mechanist argument in this section, the arguments, I think, apply more broadly to Gödelian forms of anti-mechanism. If we can, in principle, out-perform any Turing machine at the task of proving arithmetical theorems, then a natural articulation of our abilities is that we can iteratively apply reflection principles. The ability to iterate reflection principles weaker than soundness cannot guarantee that the idealised mathematician out-performs a Turing machine. Since Feferman's soundness principle is equivalent to the other plausible candidates for representing such an ability, the arguments of this section show that the Gödelian anti-mechanist should take the idealised theorem-proving power of the human mind to correspond in some way to a transfinite Feferman reflection sequence based on **PA**.

1.4 Feferman's Completeness Theorem

In this section, I'll explore some formal properties of arithmetics generated by transfinite Feferman reflection to show that this procedure really does generate the kind of model of the mind which the Gödelian wants. Transfinite reflection, in some form or another, has appeared in the literature on Gödel and Lucas before, though to my knowledge it has not previously been used in an attempt to refute Gödelain anti-mechanism, a task which will be taken up in the next sections of the chapter.[6] For

[6]The issue was originally raised by Good (1967, p.146) who emphasised that a finite machine can be constructed corresponding to $\mathbf{PA}_\omega$, meaning that the one-upmanship must be continued into the transfinite if Lucas is to refute the mechanist (which, for the record, Good did not believe was possible). The issue is also discussed by Shapiro (1998, pp.285-293), who concludes that 'running up the ordinals' doesn't help the anti-mechanist at all. As I've already argued, the anti-mechanist is to some extent *obliged* to run up the ordinals, since their model of the mind, in its mathematical

now, we must note four important properties of reflection principles and reflection sequences.

The first is that reflection principles in general are sensitive to the particular axiomatic presentation of a theory to which they are added. In ordinary cases, the addition of a new axiom to two extensionally equivalent theories (i.e. sets of axioms with the same deductive consequences) results in two new theories which are extensionally equivalent to one another. The addition of reflection principles to extensionally equivalent theories does *not*, however, preserve extensional equivalence in this way. This is because reflection principles code information about axiomatic theories (such as their consistency or soundness) *under some particular description of those theories*. Consider the following example based on Feferman's early discussion of the issue (Feferman 1960, pp.36-37): suppose we have two consistent sets of axioms, $\mathbf{A}$ and $\mathbf{B}$, both of which extend $\mathbf{PA}$. Suppose further that $\mathbf{A}$ and $\mathbf{B}$ are extensionally equivalent, but that they are axiomatized very differently to one another (though both recursively so). As a result, these two theories have non-identical canonical consistency sentences. By Gödel's second theorem, $\mathbf{A} \nvdash Con_{\mathbf{A}}$, so by construction $\mathbf{B} \nvdash Con_{\mathbf{A}}$, and vice-versa. Given the difference in the description of the sets of axioms, we can finally suppose that $Con_{\mathbf{A}}$ and $Con_{\mathbf{B}}$ are not provably equivalent. This means that when applying a reflection principle, the choice between extensionally equivalent theories is crucial, since in scenarios of this kind, $\mathbf{A} \cup \{Con_{\mathbf{A}}\}$ isn't extensionally equivalent to $\mathbf{B} \cup \{Con_{\mathbf{B}}\}$.

A second important matter to note is that the members of a reflection sequence are theories in the language of arithmetic. Hence when we formulate instances of Feferman reflection for theories constructed by transfinite iteration of that principle, like $\forall x\ Pr_{\mathbf{PA}_\omega}(\overline{\phi}(\dot{\overline{x}})) \rightarrow \forall x\ \phi(x)$, we need a means of representing transfinite ordinals arithmetically. The language of arithmetic does not include symbols like 'ω' for such ordinals, so in this language we cannot define $\mathbf{PA}_\omega$ as the ωth extension of $\mathbf{PA}$ by iterated Feferman reflection. But we need to define this theory's proof predicate in order to formulate the relevant instances of the reflection scheme. Hence the need for a coding mechanism. An ordinal 'notation' system fixes a map between the natural numbers and the order types of recursively well-ordered subsets of $\mathbb{N}$, in order

capacity, must be at least as powerful as some member of a reflection sequence with a transfinite index.

to code information about these *recursive ordinals* in the language of arithmetic. In the original work on the subject, Feferman (1962) uses Kleene's $\mathcal{O}$ notation, so we'll use that too. In this system, $\mathcal{O}$ is a subset of the natural numbers ordered by the transitive relation $<_{\mathcal{O}}$. Where $n \in \mathcal{O}$, the ordinal it represents is $|n|$, determined as follows: $0 \in \mathcal{O}$ and $|0| = 0$. If n represents α, then 2^n represents $\alpha + 1$ and $n <_{\mathcal{O}} 2^n$. Where $\{e\}$ is the e-th partial recursive function, if $\{e\}$ is total, its range is in $\mathcal{O}$, and for all n, $\{e\}(n) <_{\mathcal{O}} \{e\}(n+1)$, then $3 \cdot 5^e \in \mathcal{O}$, for all n, $\{e\}(n) <_{\mathcal{O}} 3 \cdot 5^e$, and the ordinal $|3 \cdot 5^e|$ is the supremum of the ordinals $|\{e\}(n)|$ for all n.

Thirdly, it is vital to note that in the language of arithmetic, we can't represent the structure of the recursive ordinals (i.e. ordinals $< \omega_1^{CK}$) uniquely. Each limit ordinal $< \omega_1^{CK}$ has infinitely many notations in $\mathcal{O}$, so the order $<_{\mathcal{O}}$ is partial, and branches infinitely at all and only limit ordinals. There are thus infinitely many totally ordered paths *through* $\mathcal{O}$ which assign a unique notation to each recursive ordinal. So for any given base theory and reflection principle, there are many different reflection sequences up to a given limit ordinal, corresponding to different ways of coding the indices of the theories in the sequence in order to correspond to a path within $\mathcal{O}$ up to that limit.[7]

A transfinite recursive *progression* on **PA**, for a given reflection principle, is the set $\{\mathbf{T}_n | n \in \mathcal{O}\}$. We can then construct a total recursive function f from numbers to theories that, when the argument is some $n \in \mathcal{O}$, takes as its value the theory $\mathbf{T}_{|n|}$ (under the particular description given by n).[8] Hence when presented with a machine corresponding to a particular theory by the mechanist, the Gödelian can 'easily' trump it, if it is known to be sound, by using Feferman reflection, because in Kleene's $\mathcal{O}$ it is a simple matter to find the next ordinal notation after a given one. Having done so, applying f gives us the next theory in the sequence. The anti-mechanist will then have a theory in hand stronger than the previous one which is known to be sound (since **PA** is sound and Feferman reflection is soundness-

[7]To clarify, this means that there are two distinct senses in which we can talk about the index of a theory. On the one hand, we can mean by 'the index of a theory' the ordinal position of that theory in a reflection sequence. On the other hand, talk of 'the index' might mean the ordinal *notation* used to code the required ordinal information in some presentation of the theory in the language of arithmetic. In most cases, context will make the intended sense clear. Where both senses are relevant to a single point, letters in vertical bars will stand for the ordinal index, and unadorned letters will stand for the notational index.

[8]I owe much here to Shapiro's presentation of the matter (1998, p.287).

preserving).

Finally, a theorem of Church and Kleene shows that there is no recursive enumeration of the recursive ordinals, from which it follows that $\mathcal{O}$ isn't recursively enumerable either. Consequently, the property of being a member of $\mathcal{O}$, and hence of being a member of a transfinite reflection progression, isn't definable by any formula in the language of arithmetic.

That concludes the technical preliminaries. A central mathematical result which bears on the current issues is the following:

> **Feferman's Completeness Theorem:** For any transfinite recursive progression extending $\mathbf{PA} = \mathbf{T}_0$, every true sentence of number theory is provable from $\bigcup_{n \in \mathcal{O}: |n| < \omega^{\omega^\omega}} \mathbf{T}_n$

Although a number of different results go by the name of 'Feferman's completeness theorem', the result so named above is a restatement of his theorem 5.13 (1962, p.308). Moreover, for any progression, some particular path b in $\mathcal{O}$ is such that $\bigcup_{n \in b: |n| < \omega^{\omega^{\omega+1}}} \mathbf{T}_n$ is complete, meaning that a reflection sequence up to a small ordinal proves every arithmetical truth (Franzén 2004, §14.3). So as promised, Feferman's theorem provides a small bound in the recursive ordinals on the length of the reflection sequence required to satisfy the anti-mechanist. Indeed, the bound can be reduced further to ω^{ω^2+1} (Franzén 2004a, p.386).

This result is certainly a striking one. Notably, it vindicates the use of transfinite soundness reflection as an explication of the Gödelian non-mechanical mind: Gödel's own argumentative strategy was to refute mechanism by way of his rationalistic optimism (Wang 1996, p.185), so a properly Gödelian interpretation of anti-mechanism should be one under which no arithmetical propositions are absolutely undecidable.[9] And indeed, if the mind is the sort of thing which in its mathematical respects can be modelled by a suitably rich theory obtained from a transfinite recursive progression of **PA** by Feferman reflection, then it follows by Feferman's theorem that no true number-theoretic proposition is absolutely undecidable (just take the modelling theory to be some arithmetically complete $\bigcup_{n \in b: |n| < \omega^{\omega^2+1}} \mathbf{T}_n$ for suitable b). More

[9] Of course, there are interpretations of Gödel's disjunction under which both disjuncts are true, but Gödel himself accepted the anti-mechanical disjunct and denied the existence of absolutely undecidable arithmetical propositions.

importantly, Feferman's theorem makes it easy enough to see how a dialectical presentation of anti-mechanism ascribes extraordinary mathematical abilities to the idealised reasoner. This will be the subject of the next section.

1.5 Enumerating $\mathcal{O}$

Thanks to Gödel's incompleteness theorems, no arithmetically complete theory is recursively axiomatizable. If such a theory is to be an explanatory model of our idealised arithmetical capacities, we need an account of by what non-recursive procedure an idealised agent can determine the axioms of such a theory in order that they might be used in a proof. If no account of such a procedure can be given, the non-recursive theory in question will fail to explain our idealised arithmetical abilities.

Consider, for example, the theory **TA**, or 'true arithmetic'. This theory is (non-recursively) axiomatized by every true sentence of the language of arithmetic, i.e. $\mathbf{TA} = \{\phi \mid \phi \in L_{\mathbf{Q}} \wedge \phi \text{ is true}\}$. **TA** fails as an account of idealised human proof procedures because we can't actually *use* **TA** to prove anything we didn't already know, since we have to use some other means of proof to determine what the axioms are. Even the Gödelian who thinks that the theory is *extensionally* correct as a model for our idealised arithmetical knowledge must recognise that it isn't an epistemically viable arithmetical theory like **PA**. We take **PA** to successfully model (at least part of our) arithmetical knowledge because, limitations of time and paper aside, anything provable from a canonical presentation of **PA** is thereby provable by us too. The axioms can be recognized by an effective procedure, and the tractable inference rules are ones that we can apply for ourselves.

By contrast to **TA**, an arithmetically complete theory constructed out of a Feferman reflection sequence on **PA**, which I'll call a 'Feferman arithmetic' for brevity, might have looked much more promising as a model of our idealised arithmetical capacities. We can follow proofs in **PA**, the base theory, and the axiom schema for Feferman reflection appears to be something which we can easily recognise, and use to extend any theory in order to overcome some incompleteness. Indeed, no matter what theory a mechanist might present to us, we might be tempted to think

that, if it is sound, we can extend it by Feferman reflection and thereby prove something which the machine corresponding to that theory is incapable of demonstrating. There is no corresponding temptation to think that we could defeat such a machine by inferring a consistency sentence for its corresponding theory from **TA**!

But our ability to apply Feferman reflection to a theory presented to us by the mechanist rests on our ability to recognise (by proof or other means) when natural numbers represent recursive ordinals in some fixed notation system, because such natural number notations appear in the instances of the axiom scheme for Feferman reflection for a given theory. If we cannot recognise whether a presentation of a theory given by the mechanist is indexed by a member of $\mathcal{O}$, then we cannot tell whether that theory is an extension of **PA** by a soundness-preserving reflection principle. And if we cannot determine whether a theory presented by the mechanist is sound, then we have no justification for extending it by the relevant instances of Feferman reflection in order to prove something that the machine in question cannot. The issue for the anti-mechanist is that there is no recursive procedure for identifying notations for ordinals, since $\mathcal{O}$ is not recursively enumerable. We are therefore owed an account by the anti-mechanist of what procedure we might use, under idealisation, in order to recognise notations for recursive ordinals, and hence employ the axioms of a Feferman arithmetic in our proofs.

To my knowledge, no Gödelian anti-mechanist has supplied such an argument. Lucas (1996, p.111) seems to think that we have have such an ability, but offers no argument for thinking so. He remarks that the mechanist isn't entitled to assume that we *can't* non-recursively enumerate the ordinal notations on the basis that the mind is limited to mechanical operations. The problem is that the mechanist needn't make such a question-begging inference in this case; the fact that no one has the slightest idea *how* such an enumeration could be performed is sufficient to place the burden of proof squarely on the shoulders of the anti-mechanist.

Lucas also cites Gödel and Wang as rejecting mechanism because we can enumerate ordinal notations (1996, p.111). Gödel certainly had this view, although he acknowledges that the notion of a non-recursive procedure is far from clear. Indeed, he cites 'the process of defining recursive well-orderings of the integers' as a known example of such a procedure (quoted in Wang 1974, pp.325-6). Since the recursive ordinals are exactly the order types of recursive well-orderings of the natural num-

bers, Gödel's view is unambiguous. It is notable, however, that Gödel does not offer an argument for his view here. The ascription of the view to Wang is somewhat suspect, since earlier in the chapter Lucas cites, Wang claims that considerations relating to the supplementation of theories using reflection principles 'are of little help with regard to establishing the superiority of man over machine' (Wang 1974, p.320). The anti-mechanical argument for our ability to enumerate $\mathcal{O}$ remains elusive.

Lucas does, however, offer an argument to the effect that the anti-mechanist needn't claim any such ability for the idealised mathematician. As remarked above, if $n \in \mathcal{O}$ is a notation for α, then $2^n \in \mathcal{O}$, and 2^n is a notation for $\alpha+1$. So, although enumerating $\mathcal{O}$ is a non-mechanical matter, calculating the next ordinal notation after being presented with some previous notation *is* a mechanical matter. Lucas claims, therefore, that he doesn't need to enumerate ordinal notations; rather, he just needs to calculate the *next* ordinal notation, whenever the mechanist presents him with a theory indexed by such a notation. And we have good reason to believe that he can do so (Lucas 1996, p.112).

This argument, however, is unconvincing. Even if it is a requirement of the dialectical scenario that the mechanist put forward a machine the corresponding theory of which is *actually* an extension of **PA** by iterated Feferman reflection, it seems unreasonable to require that the mechanist be able to *prove* that the their favoured machine has this property. After all, mechanism is not itself a mathematical thesis, but a hypothesis that the human mind is limited in various respects. So the mechanist, in the dialectical scenario, should be able to put forward some machine, and tentatively claim that they believe it can prove anything a human could prove. Lucas, when presented with such a machine, can out-perform it if he can determine the theory corresponding to the machine, verify that it is indexed by an ordinal notation, and then apply Feferman reflection to it get a stronger theory. Verifying the index is crucial; without doing so, we have no reason to believe that any sentence Lucas produces which the machine cannot is actually true. So Lucas doesn't simply need to know how to calculate powers of 2, as he claims. Rather he needs the ability to recognise ordinal notations when presented with them, which comes to the same as the ability to enumerate $\mathcal{O}$.

The need for the anti-mechanist to posit such an ability has proved to be just

as contentious as the initial claim that we can 'see' the truth of Gödel sentences. Turing highlights that non-mechanical 'ingenuity' is required to recognize a number as representing an ordinal (though not in that terminology) (Turing 1939, §11). Lucas cites this point, along with some remarks by Gödel, in support of his view that the human ability to recognise ordinal notations outstrips that of any machine (Lucas 1996, p.111). The problem, of course, is that whether 'we' can perform some ingenious operation which no machine could ever do is precisely the point at issue. Citations from Gödel and Turing should prompt us to take the issue seriously, but they should not on their own be taken to settle the issue.

Shapiro (1998, p.289) argues that there is a weak and a strong disambiguation of the claim that an idealised human can out-perform a machine at the game of enumerating ordinal notations. The weaker claim is that given any machine that enumerates ordinal notations, there will be some recursive ordinals it doesn't denote that a human could produce a notation for. The stronger claim is that an idealised mathematician can enumerate $\mathcal{O}$.[10]

Shapiro claims that the weak version is hopelessly vague, since it involves 'machine enumerating ordinal notations' as a parameter. For my part, if it is just the weaker claim being made by the anti-mechanist, then I can't see how the dialectic is any different from the initial Lucasian scenario. Although $\mathcal{O}$ is not recursively enumerable, for any $n \in \mathcal{O}$, $\{m|m <_{\mathcal{O}} n\}$ *is* recursively enumerable. So, for any machine that enumerates notations, there will be some recursive ordinals that it doesn't denote that a *machine* could produce a notation for. So the weak anti-mechanist thesis is too weak to distinguish humans from machines in the required fashion. To do this, the mind must be ascribed the ability to enumerate $\mathcal{O}$, or something equivalent. In other words, the Gödelian must make Shapiro's strong claim. Strictly speaking, the Gödelian must claim the weaker ability to enumerate a subset of $\mathcal{O}$ forming a path such that the Feferman reflection sequence whose members are indexed by it is arithmetically complete, but such a claim would be remarkably *ad hoc*.

With respect to the strong claim, that an idealised human reasoner simply could enumerate $\mathcal{O}$ by a non-recursive method, Shapiro has a rather different response.

[10]Essentially the same point is made in less detail by Good (1969, p.357).

He claims that this amounts to a view on which we are arithmetically omniscient, since we could simply run through the indices of a transfinite recursive progression on **PA** by Feferman reflection and come up with an arithmetically complete theory. Shapiro concludes, I assume sarcastically, that this is a 'wonderful thought' (1998, p.290).

My view is that Shapiro's argument misses the point entirely.[11] Crucially, claiming that a human reasoner could, in principle, enumerate $\mathcal{O}$ is *not* to claim of anyone that they are arithmetically omniscient (a claim which would be false). It is rather to claim that every arithmetical truth is provable by the *idealised* reasoner, and *that* was the view in play all along! I see no reason why the Gödelian should be bothered by Shapiro's response, given that the anti-mechanist view was from the start presented as an alternative to the view that there are absolutely unsolvable number-theoretic problems.

A more generous reading of Shapiro's complaint might perhaps be that *if* no arithmetical proposition is absolutely undecidable, then that can't be explained by an appeal to an ability to enumerate notations for recursive ordinals. But again, I think this would be incorrect.

Franzén has proved (2004, p.191) the existence of a primitive recursive unary function f such that for $\phi \in L_{\mathbf{Q}}$, ϕ is true if, and only if, $f(\phi) \in \mathcal{O}$. Franzén's result shows that if we could enumerate $\mathcal{O}$, then by the simple application of a primitive recursive function, we could prove any arithmetical truth, given time and paper. Hence if we had a good reason to think that we could enumerate the ordinal notations, we would have a good reason to believe that all arithmetical truths were provable. So essentially the ability to enumerate ordinal notations *could* be used to explain why all arithmetical propositions are provable.

Shapiro's criticisms miss their mark, but this doesn't change the fact that we have no reason to believe that we *can* enumerate $\mathcal{O}$, and hence no good reason to believe that a Feferman arithmetic isn't an unusable theory, just like **TA**, because we have no way of determining its axioms. Despite that, it's fair to say that any Feferman arithimetic would still be a *better* model of our arithmetical capacities than **TA**, since we can readily determine some of the axioms of the theory, namely the

[11] Shapiro's argument might be a successful *ad hominem* against Lucas. My point here is that it does not present a general problem for Gödelian anti-mechanism.

fragment equivalent to **PA** plus the instances of the Feferman reflection scheme for theories in the sequence indexed by a recognizable ordinal notation. Nevertheless, we have been given no reason to believe that we can in general make use of the proof relation in a Feferman arithmetic. Since that relation isn't recursive, and we have been given no alternative account as to how we might determine the axioms of such a theory, we have no reason to believe that the proof relation of such a theory is one we could actually make use of. And if the proof relation of a theory isn't one we could actually use to prove things, then that theory under its given presentation cannot be part of a decent model of our arithmetical knowledge, however idealised.

An objector might insist, however, that the problem with **TA** is not that it isn't recursively axiomatizable, but rather, that the given axiomatization involves a problematic notion of arithmetical truth. Since axiomatizing a Feferman arithmetic does not overtly involve such a notion, it is possible that this flaw of **TA** isn't shared by the kind of theory envisaged by the anti-mechanist. In the next section we'll see why this is mistaken, though this can only be seen in the proof of Feferman's theorem, rather than its statement.

1.6 The Failure of Autonomy

The proof of Feferman's theorem is itself extremely long, and very mathematically involved. Interested readers are advised to consult the original article; for now I shall attempt to sketch a proof of the theorem that makes manifest its philosophically important aspects.

The preliminaries to the proof will be familiar to anyone working in the general area: give a canonical presentation of **PA** and arithmetize the syntax in order to capture the required proof-theoretic notions. Then, select an ordinal notation system for representing recursive ordinals as natural numbers, and use this to define a reflection principle. The two central parts of the proof are as follows: First, define the required primitive recursive functions in order to prove the existence of a progression using Feferman's reflection principle and based on **PA**, generalizing on Turing's construction for proving the Π_1-completeness theorem for ordinal logics (see §7 below). These functions, informally speaking, verify that a given construction sat-

isfies certain properties that we require a transfinite progression to have, and serve to compute the ordinal bound on finding witnesses for these properties. Though mathematically integral to the proof, from a philosophical perspective characterising these functions and proving their existence is essentially book-work (Feferman 1962, pp.296-302). The second part of the proof involves characterizing a 'proof-description' for Shoenfield's ω-rule to obtain the bound within the progression such that for every true arithmetical ϕ, the union of earlier theories in the progression proves ϕ.

We can distinguish amongst reflection progressions a special kind, which Feferman calls *autonomous* (1962, pp.280-281). An autonomous progression is unlike the general recursive progressions previously examined, because the definition of such a progression is based on some formula, ψ, such that if $\psi(x)$ is valid, then $x \in \mathcal{O}$. In particular, for every $\mathbf{T}_n$ in an autonomous progression, some earlier theory proves $\psi(n)$ (1962, p.262). Essentially then, autonomous progressions are those that we can recognize to be reflection progressions using only techniques available during the construction of the progression by a mathematician (in the more general case, we will not have the ability to verify the indices of the theories, and hence won't know whether the construction of the progression has been successful). The formula ψ, in this scenario, functions as a kind of oracle allowing the mathematician to verify that the progression so far is indexed by a set with the required order properties.

However, proving Feferman's completeness result ineliminably relies on *non-autonomous* methods, as can be seen in the following way. Suppose we have $O \subseteq \mathcal{O}$, such that for every $d \in O$, there is some $\mathbf{T}_a$ such that $a <_{\mathcal{O}} d$ and $\mathbf{T}_a \vdash \psi(d)$. Then we can prove that $\bigcup_{d \in O} \mathbf{T}_d$ is recursively enumerable, and hence by Gödel's theorem does *not* prove every true sentence of number theory (Feferman 1962, p.262). Another way of seeing this is that if the ordinals up to ω^{ω^ω} have notations in O, then $\bigcup_{n \in O : |n| < \omega^{\omega^\omega}} \mathbf{T}_n$ is recursively axiomatizable. If a completeness theorem could be proved for autonomous progressions, then this theory would prove all true sentences of number theory, so it would witness the falsity of Gödel's theorem. The essentially non-autonomous character of the progressions required to obtain an arithmetically complete theory is crucial to seeing why the Gödelian anti-mechanist cannot formulate a satisfactory epistemology of arithmetic.

If mechanism is the view that the output of an idealised human mind is coex-

tensive with the output of some Turing machine, then anti-mechanism is the view that the idealised human mind can out-perform any Turing machine whatsoever. We are, of course entitled to an account of how this is possible, and what human theorem-proving might look like if anti-mechanism is true. I've argued that, from a Gödelian perspective, a natural account is that idealised human mathematicians can generate arithmetically complete theories, by transfinite iterated Feferman reflection from **PA**, and use them to prove things. I've already argued that without an account of how the idealised mathematician can non-recursively enumerate $\mathcal{O}$, this position is little more explanatory than the bizarre claim that the idealised mathematician can use **TA** to prove any arithmetical truth. This does not, however, preclude such a procedure from being described.

Another explanation for the uselessness of **TA** as presented is that the given presentation involves the notion of *arithmetical truth*. Since there is no recursive procedure for determining of an arbitrary formula in the language of arithmetic whether it is true, we cannot in general tell what the axioms of **TA** *are*, given that those are just the truths of arithmetic. In other words, an objector might insist that the **TA** and a Feferman arithmetic are in a disanalogous epistemic position. According to the objector, the problem with **TA** is not *merely* that it has no recursive axiomatization; after all, we should expect any anti-mechanist to think our arithmetical abilities ultimately have no recursive presentation. Rather, the problem with **TA** is that it very specifically employs the notion of arithmetical truth in the axiomatization, whereas a Feferman arithmetic does no such thing.

The problem with this suggestion, however, is that the ineliminable use of non-autonomous methods means that any Feferman arithmetic actually does share **TA**'s second flaw, the involvement of the concept of arithmetical truth in determining which sentences are axioms. It is difficult to see that this is so, since the presentation of a Feferman arithmetic does not explicitly mention arithmetical truth. Rather, the arithmetical truths are the deductive closure of a Feferman arithmetic, but the axioms are just those of **PA** plus the instances of lots of Feferman reflection schemata. The problem, put succinctly, is that making a selection of instances of Feferman reflection to build an arithmetically complete theory requires prior knowledge of certain arithmetical truths which are not provable during the construction process. This point requires some explanation however.

Consider exactly why it is that the anti-mechanist requires the idealised mathematician to have the ability to enumerate the whole of $\mathcal{O}$, rather than the weaker ability to simply follow a path through $\mathcal{O}$. After all, don't we just start with **PA**, add the Feferman reflection principle at successor stages of the sequence, and extend all of our previous theories at limits? If we need only the ability to *follow a path* through $\mathcal{O}$, the Gödelian position might seem more persuasive, given the substantial weakening of our idealised abilities.

Sadly for the Gödelian, more than this is required for the construction of a sequence of the correct kind. There are many different paths through $\mathcal{O}$, which branches infinitely at all and only limit ordinals. So after each limit ordinal in the construction process, for example in the construction of stage $\omega + 1$ in the sequence, we need to use one of infinitely many possible means of arithmetically representing the axioms of the previous theory, in this case $\mathbf{T}_\omega$ (Franzén 2004, §11.2). Since reflection principles are sensitive to the presentation of a theory, there is no guarantee that any of these choices of arithmetical representations of the ωth member of the sequence yield equivalent results further along in their respective paths. As it turns out, the choice of path is vitally important:

> **Feferman–Spector Theorem:** There are paths Z through $\mathcal{O}$ that constitute a notation for every ordinal $< \omega_1^{CK}$, such that $\bigcup_{n \in Z} \mathbf{T}_n$ is incomplete with respect to the Π_1 sentences (Feferman and Spector 1962, p.384). Moreover, there are $\aleph_0$ such paths (1962, p.389).

The Gödelian claims that the totality of what is humanly provable includes all arithmetical truths. The Feferman–Spector theorem shows that in order to construct a reflection sequence which models this property, we can't just choose any old path through $\mathcal{O}$ when selecting indices for theories in the sequence. Indeed, we must pick a path with very special properties. In order to progress along a path through $\mathcal{O}$ of the desired kind, at limit stages in the reflection sequence the choice of formula defining the axioms of theories in the progression must be made very carefully indeed. In particular, the construction must make use of highly convoluted 'definitions' of axioms that are only even recognizably such if we already assume that the sentence we are trying to prove at the given stage is true (Franzén 2004a, p.387).

The need for non-standard definitions of axioms in order to apply Feferman

reflection to theories corresponding to limit stages in our construction is deeply problematic. When using these definitions, some formula is *recognizable as an axiom only on the assumption that a given sentence is true*; the problem is that the particular sentence in question is the very sentence we wanted to prove at that stage. In consequence, we cannot even axiomatize the theory which the Gödelian supposes to simulate the workings of the mind without knowing in advance whether a given sentence we seek to prove from those axioms is true (exactly what the sentence is will depend on the path through $\mathcal{O}$ and the limit ordinal in question).

On no plausible epistemology of arithmetic is it a basic fact that we have knowledge of such truths independently of the axioms. Thanks to the failure of autonomy, the Gödelian finds themselves in the following situation: for each sentence ϕ in the language of arithmetic, some theory in the modelling reflection sequence is sound only if ϕ is true. If ϕ really is true then they have a proof; if not then the index of the theory fails to denote an ordinal and so the sequence fails to make proper sense (Franzén 2004, p.213). As Shapiro puts it, we're no better off here than simply adding ϕ to **PA** as an axiom; if it's true, then we have a proof in a sound theory, otherwise we have nothing to celebrate (Shapiro 2016, p.202).

This explains why using a Feferman arithmetic is only slightly more viable than using **TA** to model our arithmetical powers: for some sentences we have independent assurances of their truth (if, for instance they follow from **PA**), but for others all we have is that they follow from our Feferman arithmetic if they are true. This is precisely what the above presentation of **TA** tells us about such unknown sentences. The point, as we shall see in the next section, holds of weaker progressions too; Turing went so far as to claim that his Π_1-completeness theorem was 'of no value' (Turing 1939, §9) in the context of making an observation similar to Shapiro's above.

Importantly, the Gödelian cannot simply posit that we have the required enumerative abilities with respect to $\mathcal{O}$ in order to side-step the problem; this ability is, in a sense, equivalent to the ability it is supposed to explain (as shown by Franzén's function), and so is just as much in need of explanation. And although there is some motivation for thinking that we can apply reflection principles justifiably (because, for example, we know that **PA** is sound), positing a special ability to enumerate particular paths through $\mathcal{O}$ is remarkably *ad hoc*.

The arguments of this section and the previous show not only that transfinite

applications of reflection principles fail to help the Gödelian anti-mechanist. The Gödelian anti-mechanist is *obliged* to deploy them, and this in turn serves to *refute* the Gödelian picture. If Gödelian anti-mechanism were successful, then the idealised mathematician would be able to prove any arithmetical truth by deploying an arithmetical theory the axiomatization of which relies ineliminably on prior knowledge of (certain of) its theorems. And I take it that this is a position which no serious epistemology of arithmetic can tolerate.[12]

1.7 Modest Anti-Mechanism

In this brief section, I'd like to show that my arguments generalize to weaker forms of anti-mechanism that don't presuppose the in-principle decidability of every arithmetical sentence. Call an anti-mechanist *modest* if they think that our ability to overcome incompleteness by iterated reflection is restricted in some way. I'll show that any remotely strong modest position is still subject to the arguments given previously. To do so, I'll use Turing's completeness theorem:[13]

> **Turing's Π^0_1 Completeness Theorem:** For any transfinite recursive progression extending $\mathbf{PA} = \mathbf{T}_0$ using a reflection principle at least as strong as consistency reflection, for any true Π^0_1 sentence, ψ, there is some $a \in \mathcal{O}$ s.t. $\mathbf{T}_a \vdash \psi$ and $|a| = \omega + 1$.

The interest in the proof of Turing's theorem is that, just as with Feferman's theorem, we make ineliminable use of non-standard definitions (in this case of the axioms of $\mathbf{PA}_\omega$) that are recognizable as such only on the assumption that ψ is true in order to carry out the proof. In other words if we *don't* make that assumption, then we can't verify that the relevant index codes an ordinal. This is epistemologically significant because it shows that, even so low down in the arithmetical hierarchy as Π^0_1,

[12]This is not to claim that we have *no* knowledge of the consequences of the axioms independently of deduction from those axioms. We might know, for example, that *any* correct axioms arithmetical axioms prove that $2 + 2 = 4$ (see chapter 5). However, these very obvious truths are not the kind of arithmetical truths required by the anti-mechanist in this context, since such mundane propositions are provable in **PA**.

[13]The proof appears originally in (Turing 1939). Significantly improved statements and proofs can be found at (Feferman 1988, §7) and (Franzén 2004, ch.14)

and with a reflection principle as weak as consistency, we cannot model the modest Gödelian mind by means of a theory constructed from an autonomous transfinite reflection progression based on **PA**.

The reason for this is that the construction relies on the non-autonomous character of the required progression, and so the modest Gödelian must think the idealised mind has some special ability to recognize certain arithmetical truths. As Turing puts it, by means of these progressions 'it is possible to prove Fermat's last theorem (if it is true), yet the truth of the theorem would really be assumed by taking a certain [number] as an ordinal [notation]' (1939, §9). Therefore the arguments made in this chapter apply not only to full-blown Gödelian anti-mechanism, but also to more modest anti-mechanist positions.

For example, it might be suggested by a modest anti-mechanist that some arithmetical propositions are absolutely undecidable, and hence that no Feferman arithmetic can be the correct arithmetical model of the mind. However, the modest anti-mechanist might think we have a limited ability to overcome incompleteness. They might argue that the idealised human can, in principle, recognize true universal generalisations over the numbers. This would certainly be a weaker hypothesis than the full-blown anti-mechanism with which this chapter is concerned. And perhaps such a view might be appealing to some, in light of the fact that Penrose restricts his Gödelian argument to the Π^0_1 sentences (Penrose 1994, p.96).

However, a mind which could recognize, by intuition or otherwise, the truth of every true Π^0_1 sentence could only be modelled by a theory that was at least Π^0_1-complete. Turing's theorem above then shows that a mind with such abilities cannot be modelled by a theory constructed from an autonomous progression based on **PA** using any reflection principle as strong as consistency. A reflection principle at least this strong will be required, since the putative non-mechanical mind has the ability to recognize true Π^0_1 sentences, and a consistency statement for a theory is itself Π^0_1. This means that the arguments presented in the previous sections will apply to this more modest position (since a Π^0_1-complete theory will also not be recursively axiomatized, thanks to Gödel's theorems).

Of course, there may be anti-mechanist positions even more modest, on which Turing's theorem has no bearing. But the point of this section is to illustrate that any Gödelian anti-mechanism that is strong enough to be tempting will be such that

the arguments of this chapter will apply to it. Such a position will inevitably presuppose knowledge of certain complex arithmetical truths on behalf of the idealised mathematician, despite alleging to explain the idealised mathematician's knowledge of those very truths.

Conclusion

I'd like to end with a few comments on what I take to have been established in this chapter, as well as a few comments on the limitations thereof. According to the Gödelian, there are, in principle, no absolutely unprovable number-theoretic truths. Accordingly, the mind cannot be modelled, in principle, by a Turing machine.

I've argued that the Gödelian is nonetheless obliged to provide something like a model of the arithmetical abilities of the idealised mathematician, and that the Feferman arithmetics are best equipped to function as such a model. These theories are constructed via transfinite iterated soundness reflection on **PA**. However, the mathematical properties of theories so constructed presuppose that the idealised mathematician has inexplicable access to certain arithmetical truths, despite alleging to explain why those truths are provable in-principle.

Moreover, the arguments presented in this chapter apply to substantially weaker forms of anti-mechanism than full-blown Gödelianism. Hence, Gödel's incompleteness theorems do not support any remotely strong form of anti-mechanism.

I've made no attempt to engage with anti-mechanism more broadly construed. But if the mind cannot be modelled by a Turing machine, I at least hope to have shown that anti-mechanism motivated by the incompleteness theorems cannot be the correct account of why this is so.

The arguments I've presented also provide some support for the view that there are absolutely undecidable arithmetical propositions, at least indirectly. If every arithmetical truth is provable to the idealised mathematician, then they must be able to deploy some operation that cannot be performed by a Turing machine. I take myself to have shown that the most plausible account of such an operation we might perform, involving the transfinite iteration of a reflection principle, is untenable. The first attempt to reduce the incompleteness of arithmetic can therefore

be regarded only as a modest success: we do not have the ability to unrestrictedly apply reflection principles to the axioms of arithmetic in order to eliminate incompleteness. In the next chapter, the considerations raised here will be developed into a more direct argument for the existence of absolutely undecidable arithmetical propositions.

Chapter 2

Absolutely Undecidable Arithmetical Propositions

Introduction

In this chapter, I'll argue that there are absolutely undecidable propositions about the natural numbers that can be expressed in the language of arithmetic. In the previous chapter, we saw that Gödelian anti-mechanism was an unacceptable response to the disjunctive argument. In this chapter, we'll examine more closely Gödel's *rationalistic optimism*, according to which no well-defined mathematical problem is unsolvable in an 'absolute' sense. This straightforwardly implies Gödel's well-known view that no true propositions of arithmetic are absolutely undecidable. We'll take a proposition ϕ to be *absolutely* undecidable if and only if it is undecided by every formal theory recognizable by us as sound; i.e. if $\mathbf{T} \nvdash \phi$ and $\mathbf{T} \nvdash \neg\phi$ for any recognizably sound $\mathbf{T}$.[1]

Firstly, I'll examine two of Gödel's central arguments against the existence of absolutely undecidable propositions, and show that both of them, while persuasive to some degree, ultimately fail. The first, and most persuasive (addressed in §1–§3), seeks to establish that arithmetical propositions which are deductively independent of the axioms of **PA** are nevertheless just as evidently true. The argument trades on the intuitive idea that we can 'see' the truth of certain Gödel sentences, but fails for reasons related to the coding of information about transfinite recursive ordinals into the language of arithmetic.

The second argument (addressed in §4) is made firmly in the context of rationalistic optimism, and seeks to establish that some kind of scandalous inconsistency is introduced to human reason by supposing that there are absolutely undecidable

[1]Note that this includes theories where ϕ is an axiom, hence this definition does not beg the question against anyone who thinks that some axioms can only be recognized by us as valid using informal modes of verification. For if ϕ were such a proposition, and we verified it informally, we would then recognize the theory $\{s|\phi \vdash s\}$ as sound, and hence ϕ would not be absolutely undecidable according to the definition given.

propositions. I disambiguate two readings of this argument; on one its premises are simply implausible, and on the other the argument fails for familiar reasons.

I'll go on to give an argument for the existence of absolutely undecidable arithmetical propositions. I have no deductive argument that such propositions exist, but the evidence as it is overwhelmingly stands in favour of their existence. In §6, I define the *recursive ordinal selection ability*, and show that our evaluation of Gödel's disjunctive argument hinges precisely on whether we have this ability under suitable idealisation. I consider two different accounts of what this ability might amount to: the Gödelian account, and a weaker alternative. I then argue that there is no good reason to believe that we possess the relevant ability in either sense. Hence belief that we have the selection ability is based purely on faith. On the other hand, a compelling case can be made *against* our possession of the recursive ordinal selection ability, from which the existence of absolutely undecidable arithmetical propositions follows.

Finally, we'll examine the upshot of this for another alleged philosophical implication of Gödel's theorem, namely Dummett's argument that the concept *natural number* is vague because the notion of an intuitively acceptable arithmetical proof is *indefinitely extensible*. In §9–§10, I argue that Feferman's theorem refutes the Dummettian position. In §11 I offer several responses on Dummett's behalf. None of them is adequate, and I show that even in the context of Dummett's constructivism, the vagueness of the concept *natural number* cannot be sustained.

2.1 The Evidence Argument

Gödel's first (and best known) argument for the absolute decidability of all arithmetical propositions appears in handwritten notes from the 1930s (Gödel 193?) discovered in Gödel's *Nachlass* (Davis 1995, p.156). The argument there is terrifically condensed, so I'll try to spell it out with a little more clarity. Gödel (quite rightly) supposes that if an arithmetical proposition is absolutely undecidable, it will be of one of the kinds that the incompleteness theorems tell us is troublesome.

In particular Gödel identifies a special class of polynomial expressions that give

rise to unsolvable *Diophantine Problems*.[2] The Diophantine problem corresponding to such a polynomial is to determine whether the equation has solutions in the integers for arbitrary integer values of the parameters. Each recursively axiomatized theory which can express all Diophantine problems of this kind is incomplete with respect to them (Gödel 193?, p.165).[3]

Gödel goes on to argue that the fact that **PA** (and its extensions) are incomplete with respect to Diophantine problems gives us no reason whatsoever to suppose that there are propositions of this kind which are undecidable in an absolute sense. Without loss of argumentative force, we'll re-cast Gödel's argument in terms of consistency sentences, for the sake of clarity.

The 'evidence argument', as I call it, proceeds as follows. Suppose you have some particular theory which is known to be sound, such as **PA**, the standard first-order formalization of arithmetic. It follows trivially that the theory is consistent. Hence, if we then recognise some sentence, such as $Con_{\mathbf{PA}}$, as an expression of this fact, our reasons for believing that $\mathbf{PA} + Con_{\mathbf{PA}}$ is sound are just as good as those for believing that **PA** itself is sound.[4] While not all sentences which entail the consistency of **PA** will be recognizable as such, a canonical consistency sentence like $Con_{\mathbf{PA}} =_{df} \neg\exists m\ Prf_{\mathbf{PA}}(m, \ulcorner 0 = 1 \urcorner)$ certainly is. Hence we know that $\mathbf{PA} + Con_{\mathbf{PA}}$ is sound. Since that theory gives us a trivial proof of $Con_{\mathbf{PA}}$, that sentence isn't absolutely undecidable. As Gödel puts it, the kinds of undecidable sentences gener-

[2]Polynomials are equations constructed only using variables, integer coefficients, addition, multiplication, and exponentiation (to powers in the naturals). Where F is such a polynomial in $m+n$ variables, Gödel's equations are of the following form: $F(a_1, ..., a_m, x_1, ..., x_n) = 0$, with the a_is considered as parameters and the x_js considered as unknowns. A schematic example of such an equation is $a_1 x_1^{n_1} + \ldots + a_i x_i^{n_i} = 0$.

[3]As a corollary of the Matiyasevich-Davis-Robinson-Putnam theorem, we can do away with parameters and reduce the Diophantine problem to whether a certain Diophantine polynomial has any solutions. That theorem implies that for each sufficiently expressive recursively axiomatized arithmetical theory $\mathbf{T}$, there is a true (and constructible) *Diophantine sentence* $D_{\mathbf{T}}$ undecided by the theory, in the sense that $\mathbf{T} \nvdash D_{\mathbf{T}}$ and $\mathbf{T} \nvdash \neg D_{\mathbf{T}}$. The Diophantine sentence says that a certain Diophantine equation has no solutions.

[4]One reason to be suspicious here is the idea that our epistemic warrant decreases with each inferential step in a deduction, as the possibility of making errors increases with the length of an argument. I take it that such concerns should not apply in the present context, because we are concerned with *absolute* undecidability. Since we are in general idealising away from finite lifetime and supply of paper, we can suppose that the mathematician can check and re-check any argumentative move arbitrarily many times, so that warrant is ideally preserved through each deductive step, whether formal or otherwise.

ated by the incompleteness theorems are 'exactly as evident' as the theorems of the old system (in this case, **PA**) (Gödel 193?, p.164). In principle, the same argument can be run for the new theory, and so on. It follows that canonical consistency sentences, and the undecidable propositions generated by the incompleteness theorems more generally, are not absolutely undecidable propositions.[5]

Something like this argument must be correct.[6] The inference from soundness to consistency is trivial, so as long as we accept the premise that we know that **PA** is sound, it follows that we do have a proof in some sound theory of many of the undecidable propositions generated by the incompleteness theorems. The question, then, is do we have a proof in some-or-another system of *all* such propositions?[7]

2.2 Intensionality

For the purposes of this chapter, we'll take Gödel's argument to be that the consistency sentences of knowably sound theories are absolutely decidable because the result of adding a consistency sentence to any such theory will be exactly as evident as the axioms with which we started.[8]

[5]There are other interesting candidates for an absolutely undecidable proposition, like CH, and Gödel has plenty to say about such cases. However, we will here confine our attention to the arithmetical case.

[6]Indeed, in other places Gödel takes some version of the argument, though perhaps not the full version, to be essentially truistic. See, for example (Gödel 1946, p.151).

[7]The argument could similarly be run using Gödel sentences. In this, the procedure for constructing the Gödel sentence makes it obvious that it is true. The same goes for Gödel's original Diophantine formulation, especially when strengthened by the MDRP theorem, since the construction of the Diophantine sentence of a sound theory is successful only if the corresponding equation really does have no solution in the integers.

[8]We could also have framed Gödel's argument in terms of a different requirement, namely that the theory in question is what Leach-Krouse (2016, p.226) calls 'proof-constituting'. A theory is proof-constituting if a sentence being a theorem of the theory constitutes its having a proof. It's perhaps worth noting that the concepts of proof-constituting and soundness are importantly distinct. There may be sound theories which we might be tempted to deny are proof-constituting (as, for example, with the Feferman arithmetics discussed in the previous chapter). Similarly, there may be unsound theories which we nonetheless take to be proof-constituting. For example, we might think that $\mathbf{ZFC} + V = L$ is unsound, yet produces genuine proofs relating to the constructible part of the set-theoretic hierarchy. However, since **PA** is undoubtedly both sound *and* proof constituting, we needn't concern ourselves with such minutiae for the purposes of this chapter.

It's crucial to note that the evidence argument, as given by Gödel, relies on an *intensional* relation between two theories, namely that the axioms of one are exactly as evident as those of another. This is the key to seeing why Gödel's argument cannot be sustained in full generality. In describing this relation as 'intensional', all I mean is the following: if P is *recognizable* as a consistency sentence for $\mathbf{T}$, then the axioms of $\mathbf{T} + P$ are exactly as evident as those of $\mathbf{T}$. However, if P is not recognizably a consistency sentence, then there is no reason to suppose that $\mathbf{T} + P$ inherits the evidential merit of $\mathbf{T}$.[9] This means that we can assess Gödel's argument in terms of whether undecidable sentences, like consistency sentences, are always recognizable for what they are.

The claim that the relation *the axioms of ϕ are exactly as evident as the axioms of ψ* is intensional is an extremely weak one. Suppose that (a) $\mathbf{T}$ is a sound theory, (b) the consistency of $\mathbf{T}$ is sufficient for the truth of P, and (c) there is no proof or other rationally compelling reason to believe (b). Plainly, it *does not* follow that the axioms of $\mathbf{T} + P$ are exactly as evident as the axioms of $\mathbf{T}$. This is all I mean by the claim that the relation in question is intensional.

It might be objected that this forces us to adopt an *internalist* interpretation of the relevant epistemic notions which is unwarranted. I don't want to comment at all on the more general epistemological debate here, but I do think in this particular dialectic a somewhat internalist interpretation of the relevant epistemic concepts is required. The reason is that in this debate, as in our previous examination of anti-mechanism, I'm quite happy to grant the Gödelian any degree of idealisation that they please. So clause (c), that there is no proof or other rationally compelling reason to believe that the consistency of $\mathbf{T}$ is sufficient for the truth of P, is supposed to mean that the entire community of mathematicians, idealised to have as much time, paper, and brain-power as you like, has no rationally compelling argument for the fact in question. Furthermore, given the presumptive mathematical necessity involved, these situations are of the kind where the consistency of $\mathbf{T}$ is sufficient for P, but there is no *possible* proof for creatures like us that this holds.

[9] In some cases there might be *independent* compelling reasons to believe P; but Gödel's argument is supposed to be fully general, and so relies on the tacit assumption that the axioms of the extended theory are always exactly as evident as those of the old theory *for exactly the same reasons.*

Crucially, this does not beg the question against the Gödelian, since thinking that the relation *the axioms of ϕ are exactly as evident as the axioms of ψ* is intensional in this sense does not involve asserting that any situation in which (a), (b), and (c) jointly hold does, or even could, obtain.

Given the idealisation we're allowing Gödel here, any assertion that the evidence relation must be extensional is essentially a change of subject, because to demand that the epistemic notions in play are interpreted extensionally is to stop thinking about what creatures like us could achieve, and to start thinking about what some mathematically divine mind could do. This might perhaps be a worthwhile pursuit, but it isn't the concern with which we started. Even Gödel himself writes that in the context of his disjunctive argument 'the epithet "absolutely" means that they would be undecidable, not just within some particular axiomatic system, but by *any* mathematical proof that the human mind can conceive' (1951, p.310).

The intensionality of the relevant notions is much clearer when we consider that the general idea of a consistency sentence is an informal one. By 'a consistency sentence for **T**' all I mean is a sentence in the language of **T** which expresses that **T** itself is consistent. The idea of such a sentence is of course closely tied to formal ideas in the arithmetization of syntax, but it is no less informal for that. In many textbooks on the issue, the idea of a consistency sentence is made canonical. For example we may stipulate that $Con_{\mathbf{T}} =_{df} \neg\exists m\ Prf_{\mathbf{T}}(m, \ulcorner 0 = 1 \urcorner)$, for any **T**. But that definitional schema is not a functor into which the axioms of **T** can simply be inserted, because the relation $Prf_{\mathbf{T}}$ is different for different sets of axioms, and hence must be defined afresh for different sets of axioms. This is not altered by the stipulation of a canonical shape for consistency sentences to take.

Crucially, the need to define a new proof relation occurs whenever the axioms of **T** are changed, even if this makes no difference whatsoever to the theory considered as a set of sentences. Hence there might be two theories, **T** and **S**, which are extensionally equivalent in the sense of sharing a deductive closure, but which differ wildly in their 'presentation', in that very different sentences are used to axiomatize the theories (this issue was also discussed in the previous chapter). In such a case, the relation $Prf_{\mathbf{T}}$ and $Prf_{\mathbf{S}}$ might appear so different that the consistency sentences $Con_{\mathbf{T}}$ and $Con_{\mathbf{S}}$ are not obviously equivalent, even if they are both 'canonical' in the required sense. So even if we restrict our attention to consistency sentences

of a canonical form, the *exactly-as-evident* relation is still intensional. If $\mathbf{T}$ is extensionally equivalent to $\mathbf{PA}$, but radically different in presentation, the axioms of $\mathbf{PA} + Con_{\mathbf{T}}$ cannot be assumed to be exactly as evident as those of $\mathbf{PA}$. $Con_{\mathbf{PA}}$ and $Con_{\mathbf{T}}$ may be equivalent, but unless we have a compelling reason to believe that fact, we are not entitled to infer $Con_{\mathbf{T}}$ from the observation that $\mathbf{PA}$ is sound.

So, given that the relation *is exactly as evident as* is intensional, Gödel's argument will only work *if* we can always recognize, for any theory $\mathbf{T}$, that $\mathbf{T} + Con_{\mathbf{T}}$ is indeed an extension of $\mathbf{T}$ by the addition of a canonical consistency sentence for $\mathbf{T}$. As we shall see, this is an assumption to which Gödel is certainly not entitled.[10]

What makes Gödel's position on consistency extensions so persuasive is that there are abundant examples where it is clearly correct. In the first instance, $\mathbf{PA}$ is sound. If we know this (and I think we do), then we have just as much cause to believe in the soundness of $\mathbf{PA} + Con_{\mathbf{PA}}$, $\mathbf{PA} + G_{\mathbf{PA}}$ ($\mathbf{PA}$ plus its canonical Gödel sentence), and so on. I'll argue, however, that this appearance is deceptive, and ultimately breaks down.

The chain of theories that stand to one another in the relation *the axioms of ϕ are exactly as evident as those of ψ* is very long; indeed it is infinitely long. Let $\mathbf{T}_0 = \mathbf{PA}$, and $\mathbf{T}_{n+1} = \mathbf{T}_n + Con_{\mathbf{T}_n}$. For any n we can, given time and paper, verify that the axioms of $\mathbf{T}_{n+1}$ are exactly as evident as the axioms of $\mathbf{T}_n$. We can, in principle, verify that the axiomatizations and constructed consistency sentences are correct, and then the same argument that convinced us that $\mathbf{PA} + Con_{\mathbf{PA}}$ must be sound because $\mathbf{PA}$ is should also convince us that $\mathbf{T}_{n+1}$ is sound. Since $\mathbf{T}_{n+1} \vdash Con_{\mathbf{T}_n}$, we

[10]One might object that this misrepresents Gödel's views on the nature of absolute provability. He ultimately came to think that a proposition is provable *tout court* if it follows from set theory plus some true large cardinal assumptions (Gödel 1946, p.151). Hence, the restriction in this chapter to arithmetic and the neglecting of set theory perhaps fails to do justice to Gödel's thought on the matter. The reasons for restricting ourselves to arithmetic here are as follows: firstly, it is unclear to what extent the various extensions of **ZFC** by large cardinal axioms should be, or even are, regarded as knowably sound. It is by no means clear, for example, that the Riemann Hypothesis could be settled by a proof in some extension of **ZFC** by large cardinal assumptions. Hence philosophically, the significance of the discussion is unclear if framed set-theoretically. Secondly, it is widely acknowledged that the large cardinals as Gödel conceived of them *can't* successfully frame a notion of absolute provability that works in the desired way, since by the Levy–Solovay theorem, even under powerful large cardinal hypotheses the size of the continuum is still sensitive to forcing (Koellner 2010, p.202). Hence we're better off here restricting our attention to undecidable propositions that are arithmetical (unlike CH) and for which Gödel offers a more persuasive argument.

have, for any n an argument that $Con_{\mathbf{T}_n}$ isn't absolutely undecidable.

2.3 Ordinal Notations

Gödel's argument succeeds for at least arbitrary finite extensions of a knowably sound theory by the iterated addition of consistency statements. But the theory $\mathbf{T}_\omega$, which extends every $\mathbf{T}_n$ in our sequence above by $Con_{\mathbf{T}_n}$, is recursively enumerable. Hence $Con_{\mathbf{T}_\omega}$, is true and independent of the theory. The crucial question is now whether Gödel's argument is successful in the case of *transfinite* iterated addition of canonical consistency sentences.

Given the argument above that the axioms of **PA** are just as evident as those of $\mathbf{T}_n$, for any n, and the fact that $\mathbf{T}_\omega$ simply extends the $\mathbf{T}_n$s by canonical consistency statements, let's grant that the axioms of $\mathbf{T}_\omega$ are just as evident as those of **PA**. Now $\mathbf{T}_\omega$ is incomplete, but according to Gödel's argument, our reasons for believing **PA** is sound are exactly as evident as our reasons for believing in the soundness of $\mathbf{T}_\omega$, and hence are exactly as evident as our reasons for believing in the soundness of $\mathbf{T}_\omega + Con_{\mathbf{T}_\omega}$.

At this point, the situation has changed drastically. When we've iterated the addition of consistency sentences only finitely many times, we can directly write down what the theory is. For example, $\mathbf{T}_2 = \mathbf{PA} + Con_{\mathbf{PA}} + Con_{\mathbf{PA}+Con_{\mathbf{PA}}}$. Evidently the process is laborious, but there's no obstacle to checking 'in principle' whether such a theory is an extension by the iterated addition of consistency sentences of **PA**, and hence whether Gödel's argument applies to it. However, we cannot use such brute force methods in representing a theory in our chain after the addition of infinitely many consistency sentences. We also can't simply write down '$\mathbf{T}_\omega$' and formulate its consistency sentence accordingly, since all our theories are couched in the language of arithmetic, and this does not include symbols for transfinite ordinals. If we want to assert that $\mathbf{T}_\omega$ is consistent by using a canonical consistency sentence, we need to fix a presentation of that theory in order to define a proof predicate for it. Hence we need to use a notation system for representing recursive ordinals in the natural numbers, such as Kleene's $\mathcal{O}$ or an equivalent, just as we did for reflection

progressions in the previous chapter.[11]

This is a big problem for Gödel's argument, because there is no recursive procedure for deciding whether a number represents an ordinal, nor is there such a procedure for determining whether two notations represent the same ordinal. So unlike in the finite case, there is no reason to suppose that, where β is transfinite, we can effectively recognize that the βth theory in a sequence is actually an extension of **PA** by transfinitely iterated consistency reflection. In consequence, there is no reason suppose that the axioms of such a theory really are exactly as evident as those of **PA**.

To spell this out in a little more detail: recall that

$$Con_{\mathbf{T}_\alpha} =_{df} \neg\exists m\ Prf_{\mathbf{T}_\alpha}(m, \ulcorner 0 = 1 \urcorner)$$

So the formulation of the consistency sentence is itself dependent on the predicate $Prf_{\mathbf{T}_\alpha}$, which is sensitive to exactly how the theory $\mathbf{T}_\alpha$ is axiomatized. Given that we are working in the language of arithmetic, we cannot define the βth theory in our reflection sequence as the βth extension of **PA** by iterated consistency reflection if β is transfinite. But we need a means to express that some of our theories are the result of transfinite iterated addition of consistency sentences to **PA**, hence the need for a coding system which allows for the representation of recursive ordinals. Even supposing that the transfinite iterated addition of consistency sentences doesn't spoil our ability to perspicuously formulate the proof predicate here, it remains that $\mathbf{T}_{\alpha+1}$ is only as evident as $\mathbf{T}_\alpha$ if we can recognize that $Con_{\mathbf{T}_\alpha}$ is true.

Given the intended generality of Gödel's argument, our recognition of the truth of $Con_{\mathbf{T}_\alpha}$ can't hinge on any special features that sentence may have. Rather, our recognition of its truth must consist solely in our recognition *that* it is a consistency sentence for an extension of **PA** by the iterated addition of consistency sentences. This is so only if $\mathbf{T}_\alpha$ is a member of our reflection sequence; i.e if the ordinal index is denoted in the arithmetical presentation of the theory by a notation in $\mathcal{O}$ of a recursive ordinal.

[11]Though we don't know the exact date of the relevant manuscript, Gödel's argument may well have been produced years before Kleene's work on ordinal notations, so I'm not making the anachronistic argument that Gödel was wrong to ignore these issues. Rather, I am claiming merely that they can help us see why the Gödel's arguments break down.

Given that there is no recursive procedure for recognizing whether a number 'denotes' an ordinal in this fashion, we have no reason to suppose that we can always recognize the truth of a sentence expressing the consistency of a theory in the transfinite stages of our sequence, as Gödel's argument requires. In fact, this seems to be a good reason to suppose that our ability to recognize the truth of the relevant sentences might give out at some point.

So Gödel's evidence argument looses its persuasive power once we've iterated the addition of consistency sentences into the transfinite. Furthermore we *must* iterate that procedure transfinitely, since a finite version of the evidence argument fails to show that the consistency sentence for $\mathbf{T}_\omega$ isn't absolutely undecidable.

2.4 The Irrationality Argument

Gödel communicated to Hao Wang a somewhat crypic second argument against the existence of absolutely undecidable arithmetical propositions. With respect to the hypothesis that there exist such propositions, Gödel claims that 'if it were true it would mean that human reason is utterly irrational in asking questions it cannot answer while asserting emphatically that only reason can answer them. Human reason would then be very imperfect and, in some sense, even inconsistent' (Wang 1974, pp.324–5).[12] Let's call this the 'irrationality argument'.

I hope we can all agree that Gödel's meaning here is difficult to discern. He might be claiming that the mere existence of an absolutely undecidable proposition is sufficient for some kind of inconsistency in human reasoning. Against this argument, the comments made above have nothing to offer. I think there are good historical reasons to think that Gödel *did* have such an argument in mind. Though his position on undecidability fluctuated over time, he eventually settled on a position called 'rationalistic optimism' (Wang 1974, p.324-326). Narrowly speaking, this is the view that any well-posed mathematical proposition can be proved or can be refuted. More broadly, it is the view that 'for clear questions posed by reason,

[12]The quotation here is not a direct quotation from Gödel, but from Wang's paraphrase of Gödel's argument. The source can certainly be considered a reliable report of Gödel's view, but we should not make too much of the precise phrasing of this argument.

reason can find clear answers' (Gödel 1961/?, p.318). With such a strong conception of the powers of human reasoning, it is quite possible that Gödel's intended meaning is simply that the existence of an absolutely undecidable proposition would be scandalous.

At least with respect to arithmetical propositions, some form of optimism seems to have been Gödel's view throughout the majority of his career; even in the 1930s, when Gödel *did* entertain the existence of absolutely undecidable propositions, these were set-theoretic, and generally related to the continuum hypothesis.[13] He wasn't, even at this time, convinced that the incompleteness theorems suggest the existence of absolutely undecidable *number-theoretic* propositions. Tieszen (2011, p.202) argues that Gödel eventually came to take the absolute decidability of mathematical propositions (including those of set theory) to be a 'postulate of reason', and several of his writings certainly support that reading. For example, his closing remarks of the Russell paper implore us not to abandon the ideas behind Leibniz's programme for a *Characterstica Universalis* (Gödel 1944, pp.140–141). In a later paper (1961/?, p.385), he cites a broad agreement with the Kantian conception of mathematics. Tieszen traces these remarks to assertions by Kant of the explicit solvability of all problems in mathematics:

> [T]here are sciences whose nature entails that every question occurring in them must absolutely be answerable from what one knows, because the answer must arise from the same source as the question; and there it is in no way allowed to plead unavoidable ignorance, but rather a solution can be demanded (Kant 1787, A476/B504).

Additional remarks at (A480/B508) make it clear that Kant considers mathematics at large to be such a science. In the light of this historical evidence, we can make a probable judgement that Gödel's remarks to Wang are intended to be read in a rather strong sense. I don't have much to say here about such an argument, but I think it would be of little appeal to philosophers who don't entirely share Gödel's rationalistic leanings. After all, it relies on a clearly non-standard notion of inconsistency or irrationality; we don't ordinarily take the inability to answer a

[13]Indeed, he claims that the absolute undecidability of CH is 'very likely' and 'highly plausible' in (193?) itself (p.175).

clearly posed question as a symptom of either affliction. However, I think a modest interpretation of this argument is of some broader interest.

The modest claim is that there is some kind of irrationality involved in thinking that we might be presented with some axioms for a theory which are known to be sound, and think that we can't determine the truth of the canonical consistency sentence for those axioms by means of mathematical reasoning. This is much more interesting a reading of Gödel's remarks, since it appeals to the popular idea that we can, for instance, 'see' that the Gödel sentence of **PA** is true. Indeed, read this way the irrationality argument is a corollary of the evidence argument.

Moreover, the argument fails for the same reasons that the evidence argument failed. According to the picture I've sketched, though there might be an absolutely undecidable true proposition, there cannot be a recognizable *example* of such a proposition that would give rise to the irrational scenario sketched above.[14] Suppose that we were presented with a consistency sentence that was alleged to be absolutely undecidable. This proposition would specify some axioms for a theory, $\mathbf{T}_n$, which is some extension of **PA** by iterated consistency, and exhibit some consistency sentence constructed by a specified canonical method from those axioms. If we can tell what we're looking at, then such a proposition could not be an example of something absolutely undecidable: if we can recognize that $\mathbf{T}_n$ extends **PA** in the right way, which involves recognizing that n denotes an ordinal, then we can recognize its soundness, and hence the soundness of $\mathbf{T}_n + Con_{\mathbf{T}_n}$. And *this* theory trivially decides $Con_{\mathbf{T}_n}$ in the affirmative. In other words, if we can recognize that the sentence is the sort of thing that might be absolutely undecidable, we can thereby recognize its truth. On the other hand, if we *can't* recognize that $\mathbf{T}_n$ extends **PA** in the right way, then we can't in general recognize that the theory is sound, and hence couldn't recognize that $Con_{\mathbf{T}_n}$ is indeed true given that it's constructed from the axioms of $\mathbf{T}_n$. In such circumstances, there is no reason to think that we have any other means of determining the truth of the sentence. So the sentences that are candidates for truth *and* absolute undecidability, can't be recognised for what they are.

Either way, this kind of absolute undecidability does not mean that we can be in the 'irrational' position of simultaneously having a theory known to be sound and

[14] Again, attention is restricted to the arithmetical case; perhaps we *can* recognize that CH is absolutely undecidable, for instance.

not having a means of proving its consistency: if we can't prove that the theory is consistent it's *because* we can't recognize that it is sound. This will be because we can't recognize that it extends **PA** in the right way; and in general there seems to be nothing 'inconsistent' or 'irrational' in supposing that we can't always recognise whether a natural number codes an ordinal or not. This is especially so given the lack of a recursive procedure for doing this.

2.5 Gödel's Disjunction Revisited

So much for Gödel's arguments against absolute undecidability. Both have persuasive features, but neither can be made to work in full generality. Having defused them, it's worth pausing to consider the state of the dialectic with respect to the disjunctive argument. Gödel himself never offered a Lucas-Penrose type argument for anti-mechanism, but rather took it to be a consequence via disjunctive syllogism from the truth of rationalistic optimism (Shapiro 2016, p.191). Similarly, in the previous chapter, I never offered a positive argument for absolute undecidability, just another disjunctive syllogism from the disastrous epistemological picture that Gödelian anti-mechanism left us with. The arguments of Gödel's just examined in the previous few sections were similarly indirect; they aimed to ward off arguments for absolute undecidability, and could only give peripheral support for Gödel's rationalistic position. Hence, in rejecting those arguments, I have offered only a preliminary case for the existence of absolutely undecidable arithmetical propositions. My aim in the next few sections is to develop these considerations into a more positive argument for the existence of them.

In the previous chapter, we reached the conclusion that Gödelian anti-mechanism succeeds if and only if we have the ability to construct very specific sequences of notations for transfinite recursive ordinals. It now appears that Gödel's argument for the decidability of every arithmetical proposition also hinges on an ability to recognise such notations unrestrictedly, at least 'in principle' (which is to say, under conditions of extreme idealisation).

This observation lets us greatly sharpen Gödel's disjunctive conclusion. The crucial issue becomes whether we can, with respect to $\mathcal{O}$ (or to some equivalent notation

system) and under idealised conditions, select or enumerate notations for recursive ordinals in such a way that the arithmetical theories in a Feferman reflection progression indexed by these numbers are collectively complete with respect to true sentences of arithmetic.[15] Call the ability to do this the 'recursive ordinal selection ability'. Determining whether we possess this ability promises to be substantially more tractable than the issues with which we were originally faced.

There is a lack of clarity in the thesis that the human mind is a machine, and an equal lack of clarity in the claim that the mind *isn't* a machine. Due to this lack of clarity, the conclusion reached in the previous chapter was merely that anti-mechanism wasn't supported by the incompleteness theorems; certainly no definite conclusion on the general hypothesis of mechanism was reached. By contrast, the claim that we have the recursive ordinal selection ability is much more specific (although still open to multiple interpretations, as we shall see below). Furthermore, since our idealised mathematical abilities extend our *actual* mathematical abilities, we can focus the debate on whether or not our present abilities to construct ordinal notations should be taken to generalise in the required way when we remove the constraints of paper and time. This strikes me as a far less daunting task than determining the general nature of the human mind!

The revised disjunction can now be formulated as follows:

> (a) We have the recursive ordinal selection ability, and mathematical proof cannot be represented mechanically; or
>
> (b) We lack the recursive ordinal selection ability, and there are absolutely undecidable propositions

The advantage of this new disjunction is that it allows us to see precisely on what the previously unclear debates hinges. Despite that, it is true that something is lost in this formulation. Namely, it completely glosses over the wider debate outside of the arithmetical context. After all, Gödel for a time thought that there were absolutely undecidable propositions that *weren't* arithmetical in character. But we

[15] The relevant conditions of idealisation remain unchanged from the previous chapter: the idealised mathematician is assumed to have access to arbitrarily large, but finite, amounts of material, time, and energy in verifying the index of any theory in a progression in order to construct a sequence of the required kind.

shall return to the issue of set-theoretic propositions in later chapters.

For now, it's important to note that rationalistic optimism and the recursive ordinal selection ability are related by implication: if the former is true, then no arithmetical proposition is absolutely undecidable, in principle. And if there are no absolutely undecidable arithmetical propositions, then it follows that we have the recursive ordinal selection ability: Franzén (2004, p.191) has proved that there is a unary primitive recursive function f from sentences in $L_{\mathbf{Q}}$ to natural numbers such that ϕ is true iff $f(\phi) \in \mathcal{O}$. Hence if all arithmetical sentences are decidable by some means, we could simply read off the correct selection of ordinal notations via the use of this function. So, with respect to the arithmetical restriction of the disjunctive argument, we can resolve the debate entirely if we can show that human beings don't have the recursive ordinal selection ability.

I don't have an argument that Gödel's optimism is incoherent; in fact I don't think it's incoherent in the slightest. Rather, I'll argue that the evidence as it stands shows that we have no good reason to believe that we have the recursive ordinal selection ability, even in principle. For those of us who can't take the existence of such an ability on faith, the existence of an absolutely undecidable arithmetical proposition is made enormously probable by the evidence to be presented. And if there is such a proposition, it follows by *modus tollens* that rationalistic optimism is false.

2.6 Recursive Ordinal Selection

In light of the mathematical results alluded to in this chapter, the only route for the Gödelian to take in asserting the absolute decidability of every arithmetical proposition is to argue that we have the recursive ordinal selection ability. This is to argue that, under idealised conditions and with respect to a fixed coding system such as $\mathcal{O}$ or something equivalent, we can select notations for recursive ordinals such that the union of the theories in a Feferman reflection progression indexed by these notations is arithmetically complete. Having this ability is equivalent to having the ability to non-recursively enumerate the truths of arithmetic by the use of a reflection principle. There are two main senses in which we might have such an

ability, both of which will be discussed (and rejected) below.

Strong Recursive Ordinal Selection

Firstly, we might have the ability to select the correct ordinal notations in a *strong* sense. In this sense, our ability to select ordinal notations is a consequence of rationalistic optimism, and so this is the distinctively Gödelian route to take.[16] Recall that rationalistic optimism credits us with the ability to prove or refute any well-defined mathematical proposition. We can pick some branch, Δ, of a recursive progression of **PA** using Feferman reflection such that $\bigcup_{n\in\Delta:|n|<\omega^{\omega^2+1}} \mathbf{PA}_n$ is arithmetically complete, since such branches exist by Feferman's theorem. For any number then, the question '$n \in \Delta$?' is meaningful. Assuming rationalistic optimism then, we should be able to enumerate the members of Δ by enumerating the answers to '$n \in \Delta$?' for each n, and removing the numbers for which the answer is 'no'. Assuming rationalistic optimism then, we have the recursive ordinal selection ability almost trivially.

There are two central problems with this Gödelian view. We saw in the previous chapter that taking our arithmetical capabilities, even heavily idealised, to be modelled by a Feferman arithmetic involves crediting the idealised mathematician with knowledge of certain arithmetical truths which are required to correctly construct the arithmetically complete theory. But being able to deploy this theory was supposed to *explain* why the idealised mathematician had knowledge of those truths in the first place. So to whatever extent you have sympathy with the arguments of chapter 1, you should be hostile to the Gödelian view here, since in essence rationalistic optimism entails Gödelian anti-mechanism.

Moreover, this isn't all that's wrong with Gödel's view. Once we've seen that thinking about the recursive ordinal selection ability is just another way of thinking about the hypothesis that no arithmetical propositions are absolutely undecidable, it becomes apparent that the whole *structure* of this Gödelian position is problematic. This is because the 'explanation' of our ability to decide any arithmetical proposition proceeds in terms of a *prima facie* more contentious principle, namely that

[16] Not that this is the only route for denying the existence of absolutely undecidable propositions. We'll see another below. But this is the distinctively Gödelian argument which makes use of the truth of rationalistic optimism as a working hypothesis.

we have the recursive ordinal selection ability. That principle is, as I've argued, a consequence of rationalistic optimism. But that is of no argumentative support for the Gödelian, since rationalistic optimism really is *just* optimism: despite Gödel's insistence that its failure would constitute some kind of scandal to human reason, it is hard to see how anything other than faith could compel us to believe its truth.

To illustrate the problem, consider this extremely unconvincing argument that all arithmetical propositions are, in principle, decidable: all set-theoretic propositions are decidable, in principle; therefore, all arithmetical propositions are decidable, in principle. The problem with the argument is that its premise is *much* stronger than its conclusion, so it's almost trivial that the conclusion follows. Indeed, the only missing information is that all arithmetical propositions are expressible in a set-theoretic context.

Gödel's own argument may well be less egregious. However, it is more similar than we ought to be comfortable with. Given the failure of the evidence and irrationality arguments, all the Gödelian is left with is something like 'We have the recursive ordinal selection ability. Therefore all arithmetical propositions are in-principle decidable'. Feferman's theorem gives us the required suppressed premise that the recursive ordinal selection ability entails the decidability of all arithmetical propositions.

Franzén's function gives us reason to believe that the recursive ordinal selection ability is *equivalent*, in some sense, to the ability to prove any arithmetical truth, so this argument doesn't obviously deploy a stronger set-theoretic ability in order to explain a weaker arithmetical ability. But why does the rationalist think that we have the recursive ordinal selection ability in the first place? If we have it as a result of some more general ability to solve set-theoretic problems, then we've postulated a much more contentious ability in order to explain our arithmetical capacities, as in the egregious example. But the alternative is to postulate, without broader reference to set theory, that we simply have the unexplained ability to pick out the correct numbers from a notation system. But *that* is no explanation for our ability to prove any arithmetical truth, since it is simply another way of asserting it. So once it becomes clear that the selection ability is the key assumption in the Gödelian position, the argument looks to have a deeply unconvincing structure that is practically question-begging.

In summary, Gödel's position, though coherent, really has little to recommend it. We've been given no serious philosophical or mathematical reason to think that it's true. The recursive ordinal selection ability, as conceived of by the optimist, is a conceptual possibility and nothing more.

Weak Recursive Ordinal Selection

In this section, I want to discuss a weaker sense in which we might have the recursive ordinal selection ability, one that does not depend on the assumption of rationalistic optimism. The recursive ordinal selection ability requires that we can, in principle, select natural numbers such that the theories in a Feferman reflection progression based on **PA** indexed by those numbers are collectively arithmetically complete. The rationalistic optimist posits that, under a suitable idealisation, we can select or enumerate the indices of a branch in $\mathcal{O}$ such that the union of those theories in such a progression is arithmetically complete. The weaker position to be considered here makes no such claim, but nevertheless denies the existence of absolutely undecidable arithmetical propositions.

The advocate of the weak position does *not*, unlike the Gödelian, assert that given time and paper, the idealised mind can execute a non-recursive procedure. Furthermore, the weak position denies that there is some absolutely undecidable arithmetical proposition which expresses the consistency of our idealised arithmetical output.[17] The view therefore appears to lie somewhere between mechanism and rationalistic optimism.[18]

According to this intermediate position, we cannot execute a non-recursive procedure, even under the standard idealisation away from time, paper, and so on. Hence there are some arithmetical propositions which we cannot prove. Amongst such propositions will be consistency sentences of some theories in a Feferman reflection progression on **PA**. If such a sentence, $Con_{\mathbf{T}_n}$, cannot be proved, it is because we cannot recognize that $n \in \mathcal{O}$. After all, if we could recognize that fact, then we would know that $\mathbf{T}_n$ was an extension of **PA** by iterated Feferman reflection,

[17]The existence of such a proposition follows from the most straightforward expression of mechanism as the thesis that the idealised mathematical abilities of the human mind have an output coextensive with some Turing machine.

[18]Thanks to Tim Button for pressing me on the importance of this position.

and hence that it was sound. So we would also recognize that $\mathbf{T}_{2^n}$ is sound, and $\mathbf{T}_{2^n} \vdash Con_{\mathbf{T}_n}$. According to the weak position, it seems almost *accidental* that our ability to recognize ordinal notations gives out before n, rather than after n but before some m such that $n <_{\mathcal{O}} m$. The reason for this is that, although $\mathcal{O}$ is not recursively enumerable, predecessors within $\mathcal{O}$ *are* recursively enumerable. In other words, for any $b \in \mathcal{O}$, the set $\{a \mid a <_{\mathcal{O}} b\}$ is recursive. According to the weak position, this fact gives us good reason to think that under suitable idealisation, we *could* prove the undecidable sentence $Con_{\mathbf{T}_n}$.

The weak position holds that our mathematical abilities have the output of some recursive procedure. Hence, idealising to the extent that we can execute a non-recursive procedure is too far; the idealised beings in such a scenario are no longer representative of what is *humanly* provable, and hence are no longer relevant to debate about absolute undecidability. Nevertheless, since any notation in $\mathcal{O}$ has recursively enumerable predecessors, we *can*, for any $m >_{\mathcal{O}} n$, idealise to the extent that it is recognizable by us under that idealisation that $m \in \mathcal{O}$. Hence, under idealisation, we can recognize that $2^n \in \mathcal{O}$, and hence that $Con_{\mathbf{T}_n}$ is true. So there are no absolutely undecidable arithmetical propositions, since any proposition which is a plausible candidate for absolute undecidability is provable by us under some acceptable idealisation of our abilities. The indices of the theories required by such idealisations constitute a selection of notations satisfying the hypothesis that we possess the recursive ordinal selection ability.

The weak position essentially reverses the order of the quantifiers in the rationalistic optimist's thesis: the latter claims that, under some idealisation, we can prove every arithmetical truth, while that former claims that every arithmetical truth is provable by us under some-or-another idealisation. If this is correct, then we have the recursive ordinal selection ability not because we can, under idealisation select each of the required notations, but because each of the required notations can, under some idealisation, be selected by us.

Despite the intuitive appeal of such a position, it actually rests on optimism just as much as the original Gödelian position did. To see why, we need to re-examine the idea that any limitation on our ability to recognize ordinal notations would be 'accidental', insofar as any such notation could be recognized by us under a natural idealization on our current abilities. This claim rested crucially on the recursive enu-

merability of predecessors in $\mathcal{O}$, but overlooks the important fact that the ordering $<_{\mathcal{O}}$ is only partial. Suppose we accept the weak position's claim that our current arithmetical abilities have an output coextensive with some recursively enumerable set, the closure of the theory $\mathbf{T}_n$. The intermediate position claims for any m such that $n <_{\mathcal{O}} m$, under some idealisation our abilities correspond to $\mathbf{T}_m$. The claim has some intuitive appeal, but it does *not* entail that any notation is recognizable as such by us under some idealisation or another. Since $<_{\mathcal{O}}$ is only partial, all we can suppose is that any notation *which lies on a path that includes* n is recognizable by us under some idealisation.

Even if all of that is true, it does *not* entail that we have the recursive ordinal selection ability, thanks to the Feferman–Spector theorem, which we encountered in the previous chapter:

> **Feferman–Spector Theorem:** There are paths Z through $\mathcal{O}$ that constitute a notation for every ordinal $< \omega_1^{CK}$, such that $\bigcup_{n \in Z} \mathbf{T}_n$ is incomplete with respect to the Π_1 sentences (Feferman and Spector 1962, p.384). Moreover, there are $\aleph_0$ such paths (1962, p.389).

Why is this so bad for the weak position? Suppose we accept that our current abilities correspond to $\mathbf{T}_n$ and that any notation which lies on a path that includes n is recognizable by us under some idealisation. This means that for any ordinal less that ω_1^{CK}, we can recognize some notation for it. What the Feferman–Spector theorem tells us is that being able to give a notation for every recursive ordinal by enumerating a path through $\mathcal{O}$ is not on its own sufficient to generate an arithmetically complete theory by iterated Feferman reflection on **PA**. This is because there are paths through $\mathcal{O}$ such that the theories in a Feferman reflection progression which are indexed to that path are collectively incomplete.

Hence the key premises of the intermediate position could all be true, and yet we might lack the recursive ordinal selection ability. The thought was that any arithmetical proposition was provable by some natural idealization of our actual arithmetical abilities, because any ordinal notation could be recognized by us under some natural idealization. However, the fact that the relevant order on $\mathcal{O}$ is only partial means that it is only the notations *on some paths* that are recognizable by such means, and the Feferman–Spector theorem shows that being able to follow a

path through $\mathcal{O}$ does not entail having the ability to construct an arithmetically complete theory by taking the union of theories in a progression indexed to notations in that path.

Only certain paths through $\mathcal{O}$ are such that the union of theories in a Feferman reflection progression on **PA** indexed to that path are arithmetically complete. Hence, the weak position is only correct if, by some cosmic chance, the theory representing our current mathematical abilities happens to lie on one of these special paths. And there is simply no reason to suspect that this is the case.

So the weak position, much like its strong counterpart, is a bare conceptual possibility; there is no reason to think that either is true. For all its extra sophistication, the weak position rests on sheer optimism about our *present* arithmetical abilities, much as the Gödelian position requires enormous optimism about our *idealised* abilities.

2.7 Which Propositions are Absolutely Undecidable?

We've seen that two parties to the absolute undecidability debate, namely strong Gödelianism and a weaker position, are conceptual possibilities, but beyond this little can be said in their favour. Adopting them requires a certain kind of faith in our arithmetical capacities that is simply unwarranted by the evidence. Another option is to think that our arithmetical abilities are representable by the union of theories in an initial segment (or segments) of some branches of a reflection progression based on **PA**. We are capable of expanding our arithmetical abilities by going further in our iterative procedure of adding relevant instances of Feferman's reflection principle, so we can't assume that our abilities are representable in a fixed formal system.[19]

As the evidence currently stands, it would simply be a miracle if we have the recursive ordinal selection ability, even if we only have it 'in-principle'. If we are unwilling to countenance such a miracle, it follows that some arithmetical propo-

[19]It is however, *consistent* with this position that our abilities be so representable. A thorough discussion of this would take us too far afield, since it seems to me that Gödel's theorems can neither verify nor refute this kind of mechanism.

sitions are absolutely undecidable, namely those instances of Feferman reflection corresponding to theories indexed by numbers which we cannot recognize as denoting recursive ordinals, i.e. theories that we cannot recognize as sound by reflecting on the soundness of **PA**.

The question naturally follows: *which* arithmetical propositions are absolutely unprovable? Quite reasonably, one might want to see an example of such a sentence, and perhaps if appropriate a proof of its independence from a system that might represent our arithmetical capacities. The aim of this section is to give a principled excuse for my lack of an example, and gesture at some philosophical significance the necessary lack of an example might have. The reason is intimately related to the failure of Gödel's irrationality argument, as you might expect.

As argued above, Gödel's evidence argument is sound at least as far as finite iterations of reflection principles are concerned. The problem which comes to the fore in extending this argument is that there is no general method for recognizing ordinal notations within a given system (e.g. Kleene's $\mathcal{O}$), and we are obliged to use such a system since, since the theorems of $\mathbf{PA}_\omega$ can be enumerated by a Turing machine. Since we cannot iteratively add instances of Feferman reflection to **PA** in such a way that the union of the constructed theories is arithmetically complete, it follows that the axioms of some theory in our progression are *not* exactly as evident as the axioms of **PA**. But this fact means that neither I, nor anyone else, can exhibit an absolutely unprovable arithmetical truth of this kind.

As was argued in the discussion of Gödel's irrationality argument, on the picture I'm proposing, there could be no recognizable example of an absolutely undecidable arithmetical proposition, for the reason that if a true instance of Feferman reflection *is* absolutely undecidable then we cannot recognize it as asserting a soundness property of a theory which extends **PA** by iterated reflection, and which is therefore sound. If we could so recognize it, we would have sufficient justification for the soundness of a theory which trivially decides it. But if we cannot recognize it, then there is no reason to think that it is true, and so no reason to think that it expresses a true but unprovable arithmetical proposition. Either way, such a proposition could not be the kind of example one might reasonably ask for.

As I've explained it, the reason that there are absolutely undecidable true propositions of arithmetic is, to speak somewhat metaphorically, because we lose our grip

on whether a set of sentences is an axiomatization of an extension of **PA** by iterated reflection when we cannot verify that such a theory is indexed by a suitable notation or not. Any putative instance of an absolutely undecidable arithmetical proposition will present a theory and a reflection principle for it. If we can recognize that the theory is of the required kind, then reasoning just rehearsed will show that the proposition is decidable in some stronger theory the axioms of which are exactly as evident as those of **PA**. So I can't give a *counterexample* to Gödel's rationalistic faith about arithmetical propositions, because if a proposition is a recognizable counterexample, then it is not a counterexample after all. But if my arguments are sound, then the evidence overwhelmingly supports the existence of *some* unrecognizable example of a true but absolutely undecidable arithmetical proposition.

That concludes the argument for the existence of absolutely undecidable arithmetical propositions. But some light can be shed on the implications of this view by considering another major philosophical issue related to Gödel's theorems: *indefinite extensibility*.

2.8 Indefinite Extensibility

The arguments of the previous chapter have focused principally on epistemic issues related to recursive axiomatization and the recursive ordinals. I've argued that both have a principle role in addressing Gödel's disjunctive argument, because a sensible epistemology of arithmetic is sufficient to defuse Gödelian anti-mechanism and establish the existence of absolutely undecidable arithmetical propositions. It would therefore be remiss not to apply these considerations to the other central philosophical issue in the vicinity of Gödel's theorem, namely Dummett's argument for arithmetical indefinite extensibility.[20]

The central text on indefinite extensibility in relation to the natural numbers is (Dummett 1963). In that paper, the concern is to defend the working idea that the meaning of an expression is its use against the claim that Gödel's theorem shows that we have an inner perception of the intended interpretation of **PA** that cannot

[20]Indeed we can expect the preceding arguments to be relevant, for as Wright argues (1994, p.175) indefinite extensibility and anti-mechanism are intimately related positions in the debate.

be fully articulated in a formal theory. Although we'll return to the issue of intuition of mathematical objects in the next chapter, I think Dummett is essentially correct that the rival position fails because the notion of a model is not something that we can grasp independently of model theory (Dummett 1963, p.191). In other words, models are not the sort of thing that are independent of their description. I'll not quibble with Dummett on this issue, nor am I concerned to attack the programmatic identification of meaning with use. Rather, it is with the specific use to which indefinite extensibility is put in the paper that I shall take issue.

In particular, Dummett uses considerations related to indefinite extensibility to argue that the concept *natural number* is, in some sense, vague. This scandalous conclusion has its basis in three claims: firstly, that the concept *natural number* determines not only its extension, but *also* the grounds for asserting something to be true of the natural numbers. Secondly, Dummett takes it that the notion of a ground for asserting something to be true of all natural numbers is *indefinitely extensible*. Finally, he takes indefinite extensibility to be a species of vagueness. Hence the conclusion that the concept *natural number* is vague.[21] I'll argue that Dummett is mistaken, even in his own terms, to think of the grounds for asserting something to be true of all natural numbers as indefinitely extensible. In other words, I'll take Dummett's first and third claims for granted, and show that the second is false. Hence, Dummett is mistaken, even in his own terms, to conclude that the concept *natural number* is vague.

In (Dummett 1963), we are told that a concept is indefinitely extensible if 'for any definite characterisation of it, there is a natural extension of this characterisation, which yields a more inclusive concept' (1963, p.195). Dummett's archetype of an indefinitely extensible concept is the concept *ordinal*, which is supposed to be indefinitely extensible for the following reason: let **O** be the extension of some definite characterization of the concept *ordinal*. These order-types of well-orderings are

[21]There is a distinction to be drawn between the indefinite extensibility of the concept *natural number* and the indefinite extensibility of the concept *ground for asserting something about all natural numbers*. In (1963), as well as the follow-up paper (1994), Dummett does not assert that the concept *natural number* is indefinitely extensible itself (i.e. with respect to its extension), nor does he think that there might be 'borderline cases' or other indeterminacy in the application of the concept (1994, p.336). However, as Oliver emphasises (1998, p.28), certain later time-slices of Dummett *do* argue for the indefinite extensibility of the concept *natural number*. I shan't be concerned with such arguments here.

themselves naturally well-ordered, and this well-order must itself have some order type, Ω. It follows by some basic set theory that the order type of the sequence of all ordinals less than α is α itself. By the well-ordering of ordinals, it is follows that $\Omega \not< \Omega$, and hence that $\Omega \notin \mathbf{O}$. We can then form some more extensive characterization of the ordinals with the extension $\mathbf{O} \cup \{\Omega\}$.

This method of generating new ordinals is 'natural' in that it works in complete generality; we can, according to Dummett, apply it to any well-defined totality of ordinals. I won't take a stand here on the validity of Dummett's argument. All that's important for my purposes is that Dummett takes to be indefinitely extensible those concepts that are like the concept *ordinal* in this respect.

Dummett argues that the concept of *ground for asserting something about all natural numbers* is also indefinitely extensible in the sense described (1994, p.336). The reason is that the natural numbers are intimately related to their characterization as a totality over which induction is unrestrictedly valid. Induction is taken by Dummett to be a means of asserting statements about *all* natural numbers with respect to any 'well-defined' property. Indefinite extensibility appears because the notion of 'well-defined arithmetical property' is itself indefinitely extensible (1963, p.196).

Dummett's argument for the latter principle is as follows: **PA** cannot prove $Con_{\mathbf{PA}}$. However, if **PA** is sound, i.e. if all its axioms are true, then it is consistent. Hence, whatever justification we had for believing **PA** to be sound is just as good a basis for believing that $\mathbf{PA} + Con_{\mathbf{PA}}$ is sound. In this stronger system, we now define the property of being 'true-in-**PA**', and this new property can feature in the induction scheme (1963, p.195). This means we have a new ground for asserting something about all natural numbers. Since our stronger system is itself a sound extension of **PA**, it is again incomplete, and the process of forming new well-defined properties of numbers can be continued indefinitely. Because all such properties can be used for induction, any well-defined totality of grounds for asserting something about all natural numbers can, by the uniform means of adding consistency sentences to the defining theory, be extended to form a more inclusive totality of such grounds. All that is required is the insight that **PA** is itself sound. Hence the concept of *ground for asserting something about all natural numbers* is indefinitely

extensible, just as was the concept *ordinal*.[22]

As I mentioned above, Dummett takes this to constitute vagueness in the concept *natural number*. I don't know how to argue with this terminological decision, but I do think Dummett's argument misfires. The key to seeing why is to examine what is meant by 'indefinitely'.

2.9 Infinite Extensibility

In his discussion of the issue, Moore argues that with respect to any concept, not just those that Dummett marks off as indefinitely extensible, we have no hope of delimiting in advance what its possible usage and applications might be (1998, p.119). With this is mind, I don't want to be read as making any grand claims about induction over the natural numbers in general. However, I do think that Dummett's claim that the notion of a well-defined property is indefinitely extensible thanks to Gödel's theorem is straightforwardly false.

Taking the ordinals as our paradigm case of indefinite extensibility, it seems clear what 'indefinite' amounts to: the most general means we have of formulating principles relating to the length of a sequence are set-theoretic. Hence the sequence of ordinals, if it can be thought of as even having a length, is longer than any ordinal. This is demonstrated by the general method exhibited of forming an ordinal greater than any member of any putative set of all ordinals. A consequence is that we have no means of taking the ordinals as a 'definite totality' — there will always be another ordinal lurking outside any totality we might circumscribe.

The mention of a standard method of extending the initial concept is required for indefinite extensibility; as Dummett puts it 'this extension will be made according to some general principle for generating such extensions, and, typically, the extended characterisation will be formulated by reference to the previous, unextended, characterisation' (1963, pp.195–196). With respect to this process of extension, the contrast between the ordinals and the notion of an arithmetical proof is stark. Play-

[22]Dummett's original argument uses the iterated addition of the Gödel sentence of a theory. The matter is simplified without any substantive change in the argument by considering reflection on consistency.

ing the role of the standard ordinal construction in the arithmetical case is of course the application of a reflection principle.

In particular, Dummett argues that our base theory is **PA**, and that reflection upon this theory proceeds by the iterated addition of canonical consistency sentences. This leads to a sequence of theories, such that $\mathbf{T}_0 = \mathbf{PA}$ and $\mathbf{T}_{n+1} = \mathbf{T}_n + Con_{\mathbf{T}_n}$. Each time we apply the consistency reflection principle, our theory becomes more powerful than before. Dummett cautions that 'there is no ground for recourse to the conception of a mythical limit to the process of extension' (1963, p.198), and concludes that the ground for asserting something true of all numbers is indefinitely extensible.

The problem, however, is that the limit isn't a myth! According to Dummett, the application of consistency reflection is justified because the axioms of **PA** are known to be sound. For the same reason, we can apply Feferman's reflection principle.[23] That principle is plainly valid with respect to a sound arithmetical theory, no matter what the formula ϕ is: the antecedent of the principle is that, for all objects x, the predicate which codes provability from the axioms of $\mathbf{T}_\alpha$ holds of the number coding the proposition that $\phi(x)$. If $\mathbf{T}_\alpha$ is sound, then $Pr_{\mathbf{T}_\alpha}(\overline{\Psi})$ implies that Ψ. Assuming the antecedent then, we then have it that for all objects x, $\phi(x)$, which is the consequent of Feferman reflection, as required.

Even Dummett's constructivist leanings do not alter the state of play here. By the hypothesis that $\forall x\ Pr_{\mathbf{T}_\alpha}(\overline{\phi}(\dot{\overline{x}}))$, we have a means of proving, when presented with any n, that $\phi(n)$ (assuming that $\mathbf{T}_\alpha$ is sound). On a standard Brower-Heyting-Kolmogorov understanding of universal quantification, we therefore have a proof that $\forall x\ \phi(x)$. That little argument is itself a construction transforming a proof of $\forall x\ Pr_{\mathbf{T}_\alpha}(\overline{\phi}(\dot{\overline{x}}))$ into a proof of $\forall x\ \phi(x)$, and hence on the BHK interpretation of the conditional, we have a proof (though of course not in $\mathbf{T}_\alpha$ itself) that $\forall x\ Pr_{\mathbf{T}_\alpha}(\overline{\phi}(\dot{\overline{x}})) \rightarrow \forall x\ \phi(x)$. And *that* argument for the conditional claim works for any formula ϕ, constituting a proof of the Feferman reflection schema for $\mathbf{T}_\alpha$. The whole argument just given relied only the assumption that $\mathbf{T}_\alpha$ was sound, and

[23]Recall that Feferman reflection states that if, for all numbers, a theory proves the relevant instance of a formula, then that formula is true of all numbers. Formally: $\forall x\ Pr_{\mathbf{T}_\alpha}(\overline{\phi}(\dot{\overline{x}})) \rightarrow \forall x\ \phi(x)$ where $Pr_{\mathbf{T}_\alpha}$ is a provability predicate for $\mathbf{T}_\alpha$ coded in the standard Gödelian fashion; and where $\overline{\phi}(\dot{\overline{x}})$ denotes the number coding the result of substituting the numeral denoting x for the first variable appearing in ϕ.

hence even the intuitionist should accept that Feferman reflection holds for any sound arithmetical theory. This is critical to my argument against Dummett, since it allows for the deployment of Feferman's completeness theorem.

In consequence, and contrary to Dummett's claim, there *is* a limit to the process of extending our grounds for asserting truths about the natural numbers: having taking the union of theories in a Feferman reflection progression on **PA** indexed by notations for ordinals $< \omega^{\omega^{\omega}}$, there are no more truths in the language of arithmetic to be proved 'further up', not even instances of the induction schema. After such a series of applications of the reflection principle, we are left with an arithmetically complete theory. Although reaching this limit requires infinitely many applications of the reflection principle, the ordinal is *tiny*, even by the standards of countable ordinals.

We might, therefore, want to say that the notion of arithmetical proof is *infinitely* extensible, but it isn't *indefinitely* extensible, if that is to imply anything like a structure analogous to the sequence of ordinals.

2.10 Dummett on Feferman's Theorem

The original paper by Dummett appears only a year after the proof of Feferman's theorem, so it's no surprise that the former paper shows a lack of awareness of the result. In a later paper (1994) Dummett specifically discusses Feferman's theorem, albeit briefly. There, he claims that for purposes of considering indefinite extensibility, we should only be interested in *autonomous* progressions.

Recall that in each branch of an autonomous progression, one can only advance to a stage $\mathbf{T}_n$ if at some previous stage there is a proof that the coding has been successful, i.e. that n, a natural number, really does code $|n|$, an ordinal, in the system we've set up. Such proofs are not in general available in progressions as we've discussed them. Every autonomous progression on **PA** is incomplete and does not have members indexed to every recursive ordinal.[24]

[24]These ordinals are also known as 'constructive ordinals' or 'computable ordinals'. Nothing substantive hinges on the choice of name: these ordinals are exactly those which have a notation in $\mathcal{O}$, and this is determined independently of the relationship between recursiveness and computability. However, since an intuitionist like Dummett is likely to consider the name 'computable

Dummett glosses the autonomy condition on progressions as 'the formal equivalent of the requirement that we can recognize the axioms of the theory to be true' (1994, p.337). Dummett quite rightly observes that no branch of any such progression can represent all the recursive ordinals, and though the point is not made explicit, I assume that this is the reason he does not take his argument for indefinite extensibility to be under threat: since no autonomous progression exhausts the recursive ordinals, we can presumably keep extending each one in a way that vindicates Dummett's argument.

Dummett's gloss on the concept of an autonomous progression is somewhat nonstandard. A more neutral gloss would be that an autonomous progression is such that we can recognize it at all points to be indexed by a recursive ordinal. In the case under consideration, *if* we can recognize the theory as belonging to the progression, it is trivial that is sound. But that is only because **PA** is sound as well. The autonomous iteration of reflection principles is more commonly used to capture what we are committed to when we accept a theory, rather than what we can recognize as true on the basis of a theory (Feferman 1998, §4), so this characterization may seem unusual. I suspect, however, that this is harmless, since by using Franzén's function (discussed previously), if we can recognize that the theory is sound at each stage, then we can recognize that it is indexed by an ordinal notation and vice-versa. Despite that, there are two problems with Dummett's argument here.

First, the autonomy of a progression only guarantees that at each stage we can recognise the current stage as indexed by a notation for an ordinal *by means available within the theory itself so far*. But in determining what an acceptable ground for asserting truths about all natural numbers is, we *have to* go beyond what is available in the formal system so far. Dummett's own argument crucially relies on recognizing that **PA** is sound, and hence consistent, which of course requires resources that go beyond those available in **PA**. Indeed, at *every* stage of the progression, we are required to use more resources than we have available in the formal theory, since no stage of the progression proves its own canonical consistency sentence.

Even if we restrict our attention to autonomous progressions, we must still use

ordinal' as potentially misleading, it's worth bearing in mind that the argument can be run entirely in the more neutral terminology of 'constructive ordinals'. Thanks to an anonymous referee from *Philosophia Mathematica* for highlighting the need for clarification here.

resources external to the theory to satisfy Dummett. Suppose that we have some $\mathbf{T}_n$ such that for some $m <_{\mathcal{O}} n$ (where $<_{\mathcal{O}}$ is the ordering on notations induced by our coding system), $\mathbf{T}_m \vdash \Omega(n)$, where $\Omega(n)$ iff $n \in \mathcal{O}$ (i.e. iff n codes an ordinal). This does *not* tell us that $\mathbf{T}_n$ is sound, as Dummett requires. Rather it tells us that $\mathbf{T}_n$ extends $\mathbf{T}_m$ by the iterated addition of consistency sentences, which only amounts to the soundness of $\mathbf{T}_n$ if $\mathbf{T}_m$ is itself sound, which of course cannot be proved in $\mathbf{T}_m$ itself.

No argument to my knowledge has been offered as to why deploying information from outside the theory at any given stage is legitimate if that information is about the soundness of members of the progression, but illegitimate if the information is about the indices of members of the progression. This looks especially suspicious in light of the observation that recognizing the soundness of a member of the progression at transfinite stages *requires* recognizing that it is indexed by a notation for a recursive ordinal.

Second, even if we had some story about why we should only care about autonomous progressions in the course of determining whether the concept *natural number* is indefinitely extensible, Dummett would still be wrong.[25] The reason is that, although the indices of the members of an autonomous progression don't represent *all* recursive ordinals, these progressions still have a well-defined ordinal length, since all autonomous progressions break off at a recursive ordinal (Feferman 1998). Hence all such progressions are of length $< \omega_1^{CK}$, the first non-recursive ordinal. Because of this, the analogy between the concept under consideration and the sequence of ordinals breaks down completely, since there is no corresponding bound on the length of sequences of ordinals.

In short, Dummett's defence of indefinite extensibility in this case rests on an autonomy restriction which cannot justifiably be imposed in this context, and wouldn't entail the indefinite extensibility of the relevant concept even if it could be. The structure of grounds for asserting something of all natural numbers is simply *not like* the structure of the ordinals in the relevant sense.

[25]This is not to suggest that we *shouldn't* care about autonomous progressions. Autonomous progressions were developed by Feferman, Kreisel, and others in order to think about the commitments we incur from accepting a given theory, an endeavour which is surely worthwhile. All I mean to suggest is that we don't have a story about why we should care exclusively about such progressions in the debate about indefinite extensibility.

2.11 Responses on Dummett's Behalf

A very simple way of stating the problem with Dummett's argument is as follows: for a concept to be indefinitely extensible is for any definite totality of things falling under it to yield, by uniform means, a more inclusive totality of such things, typically by reference to the previous unextended characterization of the totality.[26] Feferman's theorem shows us that, in the case of extending our grounds for asserting something of all numbers there is a limit to how far the process can go.[27]

Dummett's own response is unsatisfactory, but there are two responses available on his behalf. A means for Dummett to defuse my second argument in the previous section is to re-interpret the relevant mathematics so that it doesn't have the philosophical implications which cause trouble for his indefinite extensibility argument. Wright (1985, pp.133–134) argues that the proof of Cantor's Diagonal argument, which is standardly taken to show that there is no enumeration of the sets of natural numbers, can be instead be interpreted as a proof that there is no recursive enumeration of the recursively enumerable sequences of numerals. This is a strategy that a Dummettian might be sympathetic to in order to avoid a commitment to uncountable cardinalities. A similar strategy might well be used to avoid a commitment to non-recursive ordinals — interpret any proof of their existence as a limitative theorem on what is recursively enumerable.[28] In this case the analogy with the ordinals might survive. If Dummett can avoid a commitment to ω_1^{CK} and larger ordinals, there is no reason to regard the limit on the length of progressions as a genuine ordinal value. Thus, if the Dummettian can avoid thinking that ω_1^{CK} really is an ordinal, then the argument that the concept *natural number* is indefi-

[26]In Dummett's later arguments, the 'typical' condition is more strongly insisted upon (Dummett 1991, p.318). In our particular case, given that the new system of grounds for asserting something of all natural numbers is the result of defining a truth predicate for the old system of grounds, I take it that Dummett's more stringent account doesn't change the state of play.

[27]I've argued in this chapter that we in fact can't apply Feferman reflection in such a way as to reach an arithmetically complete theory. But of course Dummett is not committed to this view. In deploying Feferman's theorem here, I am trying to show that Dummett is wrong in his own terms; I'm *not* trying to show that we actually can reflect on the soundness of **PA** in such a way as to prove every arithmetical truth.

[28]Thanks to Tim Button for making the connection between Wright's work and Dummett's here, and for offering the argument on Dummett's behalf.

nitely extensible might be salvaged.[29]

I think this is sufficient to defuse the second argument of the previous section; I doubt I could convince a sceptic that ω_1^{CK} is an ordinal any more than I could convince a finitist that $\aleph_0$ is a cardinal. But that isn't a dialectical position that I find especially worrying. It's worth noting that ω_1^{CK} is a *very small* ordinal; indeed it is countable, falling well short even of ω_1. So this proposal would severely reduce the strength of set theory in a way that few mathematicians or philosophers could accept.

The ability of the Dummettian to defuse my second argument reveals the crucial role played by Feferman's theorem in this context. Since the goal is to show that Dummett is wrong *by his own lights*, no argument against indefinite extensibility requiring ordinal numbers not acceptable to a constructivist of the appropriate kind will do, and the second argument offered *did* rely on considering ω_1^{CK} as an ordinal. By contrast, the ordinal bound provided by Feferman's theorem is intuitionistically acceptable, and hence the argument presented here has a virtue that a simple cardinality argument would lack.[30] We might be inclined, for instance, to argue that there are $\aleph_0$ true arithmetical sentences, but $\aleph_1$ countable ordinals, and hence that the notion of arithmetical proof couldn't possibly be extendible to the same extent as the ordinals. But such an argument needn't persuade a Dummettian in light of Wright's technique for re-interpreting theorems proving the existence of uncountable cardinals.[31]

By contrast, the bound given by Feferman's theorem is so small that *even* if we equate the sequence of ordinals with the sequence of recursive ordinals, the concept *natural number* is still not indefinitely extensible. Showing that it *is* would require an argument that in extending our grounds for asserting something of all numbers

[29]As an anonymous referee pointed out to me, Church's thesis 'is not particularly plausible from the intuitionistic standpoint' (Dummett 2000, p.186). Hence, there may be non-recursive ordinals (i.e. ordinals $\geq \omega_1^{CK}$) which are in some sense computable for the intuitionist. It is perhaps unlikely therefore, that Dummett would be willing to pursue the strategy outlined in order to defuse the argument. If so, then things are looking all the worse for Dummett.

[30]There is a further question about whether the proof of Feferman's theorem is intuitionistically acceptable, which is complicated by the open-endedness of the intuitionistic concept of proof. Notably, Dummett does *not* raise the issue in his discussion of Feferman's reflection techniques.

[31]Even in rejecting Church's thesis, a Dummettian might well be hesitant to regard uncountable ordinals as effectively constructible.

we are restricted to using autonomous progressions of **PA**. But we've already seen that no stable position can supply this argument, due to Dummett's need for information at each stage of the progression relating to the soundness of extensions of **PA** by the iterated addition of consistency sentences.

A second option for Dummett is to claim that a concept can only be described as indefinitely extensible 'in its own terms'. The concept *ordinal* is indefinitely extensible (according to Dummett) because any defined totality can be extended by means admissible in the general theory of ordinals. Similarly, one might think the concept of a natural number is indefinitely extensible 'in its own terms'; the argument I've put forward against this makes ineliminable use of the concept *recursive ordinal*, a concept which does not properly belong to arithmetic. Hence, Dummett's argument might escape unscathed.[32]

This argument strikes me as deeply suspicious for two reasons. Firstly, Dummett's own position relies on coding, in the language of arithmetic, concepts that are difficult to see as 'properly' arithmetical. In particular, several concepts from logic and the theory of syntax are coded into arithmetic in order to prove Gödel's theorem, and appear again in the formulation and application of consistency reflection which is required for Dummett's own argument to succeed.[33] It is simply *ad hoc* to insist that syntactic and logical concepts are legitimately arithmetizable here, but that information about the recursive ordinals is not so.

Moreover, the resulting account of the concept *ground for asserting something of all natural numbers* would be deeply unsatisfying. If the use of ordinal notations is excluded from theories which provide a ground for accepting truths about the natural numbers, then even very well-behaved theories like $\mathbf{PA}_\omega$ are not to be considered as giving us proper justification for their arithmetical consequences. If this theory did give us acceptable grounds for asserting something of all natural

[32]Put otherwise, Dummett's advocate might say that the argument for the indefinite extensibility of the concept *ordinal* is sufficiently *pure*, but the argument for the non-indefinite extensibility of the concept *natural number* is insufficiently so. This idea bears some resemblance to an idea presented by Shapiro and Wright, who consider concepts which are indefinitely extensible *relative* to some further concept. For example, the concept *real number* might be said to be indefinitely extensible relative to the concept *countable*, since any countable set of reals can be extended via a diagonal construction (Shapiro and Wright 2006, pp.266–267).

[33]For a fuller account of why these concepts go beyond what can be regarded as arithmetical in the relevant sense, see (Isaacson 1987).

numbers, then we ought to be able to reflect on its soundness; but we can't reflect on its soundness without the use of a notation denoting ω. This is a problem because $\mathbf{PA}_\omega$ merely extends, by an evidently valid reflection principle, the axioms of theories which, according to the view under consideration, *do* provide acceptable arithmetical proofs. Moreover, it can still be recursively axiomatized, just like the theories it is constructed from. So it is again *ad hoc* to exclude certain theories as illegitimate when no illegitimate property of those theories can even be gestured toward.

In conclusion, the concept *grounds for asserting something about all natural numbers* is not indefinitely extensible. More significantly, the concept *natural number* is free from whatever kind of vagueness might infect the concept *ordinal*.

Conclusion

In light of results in recursion theory, it seems untenable to claim that the concept of a ground for asserting something about numbers is indefinitely extensible. Consequently, any reason for taking the concept *natural number* to be vague, or open-ended, is undercut. The arguments I've made about indefinite extensibility are intimately related to the discussion of Gödel's evidence argument earlier in the chapter. Namely, the concept of an intuitively acceptable arithmetical proof is extensible to exactly the extent that the axioms of extensions of **PA** by the iterated addition of Feferman reflection are as evident as the axioms of the base theory, **PA**.

Given that no example of a true but absolutely undecidable arithmetical proposition can be exhibited, we are left with a peculiar species of quietism about the limits of arithmetical knowledge. The arguments of this and the previous chapter show that a rejection of Gödelian anti-mechanism and rationalistic optimism is required. Two important corollaries follow: the existence of absolutely undecidable arithmetical propositions, and the collapse of Dummett's analogy between ordinals and arithmetical proof.

This combination of views has perfectly coherent articulations, some of which are mechanistic. Benacerraf, in his discussions of the Lucas-Penrose argument (1967), notes the possibility of a position (later endorsed by Smith (2013, pp.281–283)) that

might best be called 'mechanistic quietism'. According to this position, it is consistent with the arguments given in favour of anti-mechanism that our arithmetical capabilities can be perfectly mimicked by a Turing machine, but that we don't have the ability to recognize the machine when presented with it. As mentioned above (fn.19), I'm happy to remain silent on whether this kind of mechanism is true. Perhaps the union of theories which we can reach by iterated reflection is itself recursively axiomatized, and hence associated with some Turing machine which could be taken as the definitive model of our idealised arithmetical capacities; though of course definitively specifying which machine was such would be forever beyond our abilities. But then again, perhaps the union of theories we can reach by iterated reflection is *not* recursively axiomatizable. Benacerraf provides little reason to think otherwise (Smith goes slightly further and claims that the ability to spot which machine enumerates my idealised output would be 'godlike' (2013, pp.281–282)).

Regardless of the status of mechanism, in embracing the notion that the union of theories in an initial segment (or segments) of some branches of a Feferman reflection progression on **PA** models our idealised arithmetical capacities, we've seen that a similar form of quietism is forced upon us: we can't precisely delimit our ability to recognise notations for recursive ordinals, so we can't give an example of an absolutely undecidable proposition. Moreover, it's not merely an epistemic issue; rather the very idea of exhibiting such a proposition doesn't make sense. Even if we don't embrace the mechanistic element of Benacerraf and Smith's view, we should at least acquiesce in its quietism.

There is a decent positive story to tell about why our output could be represented by the union of theories in an initial segment (or segments) of some branches of a transfinite reflection progression on **PA**, based on our knowledge that **PA** is sound together with the observation that $\mathbf{PA}_\omega$ is recursively axiomatizable. The additional point that the union of theories which we can recognize to be sound isn't arithmetically complete has been developed over the previous two chapters. With respect to arithmetical knowledge then, perhaps the significance of Gödel's theorem is best expressed as follows: the limits of our arithmetical knowledge cannot be exhibited.

Part B

Set-Theoretic Incompleteness

Chapter 3

Conceptual Platonism

Introduction

In addition to the disjunctive argument examined in the previous chapters, Gödel's Gibbs lecture contains a barrage of arguments that the incompleteness theorems support some kind of platonism. In light of the arguments in Gödel's drafts of *Is Mathematics Syntax of Language?*, a widely held view is that Gödel's theorems refute conventionalism, at least as it was understood in the 1950s. Another fairly common view is that the second incompleteness theorem is fatal for Hilbert's formalistic programme with respect to infinitary mathematics.

While I'm happy to agree that these anti-platonist views are made less credible by the incompleteness theorems, it is not so clear that the theorems can form the basis of an argument for a specific variety of platonism. The main reason is a severe lack of clarity in the nature of the position that they are supposed to be an argument for. Gödel claimed to have been a mathematical realist since 1925 (Wang 1987, pp.17–18), and was certainly a platonist at the time of his death in 1978. Five decades is a long time to hold a view, and Gödel's position was certainly not static. For example, mathematical intuition, a concept that has come to be intimately associated with Gödel, plays no role in his support for platonism in the major articles (1944), (1947), and (1951). But later, and especially in (1964), it takes on a leading role. It's further worth emphasising that much of the material we have attesting to Gödel's platonism is from reported conversation, rough notes for talks and lectures, and draft manuscripts. The result is that there is no one unique view that can be called 'Gödel's platonism'; there are only forms of platonism that are Gödelian in varying degrees.

In light of these considerations, it is difficult to know how to make progress in assessing Gödel's claims that the incompleteness theorems support platonism. That issue will be a focus of the next chapter; here I'll take up the more straightfor-

wardly interpretive task of getting clear on what Gödelian platonism involves, and in particular whether it deserves the allegations of mysticism that have been levelled against it. The methodology adopted here will be something between rational reconstruction and critical exegesis. Since there is no unique realist view attributable to the historical Gödel, I'll outline (§1–§2) a promising version of platonism that can be extracted from his various writings on the subject. I'll call this view *conceptual platonism*. Along the way, I'll also sketch what I take to be a Gödelian account of *mathematical intuition*, though we shall see that Gödel's use of this (already challenging) concept is both ambiguous and somewhat stretched. Nonetheless, there is some evidence that this reading gets Gödel right; in the next chapter we'll see that it vindicates his rather puzzling claim that our knowledge of small large cardinals (at least up to Mahlo cardinals) is founded on mathematical intuition.

In §3 I'll argue that Gödel's view, understood along these lines, involves no 'mysticism' and should be perfectly intelligible to an analytic philosopher of mathematics today. I'll end, in §4, with a discussion of an element of Gödel's platonism that is absent from the reconstruction presented here. I'll argue that Gödel's claims that some perception-like relation obtains between us and sets are largely independent from his more general view of intuition, and do not form a central pillar of his platonist perspective. This is particularly relevant to the material of the final chapter, which discusses Gödel's analogy between set theory and natural science, and his view on the non-intuitive (a.k.a. 'extrinsic') justification of larger large cardinal axioms.

3.1 Two Kinds of Platonism

With the possible exception of Dummett's work, it is often far from clear what realism with respect to a particular subject matter amounts to. The case of platonism in the philosophy of mathematics is no different, and in this context especially, realism is often taken to be characterised only by a series of platitudes, such as that mathematical objects 'really exist', that they are 'mind-independent', that they are 'acausal', 'abstract' etc. If taken as metaphors these platitudes are certainly suggestive; if taken literally then one might think they hardly amount to the articulation of a specific philosophical thesis.

Gödel himself is no stranger to these kinds of metaphors. In the Gibbs lecture he characterizes '[p]latonism or "realism" as to the mathematical objects' as the view 'that mathematical objects and facts (or at least *something* in them) exist objectively and independently of our mental acts and decisions' (1951, p.311). The statement is of course less than transparent, and further points of articulation (such as that platonism amounts to the view that mathematical objects are not located in the mind or the natural world (1951, p.312, fn.17)), are hardly more informative.

It is such remarks that lead critics like Chihara (1990, p.12), to label Gödel's platonism 'a kind of theology'. I think that the above remarks, as well some other popular examples, are intended by Gödel to be a kind of 'textbook' reminder of what platonism usually amounts to, rather than a specific thesis to be defended. Nonetheless it isn't clear just what kind of position Gödel *is* defending. The goal of this section is to try and articulate a distinctive kind of platonism that has its roots in Gödel's writings and that, I hope, is less vulnerable to charges of being metaphorical or theological. Although I'll argue that we must ultimately reject at least certain elements of Gödel's platonism (so construed), I think the result is both a reasonable reconstruction of his writing and the sort of position that might be found in mainstream analytic philosophy (though no doubt expressed in different terminology).

As the previous paragraphs indicate, platonism is typically a position formulated with respect to mathematical *objects*. It is these objects that are said to exist timelessly, mind-independently, and so on. A striking feature of Gödel's view is that his platonism is not primarily a kind of realism about mathematical objects like sets, but is rather a realism about (the content of) mathematical *concepts*, such as the concept *set*. At several places in his writings, Gödel makes it clear that he *is* a platonist about mathematical objects, in the more traditional sense, but it is clear from his remarks elsewhere that platonism about concepts is explanatorily prior in his view. Indeed at times he simply *equates* mathematical platonism with realism about mathematical concepts (1951, p.314), while at others he takes it that platonism about concepts *implies* platonism about objects:

> [I]f the meanings of the primitive terms of set theory as explained... are accepted as sound, it follows that the set-theoretical concepts and

> theorems describe some well-determined reality (Gödel 1964, p.260)

This remark makes it clear that Gödel's view is not a 'kind of theology', or at least that the central commitments are consistent with a more secular view of set theory. Indeed I think that in these remarks is the seed of a much more subtle philosophical position. Gödel's view is that a grasp of the concept *set* is the proper source of platonism about sets. And according to him, our grasp of the concept *set* is explained primarily in terms of our apprehending the truth of increasingly strong axiomatic theories of sets (1964, pp.260-261). Hence, what this platonism requires is an account of how we apprehend the truth of the axioms of set theory, and we should hope to be able to provide such an account without recourse to mystical means.

Viewed this way, Gödel's platonism is relatively minimalistic. In the context of arithmetic, Potter argues that 'Gödel thinks that once we have grasped the concept "natural number" there is nothing further involved in the claim that natural numbers exist, because he thought the concept "natural number" itself has a real content, namely the existence of the natural numbers themselves' (Potter 2001, p.343). Understanding Gödel's view this way makes clear an analogy between his platonism and Quine's; for a conceptual platonist it is quite proper to say that the existence of sets demands nothing more than the truth of sentences which existentially quantify over sets, perhaps the axioms and theorems of **ZFC**.[1]

Far from being theology then, conceptual platonism is the view that sets exist if and only if they are quantified over in sound axiomatic theories of the primitive term/concept *set*. No prayer or mystical insight is required to ascertain their existence, merely deduction from the axioms of set theory, and reflection on the concept *set*. According to the picture so sketched, the pressing philosophical question is then not how we have 'access' to causally inert abstract objects called 'sets', but rather how it is that we apprehend the truth of the axioms of set theory. In my view this

[1]There is some question as to whether first- or second-order **ZFC** should be meant here. Though a first-order formulation of the theory is standard today, there are good reasons to read Gödel as having second-order axioms in mind here. Firstly, Gödel's conception of set theory owes much to Zermelo's, whose system is most naturally taken to be second-order. Further, Gödel claims that set theory involves the primitive term *property of sets* (1947, p.181 fn17), which might suggest a second-order axiomatization with properties as the values of the second-order variables. Martin (2005, p.214) reads Gödel as meaning a second-order axiomatization; though he doesn't provide an argument, it's safe to assume he has something like these reasons in mind.

question promises to be much more tractable.

In pursing the justification of set-theoretic axioms, Gödel takes two relatively distinct routes. The former, which will be our concern for the bulk of the next chapter, focuses on what is known in the literature today as the *intrinsic* justification of axioms. The other route pursued is the justification of axioms by 'extrinsic' or quasi-scientific methods. Some of these are related to something akin to a 'perception' of sets, to be discussed in §4 of this chapter. The other considerations relate to less direct extrinsic justification based on an analogy Gödel draws between mathematics and natural science, discussion of which will be taken up in chapter 5. For now, I will flesh out the conceptual platonism at the heart of Gödel's view about intrinsically justified axioms, in particular the role that mathematical intuition plays in this account.

3.2 Mathematical Intuition

Perhaps even more so than for his platonism, Gödel is infamous for his belief in the existence of a faculty of mathematical intuition. As with platonism, it is not entirely clear what Gödel takes mathematical intuition to be, and certainly there is no well-developed *theory* of intuition in his writings. But there are (at least) two concepts at work that come under the heading of 'intuition', and disambiguating them is crucial to framing a plausible interpretation of Gödelian platonism.

As Gödel is a platonist about both sets and the concept *set*, there are correspondingly two notions of intuition, roughly corresponding to *intuition of* and *intuition that*. Roughly speaking, intuition *of* is directed toward mathematical objects, and intuition *that* is directed toward mathematical truths. Direct knowledge of mathematical objects is given by the former faculty; according to Gödel this objectual intuition is like 'a kind of perception' (1964, p.268) (though we'll see below that, in context, this remark is perhaps less substantial than it appears). Indeed, little beyond the analogy with perception is offered as to how such intuition is supposed to work. On the other hand, intuition *that* yields knowledge of mathematical axioms, and might be thought therefore to give indirect knowledge of mathematical objects via the medium of mathematical concepts.

The most compelling interpretation of Gödel is to think that intuition of is supposed to be roughly Kantian; in other words a singular representation to a thinking subject founded in perception or imagination. One reason for mistrust of the notion of intuition in Gödel's thought is that, so interpreted, it is utterly mysterious how such a faculty could do the work that Gödel ascribes to it. According to Kant, intuition founds arithmetic and geometry. The role it plays in Hilbert's philosophy is even more restricted, being used to account only for a 'finitary' fragment of arithmetic. The most prominent modern account of intuition is due to Parsons, and on his account objectual intuition does not even extend to natural numbers (2008, p.186), or to small finite sets (2008, p.214). By contrast, Gödel claims that mathematical intuition can deliver much stronger verdicts, such as the existence of Mahlo cardinals (1964, pp.260-261 and fn. 20).[2]

I think a lot of the confusion disappears if we take care to distinguish the role of *intuition that* in Gödel's philosophy. Gödel does not claim that we have intuition *of* large cardinals such as Mahlo cardinals; indeed it is hard to see how this could be possible if we understand intuition along broadly Kantian (and hence spatio-temporal) lines. Rather, Gödel claims we have an intuition *that* an axiom stating the existence of such a large cardinal is true (1964, pp.260–261). This is just as well, given that it is much easier to digest the idea that we have an intuition that an axiom stating the existence of Mahlo cardinals is true, than it is to give credit to the idea that we could have some quasi-perception of such a cardinal. Even the more palatable thought needs an argument of course; in the next chapter I'll outline such an argument on behalf of conceptual platonism.

Although distinguishing these two varieties of intuition is required to give a plausible interpretation of Gödel's platonism, it isn't always clear whether it is propositional or objectual intuition at work in his writings, in light of his platonism about concepts. For example Gödel speaks of 'an intuition which is sufficiently clear to produce the axioms of set theory' (1964, p.268). This intuition is perhaps supposed to be an intuition *that* such-and-such axioms are true. On the other hand, Gödel claims that the validity of axioms is a consequence of how things are with

[2] A cardinal κ is (strongly) inaccessible iff it is uncountable, regular, and such that $2^\lambda < \kappa$, for all $\lambda < \kappa$ (Jech 2003, p.58). A cardinal κ is (strongly) Mahlo iff the set of strongly inaccessible cardinals $< \kappa$ is stationary in κ (Kanamori 2009, p.21).

the relevant mathematical concepts, (1944, p.139 and 1951, p.321).[3] So perhaps the 'sufficiently clear' intuition referred to by Gödel is meant to be an an intuition *of* the set-theoteric concepts the facts about which entail the truth of the axioms.

Whichever interpretation was intended, it's clear enough that for Gödel, propositional intuition is delivered by the grasp we have of a concept. In sufficiently clear cases, this will deliver knowledge of the truth of axioms, some of which may contain quantification over objects such as sets. Although the term 'intuition' may sound unhappy to the modern ear, I think there's no mysticism involved in Gödel's commitments here. And as in the case of his platonism, Gödel's view here can indeed be interpreted as involving fairly minimal commitments. Thanks largely to the second incompleteness theorem, Gödel argues that mathematics cannot be the result of syntactical or semantic convention and stipulation; in other words he thinks that mathematics is importantly *non-trivial*, a view that few would take issue with today.

Potter (2001, p.340) argues that from this view, Gödel's terminological conventions with respect to thought make the existence of something fulfilling the role of mathematical intuition immediate. Gödel thinks that 'by our thinking, we cannot create any qualitatively new elements, but only reproduce and combine those that are given' (1964, p.268). If thinking is, by definition, essentially combinatorial, and mathematics is not reducible to combinatorics of syntax, then it follows that mathematical thought involves something 'extra', and *this*, whatever it may be, is called 'intuition'.

Parson's comes to a similar conclusion about the role of intuition in Gödel's thought, namely that 'the deliverances of mathematical intuition are just those mathematical propositions and inferences that we take to be evident on reflection and do not derive from others, or justify on a posteriori grounds, or explain away by a conventionalist strategy' (1995, p.59).

I think this minimalistic interpretation of Gödelian propositional intuition is essentially correct, and bears no resemblance to any kind of mystical insight. When considered as a source of evidence for the truth of mathematical propositions, intuition can be an imperfect tool. In particular, the credence intuition lends to a particular proposition need not be particularly strong, and in some cases can be

[3]In Gödel's terminology, the axioms are 'analytic'. See chapter 4 for details.

outright misleading. Gödel claims, for instance, that insufficiently clear intuition is responsible for the set-theoretic paradoxes (1951, p.321). He is of course confident that many such imperfect intuitions can be corrected over time; in a similar passage (1964, pp.267-268) he claims that the paradoxes 'are hardly any more troublesome for mathematics than deceptions of the senses are for physics'. But for Gödel, intuition is neither immediate nor does having an intuition that P entail that P is true. Parsons makes a related point in emphasising that there is a noticeable gap between acknowledging the existence of intuition and giving any credence to it (Parsons 1995, p.70).

Though there is no reason to buy into Gödel's terminological convention of calling only combinatorial thoughts 'thoughts', accepting that mathematics is non-trivial is almost sufficient for the existence of intuition in Gödel's sense. Given the explicit juxtaposition Gödel makes between intuitive sources of justification in set theory and those that we would today call 'extrinsic' or 'quasi-scientific' (1964, p.269), the only means of getting by without Gödelian intuition would be to claim that the soundness of *all* accepted mathematical axioms, together with any means of reducing incompleteness (such as the addition to a theory of its Gödel sentence) can be established exclusively by methods analogous to those employed in the sciences.

This is not the place to engage fully with this kind of radical empiricism about mathematics, but it's perhaps worth raising a point or two against it. Firstly, a primary use of intuition is the justification of consistency statements for sound formal theories. Even if extrinsic methods do have an important role to play in the epistemology of mathematics, it is implausible to claim that such methods are the only source of verification for the canonical consistency statement of **PA**, for instance, since the truth of that statement follows straightforwardly from the soundness of the theory. A second use for intuition is the verification of set-theoretic axioms, and again in this case it is implausible to claim that *all* commonly accepted axioms can be justified by exclusively extrinsic methods. Even if you think that quasi-scientific considerations justify set theory as a whole (e.g. the apparent indispensability of mathematical reasoning in the natural sciences), such considerations radically under-determine which set theory we should accept as sound. For example, even the standard set theory, **ZFC**, provides more sets than could be required for any known purposes of natural science. These considerations are not conclusive, but

they do serve at least to show quite how radical this kind of scientism really is.

If we reject this radical empiricism, then Potter and Parsons are correct to take propositional intuition in Gödel's thought as involving fairly minimal commitments: mathematical intuition is simply that which enables us to have objective, non-trivial, non-empirical mathematical knowledge. In the context of Gödel's platonism, in particular, mathematical intuition is simply the grasp of a mathematical concept. For intuition to yield knowledge of objects, the concept grasped must have the right kind of 'objective' content.[4] As Parsons highlights (1995, p.70), this doesn't entail anything as strong as Gödel's platonism. One could for example accept that we have intuition that the axioms of **PA** are true, owing to a grasp of the concept *number*, but deny that we have a similar grasp of the concept *set*, perhaps on grounds related to the paradoxes. But for those of us who don't find the concept *set* to be inherently flawed on such grounds, Gödel's account of intuition is simply that a grasp of that concept can yield mathematical knowledge without recourse to quasi-scientific methods.

Although the notion of intuition may appear to be somewhat arcane to analytic philosophers today, the idea that reflection on the concept *set* can yield set-theoretical knowledge independent of quasi-scientific considerations is relatively mainstream.[5] There are of course pressing philosophical issues in the immediate vicinity, such as how concepts are 'grasped', and what it is for an axiom (or anything else) to be true 'in virtue' of the nature of a concept. But these are issues that are pressing for a variety of philosophical positions, and the need for them even to be asked makes it clear, I think, that Gödel's view amounts to far more than the positing of some 'mysterious faculty', whatever his terminological decisions may at times suggest.

At the start of this chapter, I expressed scepticism of the idea that there was one unique theory that could be said to be Gödel's philosophy of mathematics. What is certain, however, is that he was committed to a platonistic account of mathematical

[4]Cases where a concept is not of the right kind might fail to yield mathematical knowledge in any non-trivial sense. An example of a flawed concept of this sort might be the concept of a Fregean extension.

[5]The works of Boolos (1971 and 1989) and Paseau (2007) are prominent philosophical examples of the idea. You're likely to also find a 'conceptual' argument for at least some of the axioms in any set theory textbook.

objects grounded in intuition. I've attempted to sketch a Gödelian theory which debunks the claims of mysticism surrounding Gödel's position, which are in part due to terminological choices made by Gödel. I'll call this position *conceptual platonism*. Its core commitments are:

1. **Platonism About Concepts**: Certain mathematical concepts, including but not limited to the concept *set*, have objective content.

2. **Axioms**: The content of these concepts is partially expressed in the sentences which axiomatize the concept. Such axiomatic systems can be inconsistent or incomplete; hence our grasp of the concepts involved can be insufficiently clear or not completely articulated.

3. **Platonism About Objects**: Certain mathematical objects exist, namely those that are quantified over by true sentences of theories which axiomatize a concept with objective content.

4. **Propositional Intuition**: Certain parts of our mathematical knowledge, including the result of reflection on concepts, is neither empirical, nor tautological, nor the conclusion of a deductive proof.

The theory as it stands is clearly short of a full articulation, and there are many philosophical problems lurking in the near vicinity. Most pressingly, it isn't clear what it means for a concept to have 'objective content', though Gödel's remarks suggest that this amounts to something like the axiomatization of the concept being a non-arbitrary matter (1964, p.261) where the resulting axioms 'force themselves upon us as being true' (1964, p.268). The position as I've reconstructed it is one that I think is ultimately unacceptable, for reasons related to the objectivity of the concepts, to be discussed below. My only claims for now are that it is a sensible reconstruction of Gödel's view, and that it is mysticism-free. The account sketched above omits entirely the discussion of *intuition of* mathematical objects (i.e. 'something like perception' of them), and the extrinsic, quasi-scientific justification of axioms. My view is that these elements of Gödel's thought are cleanly separable from the four elements above in a way that those four are not separable from each other. Later in this chapter, I'll argue that rejecting any substantial analogy between intuition and perception is necessary, but does little damage to the integrity

of Gödel's overall account of intrinsic justification. Quasi-scientific methods in set theory, being a major element of Gödel's thought (and undoubtedly the most well-received by later philosophers) will be the sole subject of chapter 5. It remains, however, to explore one final charge of theology against Gödel.

3.3 Gödel, Anselm, and Hilbert

I've argued that Gödel's platonism, far from being 'a kind of theology', combines elements of views in philosophy that are more-or-less mainstream: our knowledge of mathematics is sometimes non-deductive, non-empirical, and non-trivial; we should believe that sets exist because they are quantified over by axioms that we accept; the content of the concept *set* is partially expressed in the axioms of set theory, which can justified by philosophical reflection on what the concept *set* commits us to.

For all that, I think that Gödel's conceptual platonism has at its heart a serious philosophical issue, and to see why I'll examine another charge of mysticism or theology that can be levelled at Gödel's position. Once again, I think that the charge is misplaced, but nevertheless the commitments of Gödel's view here render it untenable, at least pending further analysis.

In the 11th century, Anselm of Canterbury wrote in the *Proslogium* 'So truly, therefore, do you exist, O Lord, my God, that you can not be conceived not to exist' (Anselm of Canterbury 1077/8, §3). This is the conclusion of his *ontological argument*; the very concept *God* guarantees the existence of an object falling under it. The thought is that the concept *God* subsumes all perfections, and a being which did exist would be more perfect than one which did not. Hence, God exists.

The final charge of theology that I'll discuss in relation to Gödel's work is this: conceptual platonism is essentially a mathematical version of the ontological argument. Merely by reflecting on the concept *set*, we can determine that the axiom of infinity (for instance), expresses some of its content. The axiom of infinity asserts the existence of an infinite set, and therefore, such a set exists. Gödel, just like Anselm, has deduced the existence of objects out of mere concepts, a move that a

critic might allege is mysticism *par excellence*.[6]

Of all such charges made against Gödel, this I think has the most force. A cornerstone of conceptual platonism is the idea that we gain knowledge of the soundness of axioms via conceptual reflection, and that from this, the existence of sets follows. This is seen most clearly when Gödel contrasts his interpretation of set theory with those of the constructivists and intuitionists, arguing that his is suitable for 'someone who considers mathematical objects to exist independently of our constructions and of our having an intuition of them individually, and who requires *only* that the general mathematical concepts must be sufficiently clear for us to be able to recognize their soundness and the truth of the axioms concerning them' (1964 p.258, my emphasis.).

Once again, however, Gödel's view has much in common with a mainstream view in the philosophy of mathematics, indeed one with a distinguished pedigree. Namely, his platonism incorporates elements strikingly similar to Hilbert's conception of an axiomatic theory, as discussed in his correspondence with Frege.

Hilbert's view of axiomatic theories is neatly expressed in his claim that '[i]f the arbitrarily given axioms do not contradict one another, then they are true, and the things defined by the axioms exist' (1899, p.42). Gödel's view of axiomatic systems is certainly not identical to Hilbert's and care must be taken in ensuring that the views are not conflated. Nonetheless, I think Gödel's inference from reflection on the concept *set* to the existence of sets is closer to the inference made by Hilbert than it is to the one made by Anselm, and the view of the former is certainly less 'a kind of theology' than the view of the latter! That said, it is no accident I think that Gödel thought that *some* version of the ontological argument could be made to work (Gödel 1970). Furthermore, the connection between Hilbert's view of mathematical theories and the ontological argument for the existence of God was highlighted by Frege (1900, p.47) merely eight days after Hilbert outlined his views on the subject.

The crucial points of difference between Hilbert and Gödel are that Hilbert takes the primitive terms appearing in his axioms to be *meaningless* in isolation, whereas Gödel explicitly speaks of the meaning of the primitive terms of set theory (1964, p.260). Similarly, Hilbert takes the axioms of a theory to be partially interpreted

[6]Thanks to Tim Button for playing the critic here.

syntax, but Gödel thinks they express part of the content of the concept that they axiomatize. Nonetheless, the similarity is clear; both take some axiom systems to have some property which is sufficient for the existence of some things. In the case of Hilbert, it is the syntactic consistency of the axioms which suffices for the existence of a mathematical 'system' (Hilbert 1899, p.40). For Gödel, what is required for the existence of objects appears to be more elusive, namely that the concept axiomatized is 'sound', or has objective content.

Unfortunately, Gödel is less informative than is required about what this condition on axioms amounts to. Clearly though, he does think of the axioms of set theory and of the concept *set* in this way. His remarks suggest that the objectivity of the content of the concept *set* is responsible for the non-arbitrariness of its axiomatization (1964, p.261), and later he claims that the resulting axioms 'force themselves upon us as being true' (1964, p.268). So Gödel's criteria are certainly more restrictive than Hilbert's, since we may suppose that a necessary condition on a decent axiomatization of a concept with objective content is that it be consistent.

Of course, it is to some extent a matter of taste whether or not an analogy holds good. I find myself impressed by the similarity of Gödel's view with Hilbert's, in that both maintain that certain axiomatic systems have a property sufficient for the existence of some things. Perhaps others will be less so. For exegetical purposes, all that is really important is the acknowledgement that Gödel's argumentative move, from concepts to objects, does not by itself justly invite the charge of mysticism. Amongst several central figures in the history of early analytic philosophy and mathematical logic, we find views that are similar in asserting that reflection on some aspect of mathematics is all that is required to ascertain the existence of something non-conceptual. Quite apart from Gödel or Hilbert, it was maintained by Frege that numbers exist simply as a matter of logic (1884), and by Dedekind that an infinite system exists as a consequence of the existence of any thought whatsoever (1888, pp.217-218). My point is not, of course, that we should adopt conceptual platonism merely because it sits in such illustrious company. Rather, the point is that we should not charge Gödel's view with mysticism *simply* because it takes the existence of objects to be verifiable by reflection on the relevant theories and concepts. As previously seen in this chapter, the attempt to portray Gödel as lacking a serious philosophical position to offer is found wanting.

There does, however, remain the question of the viability of this aspect of conceptual platonism. In assessing this, it's helpful to draw on a distinction sketched by Martin (2005, pp.209-210) between two differing interpretations of Gödel's remarks on the concept *set*. The first is to interpret Gödel's remarks in the 'straightforward' sense, according to which the concept *set* is what sorts sets from non-sets. In Fregean terminology, this is the first-level concept under which all and only sets fall. The second interpretation offered by Martin is to read Gödel as talking about the concept *set* in a structural sense.[7] In Fregean terminology, this is the second-level concept under which all and only first-level concepts of *set* fall. The concept *set* in this sense characterizes the structural properties that a first-level concept must have if it is to genuinely qualify as a set-concept, rather than some other kind of concept.

Martin's central criticism of Gödel is that reflection on the concept *set* can only give us information about the hierarchy in the structural sense. We might think, for instance, that reflection on the iterative conception of *set* could deliver the verdict that the axiom of pairs is true of it. At each level, we have *all possible* sets given what has come before. So if a and b both appear in V_α, the iterative conception dictates that the set $\{a, b\}$ appears at $V_{\alpha+1}$. Hence the axiom of pairs is satisfied. The axioms therefore may indeed force themselves upon us as being true at the second level, in that we may think no first-level concept could be a set-concept if it didn't satisfy the axiom of pairs.

What Martin rejects, however, is the idea that reflection upon the concept *set* can yield similar knowledge of the hierarchy in the straightforward sense. We can't, merely by reflecting on the axioms, sort sets from non-sets, and in particular we cannot determine by this kind of reasoning whether there are any sets at all. This echoes Frege's criticism of Hilbert in the geometric context, where the former complained that the axioms of the latter could not determine whether a pocketwatch is a point (Frege 1900, p.45).

The complaint is particularly vivid when we consider explicitly existential axioms, such as the axiom of infinity. Given that the iterative conception dictates that the sequence of stages goes on as far as possible, we might be able to convince ourselves that any set-structure must have a stage indexed by ω. We already have

[7] Martin calls this 'the concept *set* in my sense'. The terminology has been altered to avoid confusion.

the means to form the stages indexed by finite ordinals, so the result of the closure of the operation of forming such stages must also be a set. Even if convincing, however, such reflection can't show that *there is* such an infinite set, only that any set-structure must contain such a set.

Gödel's position, however, is that reflection on the concept *set* allows us to determine that axioms are true of the concept *set* in the straightforward sense. The meaning of the primitive terms of set theory do not *just* determine that any set-structure must contain an infinite set, though according to Gödel they do determine this. Much more than this, they determine that there is such a structure satisfying the axiom of infinity.

There is a sense in which Gödel's position is difficult to assess here, as compared to Hilbert's. The latter is quite specific that consistency is the property an axiomatic theory may have which is sufficient for the existence of the things collectively implicitly defined by the theory. For Gödel, however, it is the 'soundness' or 'objective content' of the concept axiomatized which is sufficient for the existence of some objects satisfying the axioms. Indeed, the success or failure of Gödel's project hinges on whether this notion can be made sufficiently precise.

According to Gödel, the 'criterion of truth in set theory' is the existence of a sufficiently forceful axiomatization of the concept *set* and a non-arbitrary series of extensions of that axiomatization (1964, pp.268–269). Since our intuition of the concept *set* is sufficiently clear to produce the axioms of **ZFC**, which 'force themselves upon us as being true', and since these axioms can be extended non-arbitrarily by further principles which serve to 'unfold the content of the concept' by means of further appeals to intuition, the criterion of truth is satisfied. So according to Gödel, there is no 'gap' between the truth of the axioms in the structural sense and the straightforward sense, just as for Hilbert there is no gap between the consistency of a theory and the existence of a system which is defined by it. It is the objectivity of the concept *set* which bridges the divide between structural and straightforward truths about sets, and it is in terms of this objectivity that the conceptual platonist must answer Martin's argument.

What is wrong with this line of thought? There two key points on which to put pressure. The first echoes Martin's (and Frege's) complaint that the axioms alone do not suffice to sort the sets from anything else. Even if the objectivity of the

concept *set* and 'structural' reflection on the axiom of infinity determines that $\aleph_0$ exists, for instance, there is a further question: is it my pocket-watch or not? I have some sympathy with this complaint; although the axiom of extensionality supplies identity conditions for sets in terms of other sets, it does not tell us whether $\aleph_0$ is a pocket-watch. But is this enough to show that the existence of $\aleph_0$ can't be established by reflection on the axioms?

Clearly, Martin's criticism is derivative of Frege's famous Caesar problem (1884, p.68e). Frege complains that some definitions previously given cannot decide whether Julius Caesar is a number. And certainly this simple form of the problem cannot be solved by appeal to the axioms of **ZFC**; for all they tell us, $\aleph_0$ might be any object whatsoever (other than another set). But the real force of the issue for Frege is that the indeterminacy of these identity conditions under discussion had the consequence that, while he had successfully defined the senses of expressions like 'the number 0 belongs to [the concept ϕ]', he had not succeeded in defining the expression '0' (Frege 1884, p.68e). If he had, it would follow by mere logic that $0 \neq$ Caesar.

The reason that the Caesar problem is so pressing for Frege is intimately bound up with the logicist project. Frege required that all arithmetical truths must follow from logic and definitions alone, hence the need for an explicit definition of number that ruled out troublesome identity statements like '0 = Caesar', the negation of which does not follow from an inductive definition of 'the number n belongs to [the concept ϕ]'. But it's not clear that Gödel's project faces anything like this problem.

It is true that, even by Gödel's own lights, '$\aleph_0 \neq$ Caesar' does not follow from the axioms of set theory. But for Gödel, the axioms of set theory needn't be taken (and indeed *can't* be taken) to express the full content of the concept *set*. Hence it is perfectly acceptable for Gödel to assert that reflection on the concept reveals that no set can be identical to a physical object. Whatever we may think of the notion that reflection on a concept can deliver verdicts on quite technical infinitary statements in set theory, it is less contentious to assert that if you understand the concept *set*, you understand that sets are not to be found in physical reality. When I say that '$\aleph_0$' denotes a set, part of what I mean is that it doesn't denote a thing you could bump into! So the force of Martin's complaint is not apparent to me outside of the Fregean context.

The other reason to reject the Gödelian argument is that it is simply too unclear whether or not the axioms of **ZFC** characterize a concept with the required objective content. One of the hallmarks of a concept with objective content is that its axiomatization fails to settle all of the relevant mathematical questions, necessitating the need for further appeals to intuition (1964, p.269). This principle sorts the wheat from the chaff, to some extent, since no concept with objective content can be properly axiomatized by an inconsistent theory, since such theories are not extendible by these non-arbitrary means. But the other signifier of objectivity for Gödel is that the 'axioms force themselves upon us as being true'. This is a deeply unhelpful means of sorting the theories which axiomatize a concept with objective content from those theories which do not. In the case of the concept *set* specifically, the ongoing controversy about the justification of **ZFC**, as well as its extension by large cardinal principles, ought to be a primary cause for concern. Without some clear means of distinguishing concepts with objective content from those which lack it, it is impossible to sustain the claim that if a concept has such content, then it follows that some objects satisfy it. If we do not know in what the having of such content consists, we can hardly claim such philosophically potent consequences for it.

This is ultimately why we should reject conceptual platonism as an account of mathematical objects: a crucial part of Gödel's theory is left insufficiently developed, namely the specification of what the objectivity of a concept's content amounts to. This rejection comes only to agnosticism, however. I do not currently see how an adequate account of the objectivity of a concept's content can be developed, but I have no argument that no such account can be found. Whether or not the account can be properly supplemented, I think that any charge of mysticism is misplaced. The view that mathematical entities of certain kinds are in some sense ontologically 'cheap' or 'thin' has great historical pedigree in the philosophy of mathematics, and is a mainstream position for analytic philosophers today.[8] I don't see how Gödel's version of this idea can be made viable, but that a philosophical theory has an explanatory gap can hardly qualify it as mysticism. Gödel *does* have an argument as to why the existence of sets follows from the concept set, which goes beyond

[8]Most recently, a version of this idea has been put forward by Linnebo (2018).

the mere observation that the concept is axiomatized by a theory containing an existential quantifier. The argument is unsuccessful, but that is a rather mundane philosophical problem.

3.4 Mathematical Perception

I've outlined a version of platonism and mathematical intuition reconstructed from Gödel's remarks, and defended it against charges of mysticism. Even if we can't quite accept the theory in its present state, we've reached a position where we can discuss Gödel's argument that the incompleteness theorems support platonism, and assess the question of whether intuition gives us a means to substantially reduce the incompleteness of set theory. Before moving on to those questions in the next chapter, I'd like to discuss one aspect of Gödel's thought that I've almost completely neglected in reconstructing his platonism: the hypothesised ability to perceive sets. Though I have little of substance to add to the literature that already exists on the subject, I'll try to explain why I've not given the issue any attention until now, and why I think we shouldn't take any hypothesis of mathematical perception to constitute a substantial element of Gödel's platonism.

The comments made by Gödel in regards to the perception of mathematical objects are less committal than one might suspect given the subsequent attention they have received. As we shall see, he often speaks of perception of mathematical objects and concepts analogically or metaphorically, and his remarks frequently tell against any interpretation of him as postulating a literal ability to see sets or concepts. Nonetheless, Maddy (1990) develops a full Gödelian platonism on the basis of taking perception seriously.[9] This might tempt us to think along the following lines: conceptual platonism had to be rejected because it posited the existence of sets as a consequence of facts about the concept *set*, but the argument could not be sustained due to a crippling lack of clarity. But why reject Gödel's view at this point, rather than find an alternative justification for thinking that some objects *do* satisfy the concept *set*? At this point, the hypothetical critic might point out that

[9]Maddy's reading of Gödel is very selective, as she herself is the first to admit (1990, p.78). The view as offered is Gödelian, not Gödel's.

Gödel *does* offer just such an alternative justification, namely the idea that we can perceive mathematical objects.

While I admit that the omission of a discussion of perception is an interpretative risk in developing an account of Gödelian platonism, I think it is a defensible one.[10] There are two chief reasons why, which shall be discussed in turn: firstly, the aspects of Gödel's thought that I have chosen to emphasise at the expense of perception are genuinely more prominent working parts of Gödel's philosophy of set theory; secondly, I don't see how perception, or anything much like it, could take over the role that Gödel's platonism about concepts plays in establishing his platonsim about sets, as the imagined critic suggests.

The first important thing to note is that Gödel's remarks about mathematical perception do not strongly commit him to the view that we can actually perceive sets, nor to the view that such perception justifies a belief in a structure satisfying the axioms of set theory. Indeed, he is consistently quite clear that talk of mathematical perception is not to be taken literally. In the Gibbs lecture he claims that our 'perception' of mathematical objects is nothing to do with the spatio-temporal 'world of real things' (1951, p.320) and that the objects of mathematics are completely different from the objects of the senses (1951, p.312, fn.18). He does speak of 'an objective reality...which we can only perceive or describe' (1951, p.320), but this is explicitly in relation to *concepts*, so the remark does little to support the interpretation of Gödel as thinking that platonism about sets can be established independently of our knowledge of the concept *set*. He also describes the aforementioned objective reality as 'non-sensual' (1951, p.323), and claims that from the objectivity of mathematics 'it follows at once that its objects must be totally different from sensual objects' (1951, p.312, fn.18).

Little more support is found for the view in either version of 'What is Cantor's continuum problem?' (1947 and 1964). Gödel does state that we 'have something like a perception also of the objects of set theory' (1964, p.268), but he claims in

[10]Martin (2005, p.220) asserts that Gödel would 'no doubt reject' an account of intuition which equates intuition with the understanding of a concept. This appears to be for reasons that relate to the perception of mathematical objects. But note that the reading I have urged of Gödel is *not* that intuition and understanding a concept are to be equated, but that there are two distinct notions of intuition, one of which is roughly the understanding of concepts, the other of which is roughly Kantian.

the same breath that those objects are remote from sense experience, and that our 'something like a perception' is grounded in the self-evidence of the axioms of set theory. Indeed, this particular kind of perception is explicitly stated to be mathematical intuition, and in the next paragraph, he says that what is given in intuition is '*not*, or not primarily, the sensations'. Hence even where Gödel explicitly discusses perception, he falls far short of committing himself to a robust view that sets can *literally* be perceived. In light of that, I think we should interpret Gödel's remarks about 'a kind of perception' as referring to *intuition of*, i.e. singular objectual intuition of the kind posited by Kant and Hilbert. Moreover, there is no hint that he takes this to be the sort of thing that could convince us that there are sets independently of our knowledge of the concept *set*. For this reason, I've not taken perception or 'something like a perception' of mathematical objects to be a central part of Gödel's view in reconstructing his platonism here.

Setting aside matters of interpretation, I think there are good philosophical reasons to think that Gödelian platonism cannot be salvaged by giving to perception of sets the founding role in establishing the existence of sets that I have ascribed to our knowledge of the concept *set*.[11] There are of course, all sorts of difficulties in positing a perception of even small finite sets. To use Maddy's favoured example, suppose there is a carton of three eggs before you. Do you see a set of three eggs? Three singletons each with an egg as the member? The set of three such singletons? The set whose members are the first egg and the pair of the second and third eggs? Such problems of individuation are severe enough that Parsons (2008, pp.212-214) denies even intuition of finite sets, let alone literal perception of them.

Secondly, as Maddy points out (1990, p.59), where we perceive a set of physical objects, the same spatio-temporal region will be home to any set with just those physical objects and no others in its transitive closure. This means that either our perceptual capacities are such that whenever we see an object, we see a set of rank α for any ordinal α, or our capacity to see sets is limited by the ordinal rank of a set, which is not *prima facie* a sensory property at all. The former possibility ascribes an astonishing richness to our visual capabilities, while the latter is manifestly *ad*

[11] On Gödel's behalf, of course. I should stress again that I take the inference from some property, either of the concept *set* or of its axiomatization, to the existence of sets to be illegitimate, absent a sufficiently developed account of how this could work.

hoc. Neither possibility bodes well for the defender of set-perception.

Quite aside from these difficulties, however, it is unclear that perception of sets could possibly do the philosophical work that conceptual platonism ascribes to the objectivity of the content of a concept. In other words, it is difficult to make out how perception of sets could determine that any particular set theory was satisfied by some objects. Any account of set-perception that is weak enough to be plausible will radically underdetermine which set theory is true, and hence won't assure us that the particular axiomatization of set theory delivered by our reflection on the concept set is satisfied by some objects. Conversely, any account of set-perception strong enough to ensure us that a powerful set theory such as **ZFC** is satisfied will be so strong as to be implausible. For example, can we really claim to *see*, in a purely visual sense, that the axiom of choice holds unrestrictedly, as opposed to a restriction of choice up to some large ordinal rank like V_{ω_1}? I think not.

In summary, the omission of the notion of mathematical perception as a core tenet of Gödel's platonism is interpretatively sound, and philosophically well-motivated; even if the problems with the notion could be overcome, the perception of sets is in no position to fulfil the role in Gödel's platonism that was occupied by his quasi-Hilbertian views about the objects of axiomatic theories. It is the content of the concept *set*, and the self-evidence and non-arbitrariness of its axiomatization which, according to Gödel, guarantees the existence of objects which fall under the concept and satisfy the axioms, together with a non-arbitrary series of extensions of them. So while Gödel does think that we have 'something like' a perception of mathematical objects, this plays no significant role in his arguments for platonism.

Conclusion

The central goals of this chapter were to get clear about what Gödel's platonism amounts to, and to defend it against charges of mysticism. Although I'm sceptical of claims by any one position to represent Gödel's views uniquely, I outlined a position called *conceptual platonism* reconstructed from salient remarks by Gödel in several major works. According to this position, what matters for establishing platonism about mathematical objects is not an account of how we 'interact' with

acausal objects or similar, but rather an account of how we apprehend the truth of axioms which express the content of the concepts under which those objects fall. In particular, sets are known to exist because they are quantified over in the axioms of set theory, the self-evidence and non-arbitrariness of which suffice to establish the objectivity of the interative concept *set*. I urged that the fatal flaw in this position is the lack of a developed account as to how we can distinguish, even in principle, those theories which axiomatize a concept with objective content from those which do not.

Although I don't think we can accept the view in light of that serious omission, I do think it's a sophisticated position which cannot be fairly accused of mysticism or theology. Indeed, the view shares crucial points of contact with mainstream views in the philosophy of mathematics, namely Quine's view that mathematical ontology should be assessed via accepted axioms, and Hilbert's conception of axiomatic theories. Of course, Gödelian platonism departs from both of these in important respects, but nonetheless it stands as a serious philosophical position, and not a series of mystical insights.

There is no substantive role in this account to be played by perception of mathematical objects, or anything analogous. I've argued that the omission is not interpretatively serious, since Gödel's remarks on perception have largely been taken out of context and over-stated in the literature. Moreover, it is not a philosophically promising avenue with respect to the verification of set-theoretic axioms. So even if Gödel gave the idea more credit than I have done, it is cleanly separable from the other elements of his thought, and deserves no central role in a rational reconstruction of his platonism.

In the next chapter, I'll put these interpretative efforts to work in answering two central questions about Gödel's platonism: firstly, does it justify the axiomatization of set theory in a manner that substantially reduces its incompleteness as compared to standard **ZFC**? And secondly, is the view lent substantial support by the incompleteness theorems?

Chapter 4

Intuition and Reflection Principles

Introduction

I've reconstructed a form of platonism from Gödel's remarks, *conceptual platonism*. The view is characterized by four central principles, given in §3.2. In this chapter, we'll see how conceptual platonism deploys Gödelian intuition in an attempt to establish the truth of certain *set-theoretic* reflection principles, related to, but importantly distinct from, the arithmetical reflection principles discussed in chapters 1 and 2.

In chapters 1 and 2, we saw that the use of arithmetical reflection principles could not completely eliminate the kind of incompleteness generated by Gödel's theorems. In a similar vein, I'll argue in this chapter that reflection principles cannot eliminate the distinctly set-theoretic elements of incompleteness. In §1, I'll give an interpretation of Gödel's philosophy of set theory in terms of conceptual platonism. Set-theoretic reflection principles are a crucial element in Gödel's platonism, and they do result in some reduction of the incompleteness of set theory by systematically increasing the strength of the axiomatic system, as detailed in §2. I'll outline Koellner's limitative results on what can be achieved by such reflection, and in §3 I'll go on to question whether even the little they do establish ought to be accepted. Ultimately, I'll conclude that certain reflection principles might be supported by conceptual platonism, but that incorporating them into the view requires philosophical modifications which render it less appealing as a philosophy of set theory.

In §4, I'll move on to discuss the extent to which this view of the hierarchy can be said to be supported by Gödel's theorems. I'll examine the arguments put forward in the Gibbs lecture, and conclude that some weak level of support is offered to Gödel's platonism by the incompleteness results.

4.1 Intuition in Set Theory

In this section we'll see how conceptual platonism tackles the case of set theory, and in particular what it is has to offer the programme of justifying set-theoretic axioms.

A well-known classic addressing the question of which axioms follow from the concept *set* is (Boolos 1989). There Boolos identifies three distinct 'thoughts' behind the concept *set* from which the axioms of **ZFC** follow. The first is *analyticity*, from which the axiom of extensionality follows: since it is the criterion of identity for sets, if what you're talking about doesn't satisfy extensionality, then they are trivially non-sets.[1] The second thought is the *iterative* conception, according to which sets are formed in successive stages such that at each stage every combinatorially possible set is formed, given what was 'present' at the previous stage. According to Boolos, this validates all the remaining axioms with the exception of choice and (the instances of) replacement. These, he claims, follow from the third thought, *limitation of size*: any things form a set unless there are 'too many' of them.

The usual articulation of limitation of size is in terms of a second-order axiom stating that any class is a set iff there is no bijection between the class and V, the universe of sets. The idea is certainly independent of the iterative conception; as Parsons emphasises, this way of thinking about the principle (known as *von Neumann Limitation of Size*) makes it a sorting principle for dividing sets and proper classes in a given second-order 'universe', rather than a principle for constructing sets 'from below' (2008, p.133).[2] This is similarly true of weaker principles like *Cantor Limitation of Size*, an axiom stating that a class is a set iff there is no bijection between it and On, the class of ordinals.[3] Replacement and choice are acceptable according to either principle, since they do not imply the existence of collections which can be put into one–one correspondence with On or V. So according to the limitation of size thought, the relevant sets exist, and the axioms are true.

[1]This argument is offered by Boolos only tentatively, due to Quinean anxiety about the notion of analyticity.

[2]Parsons also points out that this version of the axiom of limitation of size has the rather unintuitive consequence that V is well-ordered: by Burali-Forti's theorem, the ordinals do not form a set. By limitation of size, there is class-function from On to V that is bijective. The map therefore induces a well-ordering on V (this principle is also known as the axiom of *global choice*).

[3]The principle is weaker because it is silent on question of whether there is a bijection between the class of ordinals and the universe. Consequently it does not imply global choice.

Gödel similarly considers what axioms follow from the concept *set*, though he takes less care than Boolos does to distinguish the various thoughts behind the concept. With respect to analyticity, Gödel's view is that *all* the axioms 'might fittingly be called analytic' (1951, p.321). The thesis is not as radical as it might at first seem, since Gödel operates with a dual notion of analyticity. He distinguishes between what we might call 'narrowly' analytic propositions, and 'broadly' analytic propositions (1951, p.321). The former are characterized by Gödel as 'true owing to our definitions', or as 'tautologies', and as such are recognizably analytic in the traditional sense. Gödel is quite explicit that the axioms of mathematics as a whole are *demonstrably* not analytic in this sense (1944, p.139), since there is no decision procedure for arithmetic. On the other hand, a proposition can be be broadly analytic, in the sense that it is true 'owing to the the meaning of the concepts occurring in it' (1944, p.139), or the 'nature' of the relevant concepts (1951, p.321).

This is quite removed from the traditional conception of analyticity, since Gödel goes so far as to claim that a proposition analytic in this sense might be undecidable. Although Gödel doesn't go into enormous detail on the distinction between kinds of analyticity, the basic idea seem to be that narrowly analytic propositions are reducible to explicit tautologies by purely syntactic methods, whereas broadly analytic propositions can only be seen to be true by semantic reflection on the primitive terms which appear therein.[4] But it is in *this* sense that Gödel claims that the axioms of mathematics might be called analytic; hence the thesis is closer to being a restatement of his conceptual platonism, rather than a radically new account of analyticity. In any case, Gödel gives us no hint that analyticity, conceived of in this manner, distinguishes the axiom of extensionality from other mathematical axioms.

The iterative concept of *set* does, however, play an important role in Gödel's analysis. According to him, the primitive concepts involved in set theory are that of a set as an 'arbitrary multitude' (1964, p.262), and the concept *property of set* (1964, p.260 fn.18).[5] Gödel's concept of sets as arbitrary multitudes seems to be what we

[4]Though unusual, Gödel's view here is not completely without precedent. Ramsey argues that certain set-theoretic propositions might be both unprovable and tautological (Ramsey 1925, p.224). Of course, Ramsey is operating with a very particular notion of a tautology here.

[5]The notion of arbitrary multitude here is meant as opposed to that of a *definable* multitude. A truly arbitrary concept of multitude would perhaps include non-well-founded sets, but Gödel doesn't discuss this issue here.

would today think of as the iterative conception, namely that of sets obtained from the urelements by iterated application of the 'set of' operation (1964, p.259). At stages of the process where the set-forming operator is the powerset operation, this is 'by definition' the full powerset (1951, p.306 fn.5).

The limitation of size does not play a direct role in Gödel's analysis of the concept *set*, but a similar role is played by a different *maximality principle*, namely that the hierarchy of sets is *inexhaustible*. The idea plausibly has its origin in Cantor's distinction between the transfinite and the 'absolutely' infinite, but the mature statement of this principle is given by Gödel in terms of the concept *set*, rather than the hierarchy itself (as one might expect given his conceptual platonism). Moreover, it is stated in terms of the axiomatization of this concept, rather than its extension (also to be expected, if one accepts my arguments so far about Gödel's view). Comments suggesting that Gödel took the axiomatization of set theory, as well as the hierarchy itself, to be inexhaustible appear at least as early as (1933, p.47). But the mature statement is:

> [T]he axioms of set theory by no means form a system closed in itself, but, quite the contrary, the very concept of set on which they are based suggests their extension by new axioms which assert the existence of still further iterations of the operation "set of". (1964, p.260)

In other words, *the concept set is not exhausted by our ability to axiomatize it.*[6] In my view this is *the* core tenet of Gödel's platonism as applied to set theory, and is a nice way of cashing in the metaphor that 'concepts form a reality of their own, which we cannot create or change' (1951, p.320). The concept *set* in particular, is not simply invented by us. If it were, then it would in some sense be 'up to us' what axioms were valid with respect to it. But the non-arbitrariness of this open series of extensions, and indeed the possibility of substantive debate about what follows from the concept, shows that we do not enjoy the sort of freedom a creator would have.

[6]Parsons (2008, p.133) identifies the inexhaustibility principle as a variety of the limitation of size principle. For my part, I simply do not see how this is so, and no argument is offered by Parsons for the claim. The difference between the two principles is especially stark when it is observed that the most natural formalization of inexhaustibility is in terms of reflection principles. See below for details.

Given how I've so far characterized Gödelian platonism about sets, it's no surprise that mathematical intuition has an important role to play in the epistemology of the view. According to Gödel, we have 'an intuition which is sufficiently clear to produce the axioms of set theory and an open series of extensions' (1964, p.268), and the new axioms serve to 'unfold the content of the concept of set' (1964, p.261). The idea that successively stronger set-theoretic axioms 'unfold' the concept *set* is an important one for the plausibility of Gödel's position. As defined here, intuitive knowledge of the truth of an axiom is simply non-deductive, non-trivial, and not 'extrinsic'. The view that even the basic axioms of set theory are intuitively justified would be deeply implausible if intuition had to give *immediate* knowledge, or if the various deliverances of intuition had to be independent of one another (as perception of physical objects might plausibly be taken to be). Rather, intuitive knowledge of the axioms of set theory is supposed to be acquired in successive stages, and hence there is no requirement that the truths of various intuitable axioms be 'obvious' to the untrained eye, or be independently apparent.

Just as platonism about concepts leads to platonism about objects, the notion of the inexhaustibility of the axioms of set theory has an objectual counterpart in the conception of the set-theoretic hierarchy that it gives us. The 'absolute' infinity of the cumulative iterative hierarchy is such that it cannot be 'characterized from below'. There is a lot of metaphorical talk here, which is worth unpacking. A 'bottom-up' approach to set theory is broadly one in which the hierarchy of sets is described in terms of its members or levels, and without explicit mention of V or classes over V (as occurs, for instance, in the limitation of size axioms). In practical terms, a bottom-up selection of axioms is motivated by considerations about what *sets* are like and how they should behave, as opposed to considerations about how the universe as a whole (or the models of the theory) should look.[7] So, to say that the hierarchy cannot be 'characterized from below' is to say that in set-theoretic terms (and perhaps in a stronger sense), the height of the hierarchy is indescrib-

[7] All of this is merely heuristic, of course, and the 'bottom-up' approach to set theory cannot be precisely defined. But for an example of the approach in action, see Tait's work (2005 in particular). As he emphasises, the approach is somewhat out of fashion, with most set-theoretic work today being 'top-down', thanks to the focus on models of set theory.

able.[8] Making this more precise, we can say that no formula in the language of set theory, ϕ, can uniquely characterize the height of the hierarchy.[9] One consequence is that if ϕ *does* characterize the hierarchy, it does not do so uniquely; rather it characterizes an initial segment. Hence if ϕ is true, then it is true when reinterpreted to be just about that initial segment, rather than the hierarchy as a whole.

Axioms which, in some form or another, state that the hierarchy is inexhaustible in this sense, are called *reflection principles*. Such principles form a key part of Gödel's epistemology of set theory, as they are the articulation of the inexhaustibility principle which he takes to be intrinsically justified by our grasp of the concept *set*. In the next section, we'll take a closer look at the precise shape reflection principles take, and examine the argument that some of them are intrinsically justified by the concept *set*.

4.2 Reflection Principles

In the context of set theory, a reflection principle is an axiom which, intuitively, says that any attempt to characterize the hierarchy can only succeed in characterizing a part of it. In other words, if ϕ is true of the hierarchy, then ϕ is also true of an initial segment. The formula ϕ is 'reflected downward', and fails to characterize the hierarchy uniquely.[10] The idea of a reflection principle is therefore thoroughly informal, and the strength of any formalized axiom will vary enormously depending on language of which ϕ is a sentence, and the order of the parameters which occur in ϕ. Though the precise form of such axioms varies, Incurvati (2016, p.165) provides a very useful 'template' for reflection principles, which can be thought of as doubly-schematic:

[8] Of course, **ZFC** has models whose height can be described in set-theoretic terms as a strongly inaccessible cardinal. But for the advocate of inexhaustibility, this just shows that no such model could contain the whole hierarchy.

[9] Or rather, we can't *correctly* and uniquely characterize the height of the hierarchy. It's consistent with **ZFC** of course, to say that there is no inaccessible cardinal. But the advocate of the inexhaustibility of the hierarchy must insist that this *mischaracterizes* the height of V.

[10] This notion is therefore distinct from the reflection principles previously examined, although both kinds of principles are to be justified by the grasp we have of certain concepts. Indeed, it is perfectly reasonable to think that our grasp of the concept *set* justifies the extension of **ZFC** by reflection principles of *both* kinds.

Reflection Principle Template: Let $\phi(A_1, ..., A_k)$ be a formula in the language of mth-order set theory with parameters $A_1, ..., A_k$ of order $\leq n$. The following is an axiom:

$$\phi(A_1, ..., A_k) \rightarrow \exists\alpha\ \phi^\alpha(A_1^\alpha, ..., A_k^\alpha)$$

where α is an ordinal and ϕ^α is the result of restricting the quantifiers of ϕ to V_α.[11] For the parameters, $A_i^\alpha = A_i$ if A_i is a first-order (set) parameter, and $A_i^\alpha = A_i \cap V_\alpha$ if A_i is a second-order (class) parameter. If A_i is a third- or higher-order parameter, then $A_i^\alpha = \{B^\alpha | B \in A\}$, where A is of order $n+1$ and B ranges over classes of order n.[12] Let RP_n^m be the axiom schema which results from restricting ϕ to the language of mth-order set theory, and limiting the parameters to order $\leq n$.

Perhaps the most commonly encountered instance of this template is RP_2^1. This principle, when added to extensionality, foundation, and the separation schema yields a theory which is equivalent to **ZF** (Incurvati 2016, pp.165-166). It can therefore be deployed in the justification of set-theoretic axioms which have less of a clear connection to the iterative conception of *set*, in particular the replacement schema. But the template covers far more than just this principle of course, so by formalizing the basic idea behind the inexhaustibility of the hierarchy, we are led to a progression of increasingly powerful reflection principles. In this section we'll review the central results in the area, and examine what exactly can be delivered by reflection principles of this form. We can then take to examining which, if any, of these principles can be considered deliverances of intuition of the concept *set*, and in particular whether intuition can be said to justify the existence of certain small large cardinals as a result.

If one were inclined to think the iterative conception of set had a notion of inexhaustibility built into it, and that reflection principles were an appropriate formalized expression of this principle, it might be tempting to think that the iterative

[11] V_α is a level of the standard von Neumann universe. Where $\alpha = 0$, V_α is the (possibly empty) set of urelements. Where α is a successor ordinal, $V_\alpha = \mathcal{P}(V_{\alpha-1})$. Where α is a limit, $V_\alpha = \bigcup_{\beta<\alpha} V_\beta$.

[12] If one is concerned to rigorously keep track of the distinction between sets and classes, let higher-order variables of order n range over sets in an isomorphic copy of $V_{\beta+(n-1)}$, rather than classes of order n (Koellner 2009, p.208).

conception thereby licensed truly unlimited reflection, i.e. that RP_n^m is valid for any choice of m and n. Sadly, such a temptation leads to inconsistency. Indeed, it runs into inconsistency remarkably quickly, as demonstrated by Tait (1998, p.481). Let U be the third-order class of bounded second-order classes. If we allow third-order parameters into our favoured reflection schema, we can reflect on $\phi(U)$, the statement that every member of U is bounded. When interpreted over V, $\phi(U)$ is plainly true, but for every β the restriction $\phi^\beta(U^\beta)$ is false. This is because U^β is simply the class of all (second-order) classes over V_β, including V_β itself, which is of course unbounded in V_β. Hence, RP_3^1 is inconsistent.

The crucial question, therefore, is how to extract significant deductive strength from the informal notion of a reflection principle which avoids inconsistency. To increase the deductive power of **ZFC** via reflection principles which fit the template above, we must add a principle which reflects certain higher-order formulae, rather than merely permitting higher-order parameters. In terms of our template, we must increase m, and not just n. This is because Tait's result shows that RP_3^1 is inconsistent, and Gloede has shown that RP_2^1 does not extend the deductive strength of **ZF** (see above). Another option would be to give a more fine-grained template for reflection principles in order to avoid inconsistency. The first option will be explored first.

Bernays-Style Reflection Principles

In the second-order setting, the addition of Bernays' reflection principle, RP_2^2, affords some manner of increase in deductive power thanks to the expressive resources available in the language of second-order set theory (see Bernays 1961). Since **ZF** is equivalent to **ZF**+RP_2^1, every instance of RP_2^1 is provable in **ZF**. Clearly, every instance is therefore also provable in $\mathbf{ZF}_2$, but moreover, a *single* provable sentence can express that fact, by saying that for any A there is an α such that the structure $\langle V_\alpha, \in, A^\alpha\rangle$ is an elementary substructure of $\langle V, \in, A\rangle$ (Incurvati 2016, p.166).

This formula implies that there is an inaccessible cardinal, and using RP_2^2, we can reflect on *that* to infer the existence of a second inaccessible. Iterating this reasoning implies that there is a proper class of inaccessibles, and a very similar argument can be used to show that there is a proper class of Mahlo cardinals. So adding

the principle RP_2^2 to second-order set theory yields a genuine increase in deductive power. Hence, if the intuitive idea of behind reflection principles does follow from the iterative conception of set, a topic to which we shall return below, formalizing things this way vindicates Gödel's claim (discussed in the previous chapter) that mathematical intuition is sufficient for justifying certain large cardinal axioms, in particular axioms which assert the existence of Mahlo cardinals, and perhaps other small large cardinals.

Bernays' reflection principle can be strengthened to yield somewhat stronger large cardinal principles, by allowing reflection on higher-order formulae while restricting the order of parameters to at most 2. These principles, RP_2^n-reflection (for $n > 2$), have received little attention in the literature, but can be shown to be consistent relative to certain large cardinal assumptions. RP_2^2 yields the existence of Π_1^1-indescribable cardinals.[13] As RP_2^2, is strengthened, the existence of indescribable cardinals of increasingly high order is implied.[14]

Tait-Style Reflection Principles

Another strategy, in order to extract greater strength from the intuitive reflection principle while avoiding inconsistency, is to consider a restriction on the kinds of formulae amenable to reflection. Tait proposes to restrict the formulae on which we can reflect to what he calls 'positive formulae' (1998, p.482). A formula is positive iff it is constructed from conjunction, disjunction, and existential or universal quantification from atomic formulae *except* formulae of the form $X \neq Y$ and $X, ..., Y \notin Z$, for any higher-order predicates. We can define the following classes of formulae:

> Γ_n^m **-Formulae:** Positive formulae of the form $\forall X_1 \exists Y_1 ... \forall X_n \exists Y_n \Psi$, where Ψ is at most first order, the Y_i's are of any finite order, and the X_i's are of order m.

This allows for a different reflection schema template to be defined:

[13] A cardinal k is Π_m^n-indescribable iff for every Π_m^n sentence ϕ, the following holds: $\forall S \subseteq V_\kappa(\langle V_\kappa; \in, S\rangle \models \phi \rightarrow \exists \alpha < \kappa(\langle V_\alpha; \in, S \cap V_\alpha\rangle \models \phi))$ (Kanamori 2009, p.58). In other words, a Π_m^n-indescribable reflects down every Π_m sentence of $n + 1^{\text{th}}$-order set theory.

[14] Thanks to Luca Incurvati for providing the information about these more general RP_2^n principles.

> Γ^m_n **-Reflection:** $\forall X(\phi(X) \rightarrow \exists\beta\ \phi^\beta(X^\beta))$, where ϕ is amongst the Γ^m_n-formulae.

From the simple reflection principle, we've come to a much more complicated hierarchy of principles, which may be considered axiom candidates for extensions of **ZFC** justified on the basis of the iterative conception of set, where that notion is taken to include the thought that the hierarchy is 'absolutely infinite', or cannot be reached from below.

The work of Tait (1998, 2005) and Koellner (2009) paint an almost complete picture of the consistency strength of the Γ^m_n-reflection principles. The key results are both due to Koellner (2009), and will be rehearsed here for the purposes of discussion. The reader should consult that paper for full proofs. The central message is in the form of a 'dichotomy', which says that reflection principles in this stratification are either weak or inconsistent.

The sense of weakness here is a rather specialised one; by 'weak' Koellner means that the principle has at most the consistency strength of an axiom asserting the existence of the partition cardinal (a.k.a. 'Erdős cardinal') $\kappa(\omega)$.[15] I'd like to avoid getting bogged down in technical details, but suffice it to say that, by today's standards, $\kappa(\omega)$ is not a particularly large cardinal. Significantly for our purposes, an axiom asserting its existence is of a higher consistency strength than the upper bound on what can be obtained using Bernays' reflection principle RP^2_2, and more generally, the Bernays-style principles RP^n_2 (Kanamori 2009, p.472. Full details are at pp.59–71).

So Koellner's dichotomy is that reflection principles are either weak, in that they have the consistency strength at most of an axiom stating the existence of $\kappa(\omega)$, or they are inconsistent. The dichotomy is not strictly a theorem, since the notion of a reflection principle is informal, and is always capable of extension to cover new formal principles (a fact which Koellner himself readily acknowledges (2009, p.217)).[16] That said, Koellner's dichotomy is powerfully supported by the following

[15] Let $[\alpha]^{<\omega}$ be the union, for all n, of the n-element subsets of α. For any limit ordinal β, the Erdős cardinal $\kappa(\beta)$ is the least cardinal λ with the following property: for every $f : [\lambda]^{<\omega} \rightarrow \{0, 1\}$, f is constant on $[H]^{<\omega}$ for some β-sized subset H of λ. See (Jech 2003, pp.109 and 302) for full details.

[16] I don't want to commit here to anything as strong as that no proof can ever be found where informal notions are concerned. The point is simply that in *this* debate, the relevant informal

two theorems (2009, p.210 and p.213):

> **Theorem 1:** Suppose $\kappa(\omega)$ exists. Then for some $\delta < \kappa(\omega)$, V_δ satisfies Γ^2_n for all $n < \omega$.
>
> **Theorem 2:** Γ^3_1 reflection is inconsistent.[17]

In Koellner's stratification of reflection principles, we have it that reflection (of any complexity) on a second-order formula (in our positive language) is weak. By contrast, the next strongest reflection principle in the stratification is inconsistent. The challenge Koellner lays down is to find a different reflection principle which is justified by the iterative conception of set, that is of higher consistency strength than the existence of $\kappa(\omega)$, and is indeed consistent.

At one time, Gödel appeared to think that *all* principles of set theory could be derived from some form of reflection principle,[18] but Koellner's dichotomy strongly suggests that the principles of set theory today cannot all be justified in this way. In the next section, we'll discuss the issue of whether even the RP^n_2 and the Γ^2_n reflection principles are justified by the iterative conception; before that I'll outline the recent research on the subject of Koellner's challenge and argue that it has not been successfully met.

Koellner's Challenge

Koellner (2009, pp.217-8) discusses very strong reflection principles, though they appear to be a long way from the sort of principles that can be said to follow from the

notion appears incapable of conclusive formal treatment. The reason is that what counts as a reflection principle is highly sensitive to the linguistic resources we allow ourselves.

[17]The proof requires the use of a fourth-order parameter. So, temporarily reverting to our default template for reflection principles, this shows that RP^3_4 is inconsistent, even with Tait's restriction to reflection on positive formulae.

[18]At least, he is cited as having thought so by Hao Wang (Wang 1996, p.283). In earlier papers, such as (Gödel 1933), Gödel talks exclusively about establishing new axioms by means which look a lot like reflection principles. What's not clear to me, however, is the status of very large cardinal axioms (of the kind discussed in the next chapter) in this. He may at some point have thought that such axioms could be established by a reflection principle, and later changed his mind (necessitating the use of extrinsic methods of justification). Or perhaps he never thought such axioms could be justified by reflection at all, but changed his mind about whether there was a non-reflective admissible justification of them. The evidence does not strike me as decisive.

iterative conception of set. The same appears to be true of earlier strong reflection principles, such as those discussed by Marshall (1989), though it would of course be anachronistic to suggest that her reflection principles ought to be evaluated in terms of whether they meet Koellner's challenge.

Since 2009, however, several new reflection principles have been formulated with Koellner's challenge specifically in mind. The first such principle, given by McCallum (2017), does not answer Koellner's challenge, but appears to have strengthened it: he formulates a reflection principle which subsumes those discussed by Tait and Koellner which are not known to be inconsistent, and shows that this principle is equivalent to the existence of a remarkable cardinal.[19] The existence of such a cardinal is consistent relative to the existence of $\kappa(\omega)$ in **ZFC** (Schindler 2000, p.180), and is of a higher consistency strength than principles which can be obtained from the RP_2^n principles. Hence, if correct, McCallum's reflection principle strengthens Koellner's case by reducing the 'barrier' which candidates for intrinsically justified reflection principles appear not to be able to break.[20]

Two principles have recently been formulated which purport to offer intrinsically justified reflection principles which break the $\kappa(\omega)$ barrier. My view is that, while both the principles have some intrinsic appeal, neither of them succeed in defeating Koellner's challenge.

The first such principle is proposed by Horsten and Welch (Horsten and Welch 2016a and Welch 2017). The principle asserts that the *Global Reflection Property*

[19] A cardinal κ is remarkable iff for all regular $\theta > \kappa$ there are $\pi, M, \lambda, \sigma, N, \rho$ such that:

1. $\pi : M \to H_\theta$ is an elementary embedding (where H_θ is the set of all sets hereditarily smaller than θ.
2. M and N are countable and transitive.
3. $\pi(\lambda) = \kappa$
4. $\sigma : M \to N$ is an elementary embedding with critical point λ.
5. $\rho = M \cap On$ is a regular cardinal in N.
6. $\sigma(\lambda) > \rho$.
7. $M \in N$ and $N \models M = H_\rho$.

The significance of these cardinals is discussed in (Schindler 2000).

[20] The paper has not yet passed peer review, however, so I shall continue to speak as if the barrier in question is $\kappa(\omega)$.

(GRP) holds of the hierarchy. Let V be the hierarchy as normal, and let C be the collection of classes over V,[21] where class-talk is interpreted as quantification over parts of V, in the mereological sense (ignoring the Lewisian concerns about $\emptyset$ and the mapping from a set to its singleton). Horsten and Welch propose the following new axiom:

> **Global Reflection Property:** For some ordinal κ, there is a non-trivial elementary embedding $j : (V_\kappa, \in, V_{\kappa+1}) \to (V, \in C)$ with critical point k which is elementary for first-order formulae with class parameters. (Horsten and Welch 2016a, p.14)

The value of the principle is primarily located in its consistency strength; in particular it is far stronger than Koellner's principles and implies the existence of a proper class of Woodin cardinals.[22] The source of this consistency strength is that it says, not just that each reflected formula is reflected somewhere, but that they are *all* reflected in some particular initial segment of V. Moreover, at V_κ, the classes are indistinguishable from their respective intersections with $V_{\kappa+1}$. The principle is not 'bottom up' like the reflection principles previously examined here, in that it doesn't just assert that our language fails to distinguish the universe from some part, no matter what we say. It is 'top down', in the sense of postulating a very strong resemblance between V and an initial segment thereof.

This principle strikes me as somewhat intuitive, at least with respect to a certain conception of set theory. The problem, at least as far as the present project is concerned, is that the conception of set theory required to motivate the principle goes far beyond what is given in the iterative conception of *set*. First, the top-down nature of the principle requires us to consider the hierarchy as completed infinite totality; this much is clear from the explicit mention of V in the formulation of the principle, which is needed to give the range for the elementary embedding j. Horsten and Welch are eager to embrace such *Cantorianism*, and go to some

[21] In neither paper cited is it made explicit what kind of collection this is.

[22] A cardinal κ is Woodin iff for all $f : \kappa \to \kappa$ there is some $\alpha < \kappa$ such that $\{f(\beta) : \beta < \alpha\} \subseteq \alpha$ and an elementary embedding j from V into a transitive inner model M such that α is the critical point of j and $V_{j(f(\alpha))} \subseteq M$ (Kanamori 2009, p.360).

lengths to justify their position.[23] The alternative is a *Zermelian* conception of the set-theoretic realm as an unbounded sequence of models indexed to the ordinals. We'll return to these issues below, but for the present purposes, the details are not so important. The central issue is that a Zermelian cannot even make sense of the reflection principle put forward by Welch and Horsten, which is sufficient to show that it is not implied by the iterative conception alone: if the GRP were implied by the iterative conception of *set*, then a Zermelian interpretation would be *ruled out* by it. Whatever one may think of Zermelo's interpretation of set theory, it surely *consistent* with the iterative conception, even if the latter does not imply the former.

A second difficulty is that the motivation for the principle relies critically on the conception of proper classes as parts of V. This is because classes are postulated as the relata of a resemblance relation, which rules out any nominalist or plural interpretation of class-talk (Horsten and Welch 2016a, p.20). Moreover, conceiving of classes as governed by the transitive 'part of' relation, rather than $\in$, is crucial for motivating the restriction favoured by Horsten and Welch that the embedding j be elementary with respect to first-order formulae, but not to higher-order formulae. Not only does this deliver certain technical results that Horsten and Welch find desirable, it also ensures that the justification of GRP doesn't over-generate and justify stronger versions of GRP which may turn out to be inconsistent.

None of this would come as a surprise to Horsten and Welch, of course. Welch explicitly claims that the principle does *not* follow from the iterative conception alone (Welch 2017, p.11). But he asks us to 'swallow the Cantorian pill' and accept that there is a real distinction between sets and classes in order to motivate the principle (Welch 2017, p.7). Since it doesn't follow from the iterative conception, the GRP certainly does not defeat Koellner's challenge in letter. Moreover, as I'll argue below, the reflection principles favoured by Gödel are easier to motivate in a Zermelian setting; so for our purposes at least, we can regard Koellner's challenge as unanswered. That said, those sympathetic to Cantorianism and a merelogical concept of classes should regard the GRP as a very serious axiom candidate which offers an intrinsic justification for extremely strong large cardinal principles.

[23] I'll follow Horsten and Welch in their use of the expressions 'Cantorian' and 'Cantorianism', though I make no commitment to any related historical claims about Cantor's actual view of the hierarchy.

A second strong reflection principle has been proposed by Roberts (2017). The principle is certainly an intuitive one, and is stated quite simply: $\phi \rightarrow \exists C \phi^C$. It says that if ϕ is true of all entities of some kind, there is a set-sized collection of entities of that kind, C, and ϕ is true restricted to C (2017, p.651). Despite its simplicity, the principle is remarkably strong: its formalization in second-order set theory implies the existence of a proper class of 1-extendible cardinals (2017, pp.659–660), and is consistent if a 2-extendible cardinal exists (2017, pp.660–661).[24]

Much like Welch, Roberts seems not to have high hopes for the intrinsic justifiability of his reflection principle in terms of the concept *set*. The inconsistency of third-order reflection makes him sceptical of the intrinsic justifiability of any reflection principle, since the standard justification for reflection over-generates. Indeed, Roberts is sceptical more broadly of the notion of intrinsic justification (Roberts 2017, p.657). However, he claims that if his principle isn't intrinsically justified, then Bernays' principle probably won't be justified either (Roberts 2017, p.657). He claims that the greatest threat to the justifiability of his reflection principle is that it implies the existence of classes. However, as Roberts quite rightly points out, this is also true of Bernays' principle.[25]

While this is correct as far as it goes, I do think that an important distinction in justifiability can be drawn between Bernays' principle and Roberts'. The key reason that the latter cannot be said to follow from the iterative conception is that it ineliminably relies for its formulation on the notion of a *set-sized collection*, while Bernays' and Tait's principles do not. So, any intuitive or intrinsic motivation for the axiom must, at least implicitly, appeal to a prior distinction between collections which are set-sized, and collections which are class-sized. But the iterative conception, according to which sets are formed in stages, such that each stage contains every possible collection of what was present at the previous stages (continuing into the transfinite), makes no mention of *size* at all. Perhaps some variant of Roberts' axiom could be justified in terms of the limitation of size, by stipulating that C is

[24] A cardinal κ is β-extendible iff for some λ there is an elementary embedding $j : V_{\kappa+\beta} \rightarrow V_\lambda$ with critical point κ such that $\beta < j(\kappa)$ (Kanamori 2009, p.311). The assumption of even a 1-extendible cardinal is very strong.

[25] It is not obvious from the formulation of RP_2^2 that it implies the existence of classes. That it actually does so can be seen from the fact that the full schema in $\mathbf{ZFC}_2$ implies the axiom of global choice (Kanamori 2009, p.59), an explicitly class-theoretic principle.

set-sized iff there is no bijection between C and V, or similar. But the limitation of size is not part of the iterative conception.

It strikes me that an axiom such as Roberts', which relies on a notion of the size of a collection, is critically different from one that does not, such as Bernays'. Without a serious explanation of how the iterative conception alone, despite all appearances, really does carry with it a notion of size suitable to motivate Roberts' axiom, we can safely consider Koellner's challenge to be unmet.

Since the notion of a reflection principle is so deeply informal, it is difficult to make any final judgement on what can or cannot be achieved by means of them. Koellner's Γ^2_n principles, and Bernays' RP^n_2 principles are consistent, unlike their strengthened counterparts, and offer some deductive extension of **ZFC**. Stronger reflection principles of distinct kinds have been proposed, but cannot be said to follow from the iterative conception of *set*. Since we are concerned with the iterative conception here, for our present purposes at least, we can take reflection principles to be those at most as strong as Koellner's consistent principles, since these are the strongest known reflection axioms which *might* at least have a justification available in terms of the iterative conception. Whether or not they actually do have such a justification will now be discussed.

4.3 The Limits of Reflection

We now have two candidates for reflection principles which might be said to follow from the iterative concept of set. Both Bernays-style RP^n_2 axioms and Koellner's Γ^2_n axioms offer the possibility for the conceptual platonist to claim intuitive knowledge of large cardinal principles. In particular, both kinds of principles support the claim that *small* large cardinal axioms (i.e. those large cardinal axioms which have a consistency strength below that of an axiom asserting the existence of $\kappa(\omega)$) can be verified by non-deductive, non-conventional, and non-extrinsic means. But this is conditional on giving the reflection axioms a justification in terms of the iterative conception of set. Can this be done in terms acceptable to the conceptual platonist? Before answering this question, it's worth saying something about the general significance of Koellner's limitative results for our broader study of incompleteness.

Two Kinds of Incompleteness

The first thing to note is that for our purposes, Koellner's theorems play an analogous role to that played by Feferman's theorem in the first two chapters. We saw that Gödel's incompleteness theorems quite naturally lead one to suppose that the formal theory we started with (e.g. **PA** or **ZFC**) hadn't adequately captured the concept we wished to axiomatize. In the case of arithmetic, it seems that whatever justification we had for believing in a theory is also justification for strengthening that theory by Feferman reflection. However, the processes of iteratively strengthening our theory by such means does not suffice for the total elimination of incompleteness. Unless we can tolerate the idea that an idealized mathematician has an inexplicable knowledge of the arithmetical truths required to axiomatize a complete arithmetic by iterated reflection, then we must give up on the idea that every arithmetical proposition is decidable, even in an idealized sense.

In the case of set theory, something similar appears to hold. Whilst set theory is of course subject to Gödelian incompleteness, there is also a second kind of incompleteness that afflicts it. The axioms do not merely fail to decide consistency statements, Gödel sentences and so on, but also fail to decide other set-theoretic statements, most famously the continuum hypothesis. Closing off this kind of incompleteness via set-theoretic reflection principles similarly does not seem to be an option; rather than running into epistemological difficulties, the increased strengthening of reflection principles in the case of set theory leads to inconsistency (by Tait's theorem). Moreover, the consistent versions of the principle still leave many propositions undecided (by Koellner's results).

That there are two distinctive kinds of incompleteness at play in arithmetic and set theory can be seen more clearly in a second-order setting. Second-order arithmetic, while of course deductively incomplete, is fully categorical; it has exactly one model up to isomorphism, as proved by Dedekind (1888). There is thus a sense in which the incompleteness of arithmetic is *merely* deductive: we cannot prove all arithmetical truths, but using second-order resources we can axiomatize arithmetical concepts sufficiently precisely to *settle* all arithmetical questions.[26] The uniqueness

[26]There is a further question as to whether the deductive incompleteness of first-order arithmetic amounts to properly *arithmetical* incompleteness. According to Isaacson's thesis (1987, p.89),

(up to isomorphism) of the model of second-order arithmetic of course implies that all of its models are elementary equivalent, and for a time *this* property was known by the name 'completeness' (e.g. Wilder 1965).

By contrast, second-order set theory is deductively incomplete *and* incomplete in this model-theoretic sense. That is, second-order set theory is only *quasi-categorical*, meaning that even in the second-order setting, where statements like CH *are* settled, there remains room for models to disagree about the height of the hierarchy (Zermelo 1930).[27] And *this* distinctively set-theoretic incompleteness cannot be significantly reduced by reflection if Koellner's dichotomy argument is successful, since adding the most powerful consistent justifiable reflection principle to even a strong second-order set theory like Morse-Kelley will only show that the height of the hierarchy cannot be smaller than is required for the existence of $\kappa(\omega)$.

Gödel's theorems certainly give rise to a deductive kind of incompleteness: we cannot axiomatize our mathematical concepts in a way that will enable us to decide all questions about the consistency of the theory, for instance. By looking at incompleteness in a second-order setting, we see that there must be incompleteness of a second kind. We *can* axiomatize the concept *number* in a way that settles all arithmetical questions, even if the answers to some of those questions remain beyond us. In the case of the concept *set*, this appears to be beyond our reach, since a second-order axiomatization of the concept *set* will leave critical questions about the hierarchy not merely unanswered, but also unsettled.[28]

Thus we have two kinds of incompleteness which can afflict any sufficiently strong theory with an epistemically acceptable axiomatization. In chapter two, I argued that the first kind of incompleteness was *absolute* in the sense that no acceptable axiomatic theory sufficient for arithmetic can overcome it. In the set-theoretic case, no

the true arithmetical sentences left undecided by **PA** are all such that they code higher-order information, and do not express truths which can be seen as such in light of only their arithmetical content. The sentences which I have argued are *absolutely* undecidable are certainly of this kind.

[27]Correspondingly, this leaves certain height-sensitive statements potentially unsettled as well, such as GCH.

[28]It seems reasonable to distinguish amongst what I have called 'distinctively set-theoretic incompleteness' two different phenomena. It is significant that in the second-order setting, propositions like CH are settled, but propositions about the height of the hierarchy are not. We can set this additional intricacy to one side, however, since all I want to argue is that there is a substantial distinction between Gödelian incompleteness and incompleteness which is genuinely set-theoretical.

analogous case has yet been made; that will depend on the prospects of extrinsically justified axioms closing off all incompleteness not generated via Gödel's theorem. It does appear however, that only extrinsically justified axioms could finish the work here, given the limitative results established by Koellner. While there is perhaps some hope that a more fine-grained stratification of reflection principles could break the $\kappa(\omega)$ barrier without succumbing to inconsistency, this is nothing more than a slim hope. To my knowledge, no such stratification has been proposed, and moreover there would need to be some case made that the principles so stratified were intrinsically justified by the content of the concept *set*.

So this is the best case scenario for the conceptual platonist regarding the reduction of set-theoretic incompleteness: mathematical intuition can reduce incompleteness by at most what can be effected by positing the existence of $\kappa(\omega)$. But I should stress that this is the *best case* scenario, which only obtains if the concept *set* really does justify the strongest consistent reflection principles. But does it?[29]

The Justification of Reflection Principles

In the arithmetical case, justification for the relevant reflection principles was easily obtained; our evidence for the soundness of any theory was equally evidence for the soundness of that theory extended by the relevant instances of Feferman's reflection principle, with the minimal requirement that we could recognize the extension for what it was. The case of reflection principles in set theory is by no means so clear. Merely accepting that the axioms of **ZFC** are sound does commit us to some reflection principles, since weak principles like RP^1_2 are derivable in **ZFC** itself. But it isn't clear that stronger principles follow straightforwardly from the concept *set* or from the soundness of **ZFC**. Adding even the reflection principles very low-down in our stratification yields the existence of stationary classes of Mahlo cardinals and weakly compact cardinals,[30] cardinals which, on the face of it, take us far beyond the commitments incurred by accepting **ZFC**.

[29]We could, of course, follow Paseau (2007, p.33), in using a 'liberalized' conception of set which includes a reflection principle by stipulation. But Gödel's view is that such principles *already* follow from the concept *set* as we have it, so I'll continue in the illiberal tradition.

[30]That is, strongly inaccessible cardinals with the binary tree property. See Tait (2005, p.145–6) for details of these results.

Although this may be bad news for the conceptual platonist philosophically speaking, these results do at least provide some support for the interpretation of Gödel as deploying the notion of intuition in two distinct senses. If a decent argument can be made from the platonist's persepective that certain reflection principles can be known intuitively, and those reflection principles imply the existence of Mahlo cardinals, then we can interpret Gödel as making a cogent argument that propositional intuition delivers a verdict on their existence. On the other hand, if we interpret Gödel as operating simply with an objectual/Kantian notion of intuition, then his claims about the existence of Mahlo cardinals are difficult to interpret as anything but bizarre. However, showing that such a reflection principle has an intuitive justification might be quite a big 'if'!

Our candidates for intrinsically justified strong reflection principles are the RP_2^n principles and the Γ_n^2 principles. An obstacle to giving an intuitive justification for such stong priniciples is that it is far from clear that the iterative concept of *set* gets us even as far as the axioms of **ZFC**. Serious challenges have been raised to the idea that the standard axioms follow from the iterative conception, including choice and extensionality (Boolos 1971 and 1989), powerset (Parsons 2008), and, most often, replacement (all the above, as well as Potter 2004). Given that even the basic RP_2^1 reflection schema directly implies replacement, one might think that the case for an intuitive justification of any version of reflection is slim here. Indeed, Potter argues that no intuitive justification on the basis of a platonist understanding of the concept *set* is forthcoming (2004, p.224), and that it would be a 'coup' if such an argument could be provided. This puts Potter and Gödel directly at odds, with the latter claiming that reflection principles (amongst other things) are 'new axioms which only unfold the concept of set' (1964, p.261. See in particular fn.20 in the 1966 version).

So what exactly is Potter's complaint about reflection principles? The central problem is that even the most basic axiom schema of reflection is 'irredeemably syntactic', unlike other schemata that appear in the ordinary course of set theory. Potter thinks the motivation for adopting schemata such as separation is that we believe the second-order versions of the axioms which the schemata go proxy for. No reflection principle, however, has this character. The consistent reflection principles keep bad company, in that higher-order generalisations of them are inconsistent,

even if we restrict ourselves to the language of positive formulae. Hence, any reflection scheme that we might accept cannot be seen as a schematized version of a higher-order principle which follows from the iterative conception (assuming that the iterative conception is itself 'consistent' in some informal sense). So reflection principles are *inherently* schematic.

The technical notion of reflection on a formula is of course syntactic, but even the informal exposition of these principles is often formulated linguistically: nothing we *say* can uniquely characterize the hierarchy, as opposed to an initial segment. Formulated a bit more precisely by Fraenkel, Bar-Hillel, and Levy, 'there is no property expressible in the language of set theory which distinguishes the universe from some "temporary universes"' (1958, p.118).[31]

For Potter, that schematism is inherent in reflection principles gives us reason to believe that no satisfactory justification of any version of it can be given in platonistic terms. The thoroughly syntactic nature of reflection principles means, according to Potter, that any justification for them on intuitive grounds will take on a distinctly constructivist spin, or else will lapse into theology (2004, p.224). I think that both of these charges can be resisted by the platonist, and in doing so a positive justification for stronger reflection principles can be offered (though as we shall see, there are further problems to be dealt with). The charge of constructivism will be addressed in the next subsection; for now I'll sketch an approach which might allow the conceptual platonist to embrace reflection principles and evade the charge of theology and mystery.

Potter's Theology Argument

In what sense does Potter consider the typical justification of reflection principles to be theological? The idea is that just as the theologian might suppose that God is so far beyond finite creatures that nothing we can say could adequately represent divinity, so might the platonist suppose that the hierarchy is beyond characterization in the language of set theory. An example Potter uses is the height of the universe; presumably the platonist thinks V *does* have a height (2004, p.224), and reflection

[31]Similar informal expositions in terms of assertions, statements, formulae etc. can be found in (Horsten and Welch 2016a, p.8), (Koellner 2009, p.208), (Tait 1998, p.473), and many other locations.

principles tells us that we can't express this. The reason we can't is allegedly some kind of divine mystery.

The challenge for the platonist, then, is to offer a satisfactory picture of the hierarchy according to which it has no particular ordinal height. If the case can be made for thinking that the universe does *not* have an ordinal height, then there will be no remaining divine mystery about why we cannot say how high the hierarchy is. I think that for the platonist the right approach is to deny that the hierarchy is some particular 'definite totality' (to borrow Dummett's phrase), or well-defined higher-order object. Such a conception of set theory renders platonism about sets consistent with a denial of any particular height to the hierarchy, as we shall see below.

Potter's complaint strikes me as very persuasive in the context of a Cantorian view of set theory, according to which the hierarchy is a completed infinite totality. If we think of V as a particular (higher-order) object which comprises a model of the axioms of set theory, then it seems perfectly sensible to enquire about the height of this model, ask which ordinals it contains, and so on. After all, we can sensibly make such enquiries of ordinary models of set theory. But the various reflection principles tell us that we cannot express truths which serve to distinguish V from its initial segments, and this fact demands an explanation. On this picture, V is a model of set theory, so it must have some height or another. Hence it really does begin to seem that if we mortals cannot talk about such things, this can be little more than a divine mystery or a quirk of our language. Thankfully, Cantorianism is not the only conception of set theory available to a platonist here.

An alternative to the Cantorian conception of V as a completed totality is offered by Zermelo. According to Zermelo, the realm of set theory is not confined to some particular model of the axioms; rather there is an unbounded sequence of models of set theory indexed by 'boundary numbers' (strongly inaccessible cardinals, in today's terminology) (Zermelo 1930, p.1233). The set-theoretic realm cannot be regarded as a completed infinite totality, according to Zermelo, because the sequence of ordinals cannot be so regarded either.

This is because a key part of Zermelo's conception of set theory is that the paradoxes of näive set theory are soluble only by the elimination of proper classes. The idea is that a genuine solution of the paradoxes should not resort to higher-order

entities which are either themselves paradox-ridden, or which solve the problem by magic. Theories of the first kind would be those in which a 'Russell class' could be formed, while theories of the latter kind would include those that distinguish sets from classes only by the stipulation that classes cannot be members of classes. While theories of the latter kind would be paradox-free in a pedantic sense, we cannot accept that the paradoxes are solved in such a way without an antecedent distinction between sets and classes that has some intuitive force.

The reason why Zermelo's iterative set theory is seen as genuinely avoiding paradox is because the iterative conception has intuitive force aside from its apparent consistency. Moreover, Zermelo's conception of set theory is *class-free*, in the sense that 'what appears as an "ultrafinite non- or super-set" in one model is, in the succeeding model, a perfectly good, valid set with both a cardinal number and an ordinal type' (Zermelo 1930, p.1233). Plainly, the universe itself never appears as a set in any further models of set theory, so a platonist who adheres to Zermelo's conception of classes needn't suppose that the universe of sets itself is a well-determined model of the theory, which has either a cardinal number or an ordinal type.

For the platonist who refuses to think of V as a particular class which models the axioms of set theory, the question 'what is the height of the hierarchy?', has only the near-trivial answer that it is as high as it possibly could be.[32] No particular ordinal can be offered as the height of the hierarchy, since by reflection any non-trivial answer we offer will only succeed in characterizing the height of an initial segment of the hierarchy.[33] Hence, the hierarchy does not have some particular ineffable height, and Potter's charge of theology is defused.

Note that if we take this class-free conception seriously, we cannot even treat 'V' as a genuine singular term. I'll continue to use that expression, and related expressions such as 'the hierarchy', with their normal singular grammar, since to do otherwise would render the discussion needlessly difficult to parse. But strictly speaking, these expressions should be understood in the Zermelian context as *plural* terms (in the one case, a plural proper name, and in the other, a plural definite

[32]This idea of inexhaustibility is of course a familiar justification for reflection principles of varying strength.

[33]Assuming the answer is given is in set-theoretic terms. But it's unclear to me how that could be avoided while still specifying some *ordinal* as the putative height of the hierarchy.

description) denoting the models of set theory. As with other plural terms, such as 'the inhabitants of London', we must take care not to be mislead by the presence of the definite article and interpret the term as singularly referring to a class; particularly those of us who want to deny that the hierarchy so conceived has an order type or constitutes a model of set theory. Indeed, according to the Zermelian conception, the models of set theory do not constitute a class which is itself a model of set theory any more that the inhabitants of London constitute a class which itself inhabits London.[34]

The claim that the hierarchy does not form a totality with a definite height may have implications beyond the scope of this discussion. Tait (1998, p.478) suggests that thinking of the universe in this way should require that unrestricted quantification over V be interpreted constructively, rather than classically. One might worry here that there is a tension between platonism about sets and thinking of the universe in this way. However, recall (§1) that, for Gödel at least, inexhaustibility is a property of set *theory* just as much as it is a property of the hierarchy. So the constructive reasoning here is supposed to correspond not to an anti-realist view of sets, but rather to our inability to circumscribe once-and-for-all our means of forming axiomatic extensions of set theory. Such a a view does not straightforwardly commit one to a constructivist view of any *sets* since thinking of set theory as indefinitely extensible in this way is consistent with the logic of every segment of the hierarchy which models $\mathbf{ZFC}_2$ being classical. In other words, on Tait's proposal $\phi^\beta \vee \neg\phi^\beta$ is valid for each strong inacessible β and each formula ϕ, so this position need not be radically revisionist about set theory.[35]

While I think it is perfectly compatible with Gödelian conceptual platonism to understand talk of the hierarchy as talk about an unbounded sequence of models of set theory, which is 'absolutely infinite', or inexhaustible, there is a further question as to whether this was Gödel's own view. With characteristic caution, he does not squarely come down on either side of the debate. Horsten and Welch cite Gödel as a full-blown Cantorian, although the evidence they offer for this claim is rather thin. They refer (2016a, p.13) to an unpublished remark of Gödel's cited by Wang

[34]See (Oliver and Smiley 2016) for more on plural terms, particularly §5.3.

[35]This suggestion accords well with Zermelo's claim that in each model in the unbounded series 'the whole classical theory [of sets] is expressed' (1930, p.1223).

(1996, p.260): 'To say that the universe of all sets is an unfinished totality does not mean objective indeterminedness, but merely a subjective inability to finish it'. As it stands, this remark is consistent with the position I'm urging the Gödelian to adopt. Zermelo himself may have been a potentialist about the universe of sets, as suggested by his claim that the hierarchy 'reaches no true completion in its unrestricted advance' (1930, p.1223). But all that is required for present purposes is that the hierarchy not be thought of as a unique higher-order object such as a class which constitutes a model of the theory - whether indeterminate or otherwise. Tait, somewhat more persuasively, cites remarks of Gödel's which imply, but do not state, that V is a model of the axioms of set theory (1998, p.477). But since a Zermelian view is consistent with Gödelian platonism, even if Gödel himself was inclined to Cantorianism, we can simply regard Potter's theological argument as showing that he was wrong to be so inclined, given his acceptance of strong reflection principles.

The upshot is that reflection principles needn't be understood as saying of some well-determined higher-order object, the hierarchy, that it has some height, though we cannot say to which ordinal this height corresponds. If we regard classes merely as temporary universes, there is no mystery why we can't specify the height of the hierarchy - it is not the sort of thing to which a non-trivial ascription of height can be made. Rather, reflection principles can be understood as saying of our language that it can't uniquely pin down the hierarchy. And this is no divine mystery, it is simply a consequence of the fact that we take the closure of any principle used to form all the sets so far to form still further sets.

Since it is inexhaustibility which motivates any form of reflection principle, we have good reason to prefer the strongest consistent reflection principle - the unbounded sequence of models of set theory goes on *as far as possible*. The inexhaustibility of the hierarchy, which is the traditional justification for various reflection principles, is the key to seeing why the charge of theology is misplaced, at least in the context of a Zermelian conception of set theory.

Third-Order Platonism

What of the other part of Potter's argument against reflection principles? Even if we deny that V is some particular model of set theory, Potter's claim that reflection

principles are interminably syntactic still holds good. But should we follow Potter in thinking that this syntactic character results in justifications of these principles taking on a 'constructivist spin'? I think not; platonists as well as constructivists must speak about the hierarchy in a language, and there is no reason *a priori* to suppose that the nature of the hierarchy as understood by a platonist might not restrict our ability to speak about it in certain ways. Once we have abandoned the idea that the hierarchy is a unique well-defined class (or similar higher-order object), the way is open to a platonist to explain why various reflection principles are deliberately schematic, rather than schematized versions of a more general principles (as with the separation schema, for instance).

The central motivation for reflection principles is the notion of the *absolute* infinity, or inexhaustibility, of the universe of set theory. Given the view that classes are merely 'temporary', the notion of absolute infinity, or of inexhaustibility, cannot possibly be explained by saying that the universe is 'class-sized' or anything along similar lines. As Reinhardt puts it, 'our idea of set comes from the cumulative hierarchy, so if you are going to add a layer [of classes] at the top, it looks like you just forgot to finish the hierarchy' (Reinhardt 1974, p.196).

On this picture, classes are just sets higher up in the hierarchy, and so the universe isn't a class. Since set theory is in the business of telling us about sets, amongst which we will never find the hierarchy, we should indeed *expect* our language to be impoverished with respect to speaking of it in completely general terms. That is, we should expect any attempt to characterize it from below to result in failure. Given that set theory is the theory of sets, and not the theory of things we might say about sets, we should never expect reflection principles to be formulated as single sentences. So the reflection principles are not just 'inherently' schematic, they are *deliberately* schematic.

I think that line of thought is sufficient to establish that a platonist perspective on the hierarchy can be used to justify schematic principles that are not schematized versions of of higher-order principles. This element of Potter's criticism can be safely set to oneside. Nonetheless, reflection principles cannot be regarded as schematized versions of higher-order principles *because* some of those higher-order principles are inconsistent. So it remains to be explained how the platonist can justify as following from the concept *set* a principle which, in its most general form, is inconsistent.

Put otherwise, reflection principles keep bad company, and the platonist who would endorse strong consistent reflection principles, such as Gödel, owes us a justification which does not extend to the inconsistent principles in the vicinity.

So where do we stand? I do not think that the platonist must resort to theology or constructivism in order to justify reflection principles. Indeed, we've already seen that the inexhaustibility of the hierarchy, considered in Zermelian terms, offers a justification for strong consistent reflection principles. But reflection on a third-order parameter is inconsistent, so this justification must be supplemented with an account of why properties of properties of sets don't reflect, unlike properties of sets.[36]

This, I think, is the most serious obstacle in the way of any justification of strong reflection principles. Somehow, the case must be made that our set theory, at least as far as reflection principles are concerned, must be at most second-order. The platonist has two options here; the first of these concerns the formalization of the theory, and the second its interpretation.

The first option is for the platonist to put the theory before its subject matter, and argue directly that the formalization of set theory as a whole should not include third$^+$-order quantification (that is to say, quantification of third-order and higher). If a good argument for this can be made, then the restriction of the reflection principle to exclude higher-order parameters can be made without an *ad hoc* manoeuvre to rescue an otherwise reasonable principle from inconsistency. The best way to motivate some restriction of this kind would be to focus on the correct *theoretical primitives* required by a theory of sets. It is somewhat in philosophical vogue to suggest that the terms of a theory in general, and the primitive terms in particular, should have as their semantic values only those objects and concepts which are in some sense suitably *fundamental*.

Without getting bogged down in the details of what this 'fundamentality' should amount to in the case of set theory, it's quite clear that the first-order quantifiers should range over only sets (and, if suitably fundamental candidates can be found, urelements as well). From this point, there would be a question about whether second-order quantifiers ranging over properties of sets should be admitted to the

[36]This issue is not put forward explicitly in Potter's argument. But I take it to be a natural extension of the points he does explicitly make.

theory as well. Gödel certainly insisted (1947, p.181 fn.17) that *property of sets* was a primitive term of set theory, and we could perhaps take this as a stipulation. Alternatively we could look for historical justifications of such second-order quantification in the work of Zermelo and other early set theorists working in the iterative tradition.

From here, the strategy would finally require that considerations of fundamentality ruled out any third$^+$-order quantification. Perhaps the semantic values of such variables are too far removed from the concern of set theory for the corresponding vocabulary to qualify as legitimately fundamental. Or perhaps we could claim that quantifying over properties of properties of sets is too ideologically profligate to feature in our best account of the foundations of mathematics.

This line of thought, however, is not one I think the platonist should follow. For one thing, it isn't obvious that a suitable account of what 'fundamental vocabulary' amounts to can be given.[37] Secondly, even if such an account could be given, the fundamentality of a theory couched in a second- or first-order language could surely only count as one theoretical virtue amongst many that a formalization of set theory might have. Such a theoretical virtue could in principle be overridden by other considerations, but the platonist would require something much more substantial than this. Indeed third$^+$-order language must be completely excluded from the formalization of set theory, on pain of inconsistency or rejection of the idea behind reflection principles. If the justification of a principle like RP^2_2 is not to over-generate, it cannot simply be that virtue saves us from inconsistency. Rather, there must be some prior reason to think that properties of properties of sets don't reflect.

The alternative to focusing directly on the language would be to focus instead on the ontology of the theory. After all, there seems to be nothing stopping us adding the language of higher type theory to a theory of sets if we want to. For the platonist, the hierarchy is as it is independently of what we do or say. So what's needed is a philosophical story that distinguishes reflecting types from non-reflecting ones. A first suggestion would be to refuse to interpret the third$^+$-order part of any theory realistically. If properties of properties of sets don't actually exist, then it seems reasonable to insist that they don't reflect.

[37]For an account of why I don't think suggestions like this will work in the more general case, see Wrigley (2018).

In a sense, this proposal may seem quite natural in the context of a class-free interpretation of set theory; the values we assign to higher-order variables in formulating reflection principles are defined only relative to the restriction of the hierarchy to an initial segment. This is because we said that for a given ordinal β, first-order variables range over sets in V_β, whilst higher-order variables range over sets higher up in the hierarchy. So it's quite open to the Gödelian platonist to insist that where quantification over the hierarchy is absolutely general, there is no room to be made for third$^+$-order variables to have any values whatsoever.

A glaring issue with the proposal, however, is that a similar argument could be run with the aim of eliminating the second-order quantification in set theory that Gödel and Zermelo commit themselves to. The natural interpretation of second-order quantifiers in a class-free setting is as ranging over properties of sets. If we cannot interpret third-order variables as ranging over properties of properties of sets, for instance, then why should we be able to interpret the second-order variables as ranging over properties of sets?

The platonist could, of course, simply bite this bullet and reject second-order set theory, accepting that this is the price to pay for a strong reflection principle. There are two deep issues with this proposal, however. Firstly, RP^1_1 is extremely weak; with the axioms of extensionality, foundation, and separation, it can prove the axiom of infinity, but fails to deliver even replacement. So giving up on second-order set theory is a heavy price to pay for no gain whatsoever, if the axiom of infinity has an independent motivation from the iterative conception as is usually supposed.

Secondly, we must resist this suggestion of first-orderism if we want a satisfactory account of the different kinds of incompleteness discussed above. Given that (full) second-order logic has no sound and complete deductive system, it is of somewhat limited use for actually proving things about sets, and for that reason we might prefer our mathematics proper to be first-order. But the metatheoretic properties of second-order axiomatizations are crucial for an adequate philosophical account of incompleteness. To the committed first-orderist, there is no sensible distinction to be made between the inability of an axiomatic set theory to decide its own consistency, and its inability to uniquely determine the height of the hierarchy. But these forms of incompleteness are different and should be kept distinct, so the conceptual platonist must find a means of legitimising second-order quantification and exorcis-

ing anything third$^+$-order.

Clearly the platonist is walking a tightrope here: in an ordinary case, the use of third$^+$-order reasoning is justified where the domain is not conceived of constructively. Such reasoning is simply uncommon due to the lack of training in, and aptitude for, higher-order reasoning from which the vast majority of us suffer. There is, for instance, no obstacle *in principle* to a carpenter considering the properties of properties of wood in the construction of furniture, even if this would be unusual. So why can't the set theorist do similarly?[38] The conceptual platonist who advocates a strong reflection principle must think that reasoning about properties of properties of sets is not merely difficult or inconvenient, but for some reason cannot *possibly* be a legitimate source of knowledge about sets, while at the same time thinking that second-order reasoning is perfectly above board, because the second-order domain is suitably reflective.

I am sceptical that there is any satisfactory story to tell here. I cannot, however, rule out the possibility, as any attempt would rely on drawing a very fine distinction between properties of sets, and properties of properties of sets. As a Gödelian might put it, in order to rescue the strong reflection principles, we must think that the concepts *set* and *property of sets* have objective content, while denying this of the concept *property of properties of sets*. This is the only way I can imagine to explain why first- and second- order parameters reflect, but third$^+$-order parameters do not.

This is not an entirely happy proposal, since it is quite unclear why the concept *property of properties of sets* should not have objective content.[39] While it is perhaps true that no axioms of third-order set theory force themselves upon as true, this seems best explained by their unfamiliarity, rather than by some flaw in the relevant concept. As we saw before, the force of the axioms need not be *immediate* to us on Gödel's account.

This is certainly a problem, but it's important to note that it isn't a *new* prob-

[38]Thanks to Michael Potter for this particularly helpful example.

[39]Looking for help in Gödel's writings is of limited help here. In an early paper (1933, pp.49–50), Gödel does tell us a little about when concepts are 'objectionable'. But the example here is the concept *property of integers*, and its objectionable status is intimately linked with its impredicativity. This is a view that Gödel clearly changed his mind on, and the paper pre-dates his mature platonism by some years. In general, finding details in Gödel's papers about when a concept *lacks* objective content seems to be even more challenging than finding positive remarks on the subject.

lem. Rather, it is simply a manifestation of the same problem that we encountered in the previous exposition of conceptual platonism, namely that we have no reliable and precise method for determining whether a concept has objective content. If we had a means of so determining it, we could apply this method to the concept *property of properties of sets*. If the concept has objective content, then we would have to reject the idea behind the reflection principles as ultimately incoherent. If not, we could accept RP_2^n and Γ_n^2 for all $n < \omega$ as implied by our concepts of *set* and *property of sets*.

So, even if a suitable account of the objectivity of the content of a concept can be found, something in conceptual platonism has got to give. That is either the stronger consistent reflection principles, or third$^+$-order platonism. This represents a genuine compromise from a Gödelian perspective: Gödel explicitly says that such higher-order concepts *are* legitimate in set theory, but that they are simply redundant. He writes: 'concepts of "property of property of set" etc. can be introduced. The new axioms thus obtained, however, as to their consequences for propositions referring to limited domains of sets (such as the continuum hypothesis) are contained (as far as they are known today) in the axioms about sets.' (Gödel 1964, p.260 fn.18). The very same paragraph of text contains another footnote (in the revised 1966 version), in which Gödel mentions that the existence of Mahlo cardinals follows the concept set, and can be obtained using reflection principles. Of course, we now know that using third-order set theory with a reflection principle over third-order parameters, we can deduce *all* propositions about limited domains of sets, thanks to the inconsistency of the theory. So a conceptual platonist cannot both regard strong reflection principles as having an intuitive justification *and* be a platonist about the concept *property of properties of sets*.

Let's circle back to our central concern: Koellner's results make it very likely that only a modest reduction in set-theoretic incompleteness can be obtained from any consistent reflection principles. Some of the obstacles to a compelling justification of strong reflection principles can be overcome: in the context of a Zermelian interpretation of set theory (which may be some distance from Gödel's own position), the conceptual platonist can regard strong reflection principles as justified by the inexhaustibility of the hierarchy, without recourse to theology or constructivism. The reflection principles are not a failed approximation of more justified principles, they

are deliberately schematic, and follow from the iterative conception to the extent that the inexhaustibility of the hierarchy does. The existence of such a justification allows us to make sense of Gödel's claim that that mathematical intuition can deliver the existence of small large cardinals, since the existence of such cardinals follows from principles like RP_2^2. The more pressing problem however, is that since the time at which Gödel was writing, it has been determined that the inexhaustibility justification over-generates to inconsistent reflection principles. So even the modest reduction of incompleteness offered by reflection principles is conditional on the platonist's ability to draw a principled distinction between second-order parameters, which are reflective, and third-order parameters, which are not.

I've argued that the only means by which a platonist could achieve this is by showing that the concept *property of sets* has objective content, and that the concept *property of properties of sets* lacks it. Even if this can be done, the platonist cannot retain their full realism *and* the stronger reflection principles, which represents a significant philosophical compromise. Moreover, this serves to deepen the central mystery which bedevils conceptual platonism: which concepts have objective content, and why? Until a satisfactory answer to these questions is provided, we must be sceptical of the idea that set-theoretic reflection principles are a legitimate means of reducing set-theoretic incompleteness at all.

4.4 Platonism and Incompleteness

We now turn to the second central question of this chapter: do the incompleteness theorems support platonism (construed as conceptual platonism), as Gödel claims? I'll argue that they do, both for the reasons Gödel offers in the Gibbs lecture, and for reasons relating specifically to the form of platonism I've reconstructed from Gödel's work. However, I will argue that in both cases the support offered is unfortunately thin. The Gibbs lecture offers three distinct arguments for the claim that the incompleteness theorems support platonism. Assessing these in turn, we see that none are sufficient to *establish* platonism, although they do tell against some of its less-convincing rivals. Note that Gödel takes the incompleteness theorems to support platonism whichever disjunct of the disjunctive argument holds (1951,

p.314).

The Argument from Clarity

Gödel's first argument for platonism, contrasted with the view that mathematics is a 'free creation' of the human mind, concerns the relation between creator and creation:

> [I]f mathematics were our free creation, ignorance as to the objects we created, it is true, might still occur, but only through lack of a clear realization as to what we have really created (or perhaps, due to the practical difficulty of too complicated computations). Therefore it would have to disappear (at least in principle, although perhaps not in practice) as soon as we attain perfect clearness. However, modern developments in the foundations of mathematics have accomplished an insurmountable degree of exactness, but this has helped practically nothing for the solution of mathematical problems. (1951, p.314)

The idea is that, with respect to concepts of our own devising, a perfectly clear and exact axiomatization should be sufficient for complete knowledge. Thanks to the incompleteness theorems then, mathematical concepts (or at least, sufficiently rich mathematical concepts) are not our own 'free creation', but rather they have objective content.

Sense can be made of this argument outside the context of conceptual platonism, but it is significant that Gödel moves seamlessly between the idea of *mathematics* being our own creation, and that of mathematical *objects* being so created. Just as Gödel's platonism about mathematical concepts or mathematical truth is the source of his platonism about mathematical objects, his hypothetical anti-platonism about mathematical objects is founded on the supposition that mathematical concepts lack objective content.

Though this is Gödel's most compelling argument that the incompleteness theorems support platonism, it might initially seem that there is a problem with the inference made by Gödel, that if we created mathematical objects, then conceptual clarity with respect to them would result in all questions about them being decidable

in-principle. It seems perfectly consistent to think that sets are 'of our own making', and that perfect conceptual clarity would not result in complete knowledge about them. Chairs, for instance, are of our own making, and yet perfectly clear blueprints for the construction of a chair are not sufficient for a complete theory of its physical properties. As Gödel points out (1951, p.312), this argument should not be pressed by the creationist; the reason that clarity about physical constructions is insufficient for knowledge of their physical properties is precisely because the physical material from which they are constructed is objective.

Even in the mathematical case though, for a creationist who does not take mathematical objects to be constructed from something objective, it might seem dubious to suppose that perfect clarity would be sufficient for complete knowledge. For example, no amount of reflection on the concept *set* could decide the cardinality of the set of urelements.

I think such a response to Gödel would be much too quick. Our commitment to the soundness of some sufficiently strong mathematical theory commits us to accepting as further axioms some sentences expressing this soundness (e.g. instances of Feferman's reflection principle), as well as sentences expressing consequences of the system's soundness (e.g. its Gödel sentence and canonical consistency sentence). Despite Gödel's focus in the above passage on the objects of a mathematical theory, the problem is much more pressing when formulated in terms of axiomatic considerations: can the philosopher who believes that mathematics is our own free creation make sense of the idea that certain axioms are 'missing' from a theory?

It's quite important that the missing axioms are ones that we are obligated to regard as valid. The formulation of **ZFC** does not, for instance, force us to regard any axiom candidate stating the number of the urelements as valid, whether or not the other axioms are sound. But accepting that the axioms are sound *does* force us to regard an axiom expressing the consistency of **ZFC** as valid. It is not so obvious that the advocate of free creation can make sense of this situation, or of this distinction between such different kinds of axiom candidates. Why can't the free creator simply stop at some desired theory, such as **ZFC** and refuse to accept further axioms like the relevant Gödel sentence?

The kind of response that we can offer to Gödel on the creationist's behalf will depend critically on the kind of creationist view being considered. It's plausible

that Gödel's argument roundly defeats the crudest kind of creationism, whereby the truth of any axiom in mathematics is up to the mathematician, on a case-by-case basis. It simply can't be that the axioms of some theory are sound, and that its consistency sentence isn't valid. There are, however, issues with offering Gödel's argument as a response to platonism's rivals that are less permissive than this crude creationsim.

The turning point of the issue is whether our hypothetical creationist can make sense of the notion of *unfinished* creation, in the sense that the construction of objects answering to the theory has not been achieved to the intended extent. Consider a creationst who believes that the content of the concept *set* is at least partially determined in advance, so that the creative process is not completely free,[40] or at least that the creative process is constrained by previous creative decisions made. Is there any obstacle to such a creationist believing that mathematical existence is just construction by mathematical proof?[41] It seems to me that there is no such obstacle.

According to this imagined creationist, the concept being axiomatized would function like a blueprint, and the construction of objects would proceed by showing them to satisfy increasingly strong approximations to the blueprint. Axiom candidates which are not made valid by our constructions can be distinguished into kinds by this creationist similarly to how they can be distinguished by the platonist. Obligatory axioms, such as Gödel sentences, are those which figure in the blueprint and must later in the creative process be valid with respect to our creations. Optional axioms, such as those asserting the number of urelements, are those which needn't guide our creative process at any stage.

Gödel, of course, would not be satisfied with such a creationist position, since according to him the concept acting as blueprint would, at least in the case of the concept *set*, already determine that some objects fell under it, making any creative process redundant. But this particular response to the creationist is not at all motivated by the incompleteness theorems, rather it is a consequence of Gödel's view

[40]This needn't entail that the concept has anything like 'real content' in the platonist's sense. It might, for instance, simply be that our established patterns of use of the *term* 'natural number' determine that nothing could count as an axiomatization of the concept it expresses without meeting certain constraints.

[41]The notion of proof at work here needn't correspond to proof in any given formal system.

that platonism about the objects of mathematics is entailed by platonism about the concepts. It is not much to say that, on the assumption of conceptual platonism, the incompleteness theorems support conceptual platonism.

However, this is still something of a victory for Gödel; as we've seen the incompleteness theorems do seem to rule out a kind of view where the validity of mathematical axiom candidates is simply a matter of decision on a case-by-case basis. To that extent then, the theorems support platonism about mathematical concepts. Our grasp of such concepts is reflected in the axioms we take to be valid with respect to it. If those axioms aren't 'up to us', then there should be something in the content of the concept which is correspondingly not at our discretion. But it isn't obvious that one can't think that mathematical concepts are to some extent *determinate* without being a platonist with respect to them. That we can only axiomatize arithmetic and set theory in certain ways could perhaps be explained in terms of linguistic behaviour or psychology, for instance, without recourse to Gödelian platonism.

The significance of the clarity argument for platonism will then rest on whether rival philosophical views, such as formalism and intuitionism, are committed to the creationist principle the incompleteness theorems refute. But on the face of it, the main rivals are committed to no such principle. For example, the intuitionist should not think that we are entitled to choose the axiom of foundation to be true, since it violates the principles of constructive reasoning. According to the formalist, we can't freely choose which axioms are valid where they have implications for the 'real' or 'concrete' parts of mathematics, whatever those may be identified as.

Quite apart from this central difficulty, the argument made by Gödel is in tension with his views elsewhere in the paper. Despite claiming here that we have attained perfect conceptual clarity in order to push the creationist into a corner, his later remarks contradict this. He claims, only a few pages later, that it is 'undeniable that this knowledge [of the world of concepts], in certain cases, not only is incomplete, but even indistinct' (1951, p.321). This may be unfortunate, but it does not strike me as particularly damaging to Gödel's case; the crude creationist will have a hard time explaining our obligation to regard certain axioms as valid, regardless of the current state of our conceptual knowledge.

In conclusion, Gödel's argument can claim some success, in that the incomplete-

ness theorems refute the crudest form of anti-platonism as imagined by Gödel. A further anti-platonist view which seems to come under serious pressure from the clarity argument is a Wittgensteinian identification of mathematical truth with formal provability. But it is unclear that we should regard the theorems as supporting Gödel's platonism over other anti-platonist positions that place greater constraints on the creative abilities of the mathematician.

The Argument from Freedom

Gödel's second argument for platonism concerns the kind of freedom that a creator should allegedly enjoy with respect to their creation:

> [T]he activity of the mathematician shows very little of the freedom a creator should enjoy. Even if, for example, the axioms about integers were a free invention, still it must be admitted that the mathematician, after he has imagined the first few properties of his objects, is at an end with his creative ability, and he is not in a position to create the validity of the theorems at his will. If anything like creation exists at all in mathematics, then what any theorem does is exactly to restrict the freedom of creation. That, however, which restricts it must evidently exist independently of the creation. (1951, p.314)

Gödel goes on to clarify in a footnote that the restrictions cannot themselves be freely chosen, because a genuinely free creator could will such restrictions to be satisfied (e.g. consistency) and will simultaneously that certain sentences be theorems. Gödel's argument here is persuasive for a similar reason to the previous, namely that we plainly do not enjoy the kind of radical freedom envisaged in this passage.

The main problem, however, with this argument is also similar to that of the previous argument; it lends only limited credibility to a platonist interpretation of mathematics in virtue of its conclusion being consistent with the vast majority of platonism's major rivals. No anti-realist position in the offing is incompatible with the existence of theorems in mathematics. The main target, conventionalism, is only taken by Gödel to be inconsistent with certain theorems, namely the incompleteness theorems (as argued later in (1953/9)), so it is difficult to see what this argument

adds to his case against Carnap and co. With respect to other anti-platonist positions, all of them incorporate from the beginning some serious external restrictions on the 'freedom' of the mathematician.

In Hilbert's formalism, for instance, infinitary theories are constrained by the requirement to be conservative over finitary mathematics, and it is no part of Hilbert's account of the latter area that we have any freedom over what the theorems are. In the case of intuitionism, the creative powers of the mathematician are from the outset constrained by the canons of constructivist reasoning; if it weren't then there would be nothing to stop the intuitionist constructing the reals via standard methods, for example, and the disagreement between the intuitionist and the 'classicist' would never get off the ground.

None of this is to say that the argument from freedom is not a good starting point for an argument against these positions. The idea that theorems provide independent limitations on our creative freedom is certainly a persuasive one. But more is needed to persuade the formalist and the intuitionist that those theorems which Gödel, or a classicist more generally, takes to be limiting our creative powers should be regarded as such. For instance, **ZFC** is radically non-conservative over finitary mathematics, and classical set theory is deeply non-constructive. So what is needed for the argument from freedom to work is an account of why the theorems of **ZFC**, and other such theories, should be regarded as genuine theorems, regardless of one's prior position on the nature of mathematical objects. But even if Gödel's argument isn't sufficient to resolve the debate between platonism and its rivals, it perhaps offers a fresh light in which to view it.

The second argument is therefore, like Gödel's first, a partial success: it refutes the crudest form of anti-platonism, but it seems to miss the bigger targets against which Gödel is primarily concerned to defend platonism.

The Argument from Strangeness

Gödel's third argument is perhaps the most puzzling; it concerns the relationship between integers and sets of integers:

> [I]f mathematical objects are our creations, then evidently integers and sets of integers will have to be two different creations, the first of which

> does not necessitate the second. However, in order to prove certain propositions about integers, the concept of set of integers is necessary. So here, in order to find out what properties *we* have given to certain objects of our imagination, [we] must first create other objects—a very strange situation indeed! (1951, p.314)

This argument, though certainly interesting, is difficult to assess. Not least because, despite its being one of the 'main arguments' for platonism (1951 p.341), its conclusion is merely that the situation would be very 'strange' if some form of anti-realism were true. But something stronger can be made out of this passage, I think. Can the radical creationist make sense of the idea that the integers are *wholly* their own creation, and yet think that certain propositions about them remain opaque without the creation of some distinct further objects? Much like the previous arguments, I think this one does force the creationist to cede some ground. But that said, there are still a number of points on which the argument can be criticized. For one thing, it is quite normal to rely on one creation in order to ascertain the properties of another; for instance the use of a meter stick to measure the dimensions of a table. A likely reply from Gödel would be that this response already gives the game to the platonist, since tables are built from something objective (wood, metal etc.), so if the analogy between the table and the integers hold good, then there is 'something objective' in the latter after all.

Even so, a creationist of the sort envisaged by Gödel here could protest against the assertion that the creation of the integers does not necessitate the creation of sets of integers. Perhaps according to the creationist, the integers *are* sets, and the concept *set* dictates that we always construct further sets where possible.[42] Hence the (set-theoretic) construction of the integers simply does necessitate the construction of the sets of integers.

This creator does not, of course, have the absolutely radical freedom imagined by Gödel, since the creator is bound by some maximality principle satisfied by the relevant mathematical concepts. Furthermore, if the response I've provided is the

[42]The problem of multiple reductions doesn't occur in this context. Given the sort of mental construction envisaged by Gödel, where the objects are 'imagined' into existence, presumably the mathematician can simply *imagine* that the integers are one specific set-theoretic construction satisfying the relevant axioms, rather than another.

best response on behalf of the anti-platonists, then Gödel's argument seems to imply that the *only* mathematical objects we can construct are set-theoretic, on pain of strangeness. But this does little to seriously undermine the anti-realist's case, since any standard mathematical objects taken to be constructed can just be identified with certain sets.

Once again, the argument is a partial victory, showing that the creationist must, in order to avoid an explanatorily strange situation, regard their possible creations as of a single homogeneous kind, sets. Further, they must regard the concept *set* as satisfying some maximality principle (though perhaps only a fairly weak one). But this need not mean that the concept has any 'objective content' in anything like the Gödelian sense, if the constraints on the concept *set* arise merely from linguistic convention, mathematical practice, human psychology, or patterns of language use.

In summary then, the arguments which Gödel takes to be the main arguments in favour of platonism (1951, p.314), constitute a successful refutation of a kind of anti-platonism, according to which axioms and theorems can be created at will, such that our knowledge of one kind of mathematical object does not depend upon knowledge of any other. But as support for platonism, the arguments are somewhat unconvincing. In the main, this is because the commitments that Gödel ascribes to his opponent are not commitments recognizable in many of platonism's major rivals. Hence, it is entirely unclear that the incompleteness theorems substantially support platonism, as Gödel claims.

Even if we accept Gödel's refutation of conventionalism (which *is* thoroughly grounded in the incompleteness theorems), it doesn't seem as if that argument lends any more support to platonism than it does to any of its non-conventionalist rivals. Gödel's arguments establish that the axiomatization of a (sufficiently rich) mathematical concept is not entirely arbitrary, and perhaps also that mental constructions must all be of a single kind. These constraints seem consistent with the major anti-platonist positions. But all of these arguments rely on a very generic conception of platonism and its rivals. To properly assess Gödel's claim, we need to examine the possibility of an argument based on the specific features of conceptual platonism.

Conceptual Platonism and Incompleteness

As I've reconstructed Gödel's platonism, the vindication of his claim that the incompleteness theorems support it would require showing that those theorems should increase our confidence that the concept *set* has objective content. Gödel is not *just* a platonist about sets; he also takes it that the urelements are integers (1964, pp.258-9), so some weaker support for his platonism could be had if the incompleteness theorems supported our belief in the objectivity of the content of the concept *integer*. I'll deviate slightly from Gödel here and explore the possibility that conceptual platonism gains some support by increasing our confidence that the concept *natural number* has objective content instead; plainly nothing philosophically significant is sacrificed here, and some gain in clarity is had, given the familiarity of **PA**.

The difficulty, of course, is in assessing just how it is we become confident that a given concept has objective content. In the context of set theory, we have seen that Gödel cites two factors: that the axioms force themselves upon us as being true, and that the axiomatization of the concept is a non-arbitrary matter. Presumably, these criteria for judging the concept *set* also apply to the concept *natural number*.

With respect to the first criterion, it is difficult to see how the incompleteness of some axiomatization of a concept could further any self-evident force those axioms might have. After all, the force of the axioms is a consequence of our pre-axiomatic conception of the relevant concept, and nothing at all to do with metatheoretic properties like incompleteness. But the two criteria together offer more promise.

In the spirit of the previous three arguments, we might offer something along the following lines: if the concept of *set* or *number* weren't objective, it would be (at least to a substantial extent) 'up to us' how to axiomatize it. However, it is *not* up to us how to axiomatize those concepts. In the first instance, the axioms of **ZFC** and **PA** strike us as intuitively correct. Since the axioms in question are evidently true, it follows that they are consistent, and hence the incompleteness theorems show that arithmetic and set theory are incomplete. The extension of these axiom systems is therefore a non-arbitrary matter; thanks to the incompleteness of our axiomatization, we are rationally constrained to supplement the axioms by further principles that can be derived via reflection on the concept in question which reduce this incompleteness.

So the idea is that, if a theory has self-evident axioms, the incompleteness theorems give us further reason to think that the concept axiomatized has objective content. If we accept that argument, then to some extent the incompleteness theorems do support platonism. In the specific cases where we have a concept whose axiomatization is intuitively sound, and hence consistent, the incompleteness theorems show that the axiom system will be extendible by non-arbitrary means via reflection principles, provided that the axiomatization is sufficiently strong.

But this degree of support for platonism is unfortuantely rather weak. Namely, it only applies to cases where we are *already* convinced that a concept has an axiomatization the axioms of which 'force themselves on us as being true'. This is required to infer the consistency of the axioms, which is required in turn for the incompleteness theorems to be relevant. But having such self-evident axioms was already one of central means by which to judge that the axiomatized concept has objective content. Hence, the incompleteness theorems will support conceptual platonism only when there is already good cause to be a platonist with respect to the concept in question. And we've already seen such 'good cause' is difficult to come by, since the central weakness of conceptual platonism is the lack of clarity provided by Gödel on what the objectivity of concepts amounts to.

Another thing to note about the argument offered on Gödel's behalf is that it is substantially more convincing in the case of platonism about numbers than of sets. This is primarily because the intuitive force of the axioms is *much* greater in the case of arithmetic than of set theory, as demonstrated by the ongoing controversy over the proper axiomatization of the concept *set*.

To conclude, Gödel's claim that the incompleteness theorems support platonism can be modestly sustained. Although the form of anti-platonism which the incompleteness theorems refute is perhaps not a serious candidate for an account of mathematics, the incompleteness theorems do impose constraints on our account of mathematical objects. Those constraints are certainly satisfied by platonism, but are also satisfied by a number of serious rivals. We've also seen that the theorems more particularly support one specific kind of platonism, namely conceptual platonism. Unfortunately they do very little to remedy the lack of clarity from which that position suffers.

Another serious limitation is that the incompleteness phenomenon only lends

support to conceptual platonism in cases where we *already* have good reason to be conceptual platonists, lessening the importance of the support offered. Finally, where it does lend such support, it is more substantial in the case of arithmetic than set theory. My view is that such slender support cannot possibly outweigh the serious lack of clarity from which conceptual platonism suffers in its account of the content of concepts. If such a gap could be filled in adequately, then some further support for the position is forthcoming on the basis of the incompleteness theorems. But this support is somewhat less substantial than we might have hoped for.

Conclusion

This chapter had two central goals, each of which shall be reviewed in turn. The first was to assess whether reflection principles were any more effective at reducing set-theoretic incompleteness than were soundness reflection principles at reducing Gödelian incompleteness in arithmetic. The second was to assess whether (conceptual) platonism is supported by the incompleteness theorems, which is one of the major claims made by Gödel in the Gibbs lecture.

Firstly, we saw at work Gödel's claim that certain set-theoretic axioms can be justified by appeal to propositional intuition. Far from being mystical, this is simply the view that reflecting on the iterative conception of set can go some way towards telling us which axioms are justified. In particular, we saw that Gödel takes *inexhaustibility* to be a crucial part of the concept *set*, and that this principle potentially licences a strong series of reflection principles. However, we saw that such axioms could not succeed in eliminating incompleteness, since Koellner's theorems make it extremely likely that the strongest consistent reflection principles cannot justify axioms stronger than those asserting the existence of the Erdős cardinal $\kappa(\omega)$. This leaves plenty of incompleteness, such as the undecidability of axioms asserting the existence of even larger cardinals, together with strong reasons to believe that such undecidability cannot be removed by appeal to axioms with only intrinsic or intuitive justification.

We then turned to the question of whether reflection on the concept *set* can licence even such a modest reduction in incompleteness. I defended consistent higher-

order reflection principles against Potter's charge of theologicality, showing that the priciples are well-motivated according to a class-free conception of set theory according to which the hierarchy is an unbounded series of models, and not some particular higher-order object.

A more serious problem is how to account for the supposed fact that first- and second-order reflection follow from the concept *set*, but that higher-order reflection principles do not. I argued that restricting our attention to vocabulary demarcated as 'fundamental' would be unlikely to solve this problem, and suggested that the Gödelian would be better-served by abandoning platonism about properties of properties of sets, on the ground that we have an insufficiently clear conception of the axioms of third-order set theory. The proposal is not wholly satisfactory, since it exploits the ambiguity in the conceptual platonist's position regarding which concepts have objective content. My own inclination is to think that the reflection principles should be rejected as axiom candidates, but I conceded that the conceptual platonist has some means of justifying the existence of (at most) $\kappa(\omega)$ by propositional intuition, if they are willing to pay the philosophical price and abandon platonism about third$^+$-order objects. But even this proposal awaits a satisfactory clarification of the idea of the objectivity of a concept's content. So for now, even these relatively modest reflection principles must be regarded as without sufficient justification.

Turning to our second question, we saw that even by Gödel's own lights, it is mistaken to think that the incompleteness theorems provide very substantial support for conceptual platonism. The main arguments offered in the Gibbs lecture were shown to be quite persuasive, but not against platonism's most credible rivals. An attempt to construct a tailor-made argument from the incompleteness theorems to conceptual platonism met with equally limited success. Even here, the best that could be achieved was the conclusion that the incompleteness theorems should strengthen conviction in conceptual platonism when a substantial level of support for that position is already available. Worse still, the argument from the incompleteness theorems is more convincing in the case of platonism about the numbers, falling short of substantially increasing our confidence in the existence of sets.

So far then, we've seen that there are at least two sources of incompleteness which can afflict a formal theory, and that neither can be eliminated by the use of reflection principles. In the next chapter, I'll examine whether the distinctively set-theoretic

element of incompleteness can be justifiably reduced by the use of large cardinal axioms stronger than any consistent reflection principle, a programme launched by Gödel on the basis of a perceived analogy between mathematics and natural science.

Chapter 5

Quasi-Scientific Methods of Justification in Set Theory

Introduction

In the previous chapter, I argued that the conceptual platonist could deploy intrinsic methods of justification to posit the truth of axioms asserting the existence of cardinal numbers of at most the consistency strength of the axiom that there is a partition cardinal $\kappa(\omega)$ (though not, as we saw, without some cost). But stronger axioms for larger cardinals certainly can be formulated, and according to Gödel's view, justified by *extrinsic*, or quasi-scientific means. In this chapter, I'll refer to such axioms which cannot, on the Gödelian view previously sketched, be justified by appeals to intuition as *large large cardinal axioms*.[1]

According to Gödel, platonism about set theory is the primary foundation for the use of quasi-scientific methods in justifying large large cardinal axioms. In the case of the natural sciences, the real existence of the objects concerned justifies the use of 'probabilistic' or inductive methods to reach decisions about the nature of those objects, which would not make sense if they were regarded as useful fictions or mental constructions. According to Gödel, the same realistic attitude to the objects of set theory establishes an analogy between mathematics and natural science which is sufficient to employ analogous methods in establishing an over-all picture of the

[1]It is important to bear in mind that such talk is loose, in an important sense. Large cardinal axioms *appear* to be linearly ordered by consistency strength, but there is no theorem to this effect (and there is no agreed definition of 'large cardinal property' which would be required for the formulation of such a theorem). Secondly, although I'll speak of large large cardinals and large large cardinal axioms more-or-less interchangeably, it is important to remember that the ordering of such axioms by consistency strength is not identical to the ordering of the cardinals concerned by size. A cardinal κ is *huge* iff it is the critical point of a non-trivial elementary embedding j from V into a transitive inner model M containing all functions $f : j(\kappa) \to M$ (Kanamori 2009, p.331). κ is supercompact iff for all $\lambda \geq \kappa$, there is some elementary embedding j from V to a transitive inner model M containing all $f : \lambda \to M$ and $j(\kappa) > \lambda$ (Kanamori 2009, p.298). If **ZFC** plus an axiom for the existence of a huge cardinal is consistent, then so is **ZFC** plus an axiom for the existence of a supercompact cardinal. However, if cardinals of each kind exist, then the least supercompact cardinal is far larger than the least huge cardinal (Jech 2003, p.381).

hierarchy.

The §1, I'll introduce a strong analogy that Gödel draws between sets and material bodies, namely that the former are required to make sense of our mathematical experience in the same way that the later are required to make sense of our empirical experience. I'll argue that no such analogy can be used to justify a belief in large large cardinals.

In §2, I'll introduce Russell's regressive method, which accords well with part of Gödel's thinking on the justification of axioms in set theory, whereby axioms are verified by permitting the deduction of elementary mathematical 'data', just as laws of nature in the sciences are justified by facilitating the prediction of data drawn from sense experience.

In §3, I'll examine the various options for what might constitute the mathematical data for the purposes of Gödel's analogy. These include the deliverances of so-called mathematical perception, the theorems of ordinary mathematics, and Π^0_1 arithmetical consequences. I'll argue that of these candidates, some selection of Π^0_1 sentences offers the only plausible option. Not all sentences of this form can act as data, but a reasonable delineation of some privileged such sentences can be isolated (though this delineation is perhaps not sharp).

In §4, I'll then argue that, on this construal of the data, no large large cardinal axiom gains any *strictly regressive support* by accounting for the data. By that, I mean that no large large cardinal axiom is such that it permits the deduction of a datum that cannot be deduced with help only from weaker assumptions. Hence we can't regard such posits as analogous to laws of nature with strictly regressive support.

So the only plausible respect in which they could possibly be justified by quasi-scientific means is by being regarded as principles which seek to maximise the theoretical virtues of set theories to which they might be added. Though I don't have a means of weighing and evaluating the contribution of various such virtues, I'll make the case in §5 that under no scheme for evaluating theoretical virtues should we expect large large cardinal axioms to perform well, if the virtues in question are broadly scientific as Gödel suggests. Indeed, the closer the analogy between mathematics and science, the less well-supported by the analogy are large cardinal axioms, and hence the prospects for justification of these principles by analogical reasoning

are bleak.

5.1 The Material Bodies Analogy

The use of quasi-scientific methods for justifying axioms of set theory is now commonplace in the philosophy of mathematics. Most famously, the indispensability arguments of Quine (implicit in his (1951a)) and Putnam (explicit in his (1975)) justify the truth of set-theoretic axioms by examining the role they play in formulating adequate theories in natural science. Maddy's later work (e.g. (1997)) seeks to legitimise large large cardinal axioms via the empirical study of the behaviour of actual set theorists. In contrast with the Quine–Putnam approach, which attempts to found mathematical platonism on an empirical basis, Gödel's use of quasi-scientific methods is largely internal to mathematics. The role of large cardinals is examined in terms of their contribution to a wider *mathematical* theory; little more than lip service is paid by Gödel to the applications of such theories in the sciences. And in contrast to Maddy's approach, which sees the methods of set theorists as essentially autonomous, Gödel attempts to justify set-theoretic modes of theory choice by showing them to be analogous to sound methods found in the natural sciences. So Gödel's approach to the problem has not survived the decades in its original form, despite the fact that this aspect of his thought has undoubtedly had the greatest impact in later analytic philosophy, far eclipsing the reception of his anti-mechanism, rationalistic optimism, and conceptual platonism.

Gödel's quasi-scientific approach does not annex mathematics to the sciences, and nor does it insulate the former from the latter. Rather, it *imports* some elements of scientific methodology into mathematics, by finding a structural similarity between the two. There is not, in Gödel's remarks, a unique analogy to this effect, so first I'd like to disambiguate two distinct analogical arguments presented by Gödel.

The analogical argument that is the primary concern of this chapter takes it that large cardinal *axioms* play the role in a mathematical theory that laws of nature play in a scientific theory. It is common enough to conceive of natural-scientific propositions as being divided into two broad kinds (whether or not we take those kinds to be disjoint or sharply delimited): the data and the laws. On this standard concep-

tion, the data are empirical propositions that we take to be the facts, and the laws are those propositions which are formulated in order to predict the facts. Indeed, the prediction of the data is the primary means of verifying these laws; whether or not they are intrinsically plausible, we take them to be true if they predict all the data and don't predict anything false. By analogy, certain large cardinal axioms are supposed to be verified by 'predicting' (which is to say, deductively implying) mathematical propositions of some privileged kind identified as the data. This is broadly the Russellian view of the matter, and we shall return to it in due course.

Distinctly, Gödel sometimes speaks as if sets themselves, that is, the particular objects asserted to exist by the axioms, play the role in our understanding of a mathematical theory that physical objects play in understanding our phenomenal experience (1944, p.128). It may seem that these views are not substantially different; perhaps it matters little to natural science, for instance, whether we posit the existence of physical bodies or assent to the truth of sentences asserting them to exist. I'll argue that there is, however, a substantial difference between the two views in the case of large large cardinals, and that an analogy between mathematical and physical objects cannot justify any large large cardinal axioms.

The analogy between sets and material bodies first appears when Gödel discusses his platonism about sets (though not large large cardinals in particular) and properties of sets. He writes:

> It seems to me that the assumption of such objects is quite as legitimate as the assumption of physical bodies, and there is quite as much reason to believe in their existence. They are in the same sense necessary to a satisfactory system of mathematics as physical bodies are necessary for a satisfactory theory of our sense perceptions (1944, p.128).

Although Gödel does not elaborate here on the sense in which he thinks the assumption of physical bodies is necessary for a satisfactory theory of perception, I think it is safe to assume that it is something along the lines of the following now-standard explanation from Russell:

> [A]lthough this is not logically impossible [that there are no physical bodies], there is no reason whatsoever to suppose that it is true; and

> it is, in fact, a less simple hypothesis, viewed as a means of accounting for the facts of our own life, than the common-sense hypothesis that there really are objects independent of us, whose action on us causes our sensations (1912, p.10).

Russell describes the 'simplicity' as stemming from the fact that it would be a 'miracle' (1912, p.9) if objects came and went from existence as we started and finished perceiving them. Obviating the need to believe in this miracle by positing physical bodies is described as a 'natural' theoretical move, rather than an account of how we acquired our belief that there are physical objects (1912, p.11).

Can a similar account be given of the posit that there are sets in general, and large large cardinals in particular? In an early presentation of the regressive method (which will shortly be examined in more detail), Russell (1907, p.573), takes it that 'accounting for' or 'predicting' the relevant data amounts, in the case of mathematics, to proving some given privileged propositions. In the case of large cardinals, we have it that the addition to **ZFC** of an axiom stating that there is a cardinal of some particular kind allows for the deduction of certain sentences which are not provable from **ZFC** alone. Some of those sentences might plausibly count as data, while others should not be considered as such. In particular large cardinal axioms have set-theoretic consequences (which may or may not count as data depending on the case in hand), but also arithmetical consequences which are much more plausible candidates for data. Consider, for example, the theory $\mathbf{ZFC} + \exists x\ x$ is measurable.[2] This proves that there is an inaccessible cardinal, which can hardly count as an elementary datum, but it also proves Π^0_1 arithmetical sentences not provable in **ZFC**, for example $Con_{\mathbf{ZFC}}$, which are much more plausible candidates for mathematical data.

Suppose I posit the existence of a measurable cardinal in order to prove some Π^0_1 arithmetical sentence which I believe to be true. There is an important respect in which positing material bodies in order to systematize our sense data differs radically from this. Consider, for example, the case of positing a table to account for the coherence and continuity of my table-ish experiences with respect to leaving and

[2] κ is measurable iff it is the critical point of some elementary embedding from V to a transitive class M (Martin and Steel 1989, p.73).

re-entering a particular room. In the case of the table-posit, the particular material body being posited, *that* table, plays a crucial role in the systematizing of my table-ish experiences. If 'accounting for the facts of our own life' (as Russell puts it) is to be made any simpler by this posit, it is because *that* particular table is there. The mere truth of the sentence 'there are tables' is insufficient for such purposes. Accounting for the facts of my experience is no simpler, for instance, if there is a table somewhere else, but that *this* is merely a series of sense data. Indeed that seems to rather complicate the story if some table-ish experiences are of actual tables, and some are merely of sense data. The existential generalisation over tables does not on its own systematize our experience, it is the particular posits of particular tables that perform such a function on a case-by-case basis.

In the case of the measurable cardinal, however, things are not so. It is merely the increase in the strength of our set theory that accounts for the elementary arithmetical consequences. Although it might seem natural to think that it is the least measurable cardinal which accounts for the arithmetical data in this scenario, in truth no *particular* measurable cardinal explains the arithmetical consequences. Unlike in the case of tables, the existence of *any* witness to the existential generalisation will do the job. Worse still, the role played by even the posit of a measurable cardinal is dispensable with respect to the elementary consequences, because it is only the consistency strength of the assertion which matters. For example, the addition of a measurable cardinal allows us to prove $Con_{\mathbf{ZFC}}$. But positing any stronger axiom of infinity, such as the existence of a Woodin cardinal, would do the job just as well. Arguably, it would do the job *better* since it would prove further Π^0_1 arithmetical sentences which are not accounted for by the weaker theory (e.g. it proves $Con_{\mathbf{ZFC}+\exists x\ x \text{ is measurable}}$).[3]

Something quite substantial is at stake here, because *if* large cardinals really were required to make sense of mathematics in the same way that material bodies are required to make sense of our ordinary sense experience, that would afford to them a *massive* degree of quasi-scientific justification, because the existence of ma-

[3]As long as we are considering the facts of common experience here, the argument can even be re-run with respect to theoretical physical entities. For example, some *particular* arrangement of elementary particles at a particular spatial location is required to explain why I see a table every time I go into the room. The mere existence of some such particles somewhere is insufficient.

terial bodies is a far more certain proposition than any particular natural law in the sciences. Even well-established scientific laws are at times overthrown or precisified in the face of new experimental data, as occurred with Newtonian mechanics. Moreover, even the most well-established scientific laws can rest uneasily with one another, as is the case with general relativity and quantum mechanics. By contrast, our belief in material bodies is practically certain. Most of us are inclined to agree with Russell that we can't *prove* that we are not dreaming; and yet I am more certain, for instance, that the experience of writing this chapter is veridical than I am of any philosophical conclusion reached in it. Since elementary mathematics is no less evident than the experiences we have in the ordinary course of life, if large cardinals were like material bodies in this sense, then we could be overwhelmingly confident that they exist.

This is not to say that Gödel ever seriously entertained the justification of large large cardinal axioms in this sense; where he makes these kinds of assertions he is speaking of sets *generally*. Although this of course includes such cardinals if they exist, a more charitable reading of the passage would interpret these remarks as directed toward the elementary parts of set theory. It is an interesting question which parts of set theory, if any, can be justified by appeal to an analogy with material bodies, though for my purposes it is redundant.[4] Although Gödel's argument for the existence of small large cardinals on the basis of the iterative conception of *set* is in certain respects problematic, the aim in this chapter is to examine whether larger cardinal axioms can be justified by quasi-scientific methods. The material bodies analogy, even though not offered as a specific defence of large large cardinal axioms, promises a huge degree of regressive support for the existence of certain sets, so it is significant that it cannot be used to justify large large cardinals in particular.[5]

[4]Chihara (1982) takes very seriously Gödel's claim that we have *as much* reason to believe in sets as in material bodies, and rejects it wholesale. I think this conclusion is probably correct, but Chihara gives little consideration to the possibility that we have a good justification for the belief in sets of a similar kind as the justification of the belief in material bodies, even if the quality of the justification is not equal in both cases. Given what Gödel says elsewhere about quasi-scientific justification, it seems probable to me that he suffers from an uncharacteristic lapse of caution in the passage quoted by Chihara, and that the slightly weaker position is Gödel's own. Even if not, it strikes me as an interesting possibility in its own right.

[5]Maddy (1990, p.31) takes it that the *primary* function of intuition in Gödel's epistemology is to provide intuitive data that is accounted for by mathematical theories by way of analogy to physical bodies. Needless to say, this assessment does not accord with the reading of Gödel offered

Rather, a more modest quasi-scientific justification for large large cardinal axioms must be sought, in the form of something like Russell's regressive method in which axioms are taken to be analogous to scientific laws.

5.2 Gödel and the Regressive Method

The version of the science–mathematics analogy that Gödel draws on most heavily has its origins in Russell's regressive method of finding justification for axioms. The similarity between Gödel and Russell runs deep here. Russell had in 1907 drawn a 'close analogy between the methods of pure mathematics and the methods of the sciences of observation' (1907, p.572). Here Russell claims that mathematics, like every science, has a body of commonly accepted propositions for which broader theory is supposed to account. These are known as 'data' or 'facts'. In the empirical sciences, the facts are accounted for by proposing laws of nature which collectively predict them. Analogously, in mathematics the most elementary facts are accounted for by proposing axioms which deductively imply them.

Gödel cites firm approval of this method, and predicts that it will be even more successful in the future (1944, p.121), so a more thorough analysis of Russell will assist in our evaluation of Gödel here. Of course in the 1944 article, Gödel is discussing sets generally, not large large cardinals in particular (as noted above in connection with his first analogy). Nonetheless, we'll see whether the considerations at work in the regressive method can be put to use in justifying large large cardinal axioms.

Russell sharply separates the epistemological problem, which is the problem of the present chapter, from the psychological and historical (in some cases prehistorical) problem of how we come to believe the propositions identified as data. The *empirical* premises of a belief are those propositions which cause us to believe the data, whereas the *logical* premises are logically less complex propositions from which the data is to be deduced. Take, for example, the proposition that $2 + 2 = 4$, which ought to count as common fact if anything does. Russell conjectures that the empirical premises of this belief will be various beliefs acquired from everyday life, such as ancient shepherds repeatedly noticing that two pairs of sheep is always four

here.

sheep, and similar. By contrast, the logical premises of this data will be formulae in a system of mathematical logic or axiomatic arithmetic from which '$2 + 2 = 4$' can be derived.

The point of interest for Russell is that, for the greater part of mathematics, the simple picture on which the empirical premises and logical premises coincide holds good. In other words, we believe a mathematical proposition precisely because we have a proof of it from simpler propositions which we already accept. With respect to elementary propositions, including $2 + 2 = 4$, however, this is plainly a misleading picture, since the truths of elementary arithmetic are far more evident than the axioms of any system from which they could be derived. This leads Russell to conclude that the method of discovering and justifying foundational principles in mathematics is 'substantially the same as the method of discovering general laws in any other science' (1907, p.573).

Given the similarity of methods of justification, it is unsurprising that for Russell the degree of verification obtained by axioms in mathematics is alike to the degree which may be claimed for the laws of physics. As he puts it 'when the general laws are neither obvious, nor demonstrably the only possible hypotheses to account for the [data] then the general laws remain merely probable' (1907, p.573).

There is some lack of clarity as to *which* general laws can be, or should be, justified regressively, and hence potentially without complete certainty. In the original paper, Russell seems to think that general logical laws like $\phi \to \phi$ can be justified in this way (1907, p.576). This strikes me as somewhat bizarre, since such a law seems as evident as a proposition about one's present sense data. Later on, however, Russell appears to shift into thinking that the laws of logic are self-evident upon reflection, and do not require regressive justification (Russell 1914, pp.70–71). This idea is much more appealing, since the most obvious logical laws can then be themselves considered as data on a par with elementary mathematical propositions. But it is still unclear where exactly to draw the boundary between generalities like $\phi \to \phi$ which may be considered part of the data, and generalities which are designed to account for the data, like the Peano axioms. Despite this, the proposal constitutes a radically non-traditional epistemology of axiomatic systems, in that axioms may be afforded fallible justification in the absence of *any* intuitive evidence.

Gödel's view is in many respects similar to Russell's. An element of Gödel's ap-

proach to the issue that differs from Russell's is that he is more concerned with the verification of axioms by the enhancement of what today would be called 'theoretical virtues'. In discussions of the regressive method, Russell focuses on confirmation which flows from two sources: the obvious truths which axioms or laws entail, and the obvious falsehoods which they do not (1907, p.578). Where an axiom candidate is justified because it accounts for some data which have no proof in the unsupplemented theory, I'll call such justification *strictly regressive*. But there is another respect in which posits in science can contribute to verification of a theory beyond its observational consequences, by discriminating between the virtues of competing empirically adequate theories.

To take a famous example, Einstein's theory of general relativity describes spacetime as curved. Logically speaking, we could maintain that spacetime is actually flat, and posit compensating fields. The result is a theory which has the same observational consequences as Einstein's, but which preservs our pre-theoretic Euclidean conception of the geometry of physical space. However, corresponding to each relativistic model there are infinitely many distinct but empirically indistinguishable alternative Euclidean worlds. So despite their empirical equivalence, the Euclidean alternative is not seen as a genuine competitor to general relativity, because of the latter's tremendous advantage in terms of simplicity, naturalness, and other theoretical virtues (Sklar 1992, pp.62–63).

Russell does give consideration to theoretical virtue in cases like this, where there are multiple candidate hypotheses which can account for certain data. For instance, in (1908, pp.242–243), he adopts the axiom of reducibility on the grounds that it does the work required of a theory of classes, but is considerably more convenient than a theory of classes suitably modified to avoid the paradoxes. More generally, he emphasises that axiomatic theories which predict the data serve to organize our knowledge and make it more manageable (1907, p.580). It is not apparent to me, however, that Russell regarded it as possible to justify mathematical axioms *solely* on the grounds that they substantially enhance virtue. He never, to my knowledge, offers an explicit justification for adding an axiom to a theory which has *no* strictly regressive support (i.e. one that is not sufficient to take account of any data not accounted for by the unsupplemented theory), but which does substantially enhance the the theoretical virtues of the unsupplemented theory. On the other hand, Gödel's

remarks strongly suggest that he does believe such justification to be possible (see §5). In general, Gödel places much more emphasis on this element of the analogy between mathematics and science than does Russell.

We'll postpone for now the discussion of what such theoretical virtues might be in the mathematical case. The important point is that Gödel is quite alive to the degree of revisionism in this epistemological picture. At the time of writing (1964), Gödel was sceptical about mathematicians' present ability to verify large cardinal axioms by quasi-scientific methods, though he does claim that in principle they could be verified 'at least in the same sense as any well-established physical theory', even in cases where the axioms entirely lack intuitive justification (1964, p.261). Gödel, much like Russell, is clear that certain axioms may possess both intuitive and quasi-scientific justification (1944, p.121), but the main focus is on axioms without *any* intuitive force, such as large large cardinal principles. The verification of such axioms is described as being 'only probable' (1964, p.269).

This all stands in sharp contrast to the epistemology of intuition discussed in the previous chapter; although Gödel did not take intuition to be infallible, it seems he thought that the existence of certain large cardinals could be established definitively by such methods. If however, our mathematics were to make use of axioms possessing *only* quasi-scientific justification, 'mathematics may lose a good deal of its "absolute certainty"' (1944, p.121). Gödel's confidence in his revisionist epistemology is such that in the Gibbs lecture he even claims that the monopoly of deriving 'everything by cogent proofs from the definitions' may turn out to be 'as mistaken in mathematics as it was in physics' (1951, p.313).

In summary, Gödel's introduction of quasi-scientific methods into the theory of large cardinals constitutes a stark departure from his more traditional epistemology of arithmetic and the more basic elements of set theory. He hopes that some justification of set theory can be offered based on two analogies. The first is that sets help us systematize mathematical experience just as material bodies do our sensory experience. We saw that this analogy was ill-founded, at least in the case of large large cardinals. The second analogy is between large large cardinal axioms and scientific laws, derived from the work of Russell. The central elements of the analogy are as follows:

1. Certain mathematical truths stand to set theory as elementary data stand to physical theories.

2. Positing large cardinals can account for this data, similarly to how laws of nature can account for the data of scientific theories.

3. Such posits can be justified either by being necessary for the deduction of elementary data, or by enhancing the theoretical virtues of theory to which they are added.

4. Consideration of such theoretical virtues can be so significant as to admit into mathematics axioms (and hence theorems) which only have a probable justification.

The next few sections of this chapter will aim to clarify the central elements of this analogy: what mathematical truths count as data? In what way can large cardinals account for this data? How do large cardinal axioms enhance theoretical virtue?

5.3 Mathematical Data

If large large cardinal axioms are to find their justification in accounting for the mathematical data in a certain way, then some delineation of which mathematical propositions constitute that data is plainly required. Though such a delineation need not be made entirely precise in order to see the force of Gödel's analogy, we clearly need some account of what the large cardinal axioms are supposed to be accounting for if the analogy with the natural sciences is to be informative at all. At the very least, some delineation of the data is required for the account to be non-trivial; if *any* mathematical truth qualified as data, then any true large cardinal axiom would be self-certifying in a way that could not be considered scientifically respectable. In this section of the chapter, I'll examine various possible accounts of mathematical data that are suggested in Gödel's writings and elsewhere, and argue that only one of these has any hope of being plausibly seen as analogous to data in natural science, if Gödel's argument is to be of any use in justifying large large cardinal axioms.

Perception and Objectual Intuition

Given the attention that has been given to Gödel's remarks on 'a kind of perception' in mathematics by later readers, we might expect that the data to be accounted for are mathematical perceptions, and set theory is verified to the extent that it 'predicts', i.e. proves, the propositions which we can perceive to be true. On such a view, there would indeed be an overwhelming analogy between mathematical and scientific theory, namely that both of them are a means of systematizing and streamlining the data we experience into a cohesive theory.[6]

I've already argued (in chapter 3) that we should understand Gödel's remarks about perception in mathematics as referring to Kantian or Hilbertian intuition, and not to perception in a literal sense. Moreover, I've argued that *intuition of* (i.e. singular objectual intuition) doesn't play a significant role in Gödel's platonistic epistemology. That said, some view whereby this faculty provides data to be accounted for regressively may still be worth considering, given the enormous degree to which it renders science and mathematics analogous. A clear statement of the view is given by Maddy (1990, pp.44–45). She claims that if we are persuaded of some kind of platonism or realism, then we should expect scientific and mathematical epistemology to be analogous. Since some scientific beliefs are pre-theoretical and non-inferential, so too should we expect this in mathematics. In science, these beliefs are formed by perception, and so in mathematics they should also be formed by perception, or something perception-like. The real problem with this view, at least with respect to the justification of large large cardinal axioms, is that it is unstable between the two main ways of thinking about mathematical perception: on one account it is far too weak, and on the other it is so strong as to be trivial.

Firstly, we might imagine that deliverances of singular objectual intuition, something like perception (but not perception itself), must be accounted for by a mathematical theory. That is, we need to provide a formal theory $\mathbf{T}$ such that $\phi \in \mathbf{T}$ if the truth of ϕ is apparent given intuition of the objects concerned (much as the

[6]This is not to be confused with Quine's view (Quine 1951a, p.45), according to which mathematics is also an attempt to systematize and streamline the data of experience. On Quine's view, there is a single kind of data, which is accounted for by scientific (including mathematical) theorizing as a whole. The view presently being considered however, posits *two* kinds of data, which are accounted for by the natural sciences on the one hand, and mathematics on the other.

truth of colour-ascriptions are made apparent by looking at the relevant objects in favourable visual conditions). There are of course questions to be raised about how such a faculty of intuition might function, but on anything analogous to Hilbert's view of intuition, what is given by this faculty will be such a tiny fraction of mathematics that no large cardinal axioms will have a role to play in accounting for it. If for example, deliverances of intuition concerning number are captured by primitive recursive arithmetic (i.e. quantifier-free arithmetic), then no objects other than the natural numbers are required to explain the data.[7] Of course, set theory with a large cardinal axiom would *also* explain this, and would also solve lots of open set-theoretic problems besides. But such an explanation would surely fall foul of considerations of simplicity and economy of both ontology and ideology.

For example, Newtonian mechanics is a simpler theory than Newtonian mechanics plus evolution by natural selection. In a perfectly Newtonian world with no living creatures, the supplemented theory would do all the explaining of the base theory, and would additionally answer lots of questions about the heritable traits of living things. But that would not make the theory any *better* because, by our assumption, there is no data for the complicated theory to account for that the basic theory could not. And similarly for large cardinal axioms (indeed, set theory in general), if we take the data to be limited in advance to what is given in singular objectual intuition.[8]

Alternatively, we might follow Maddy and think that sets themselves can be literally perceived, with no need for the surrogate faculty of intuition. As discussed in chapter 3, this gives a much richer relation between us and the objects of mathemat-

[7]The qualification 'concerning number' is required to avoid questions that could be raised about *geometric* intuition, which plausibly requires resources going beyond those available in primitive recursive arithmetic. Though there are interesting questions about such cases, here is not the place to discuss them. Complications immediately arise concerning geometrical intuition when one considers the modern conception of geometry as lacking an intended interpretation, or the possibility that geometric intuition could be explained as a spatio-perceptual faculty, rather than a genuinely mathematical one.

[8]I don't want to commit myself to the view that singular objectual intuition *is* captured by primitive recursive arithmetic. But it does seem to be a reasonable approximation (the classic presentation of this view is (Tait 1981), though Tait rejects the Hilbertian claim that the *security* of finitary arithmetic is grounded on our ability to represent its objects in intuition). Furthermore, I expect any account of objectual mathematical intuition in the vicinity would equally support a slightly modified argument to the same conclusion.

ics than traditional Kantian or Hilbertian intuition, but indeed the relation is much *too* rich. Since perception requires a causal connection between the perceiver and the object of perception, the only sets we can see, according to Maddy, are those with physical objects in their transitive closure. Moreover, *any* sets with the same physical objects in their transitive closure are co-located. A consequence of this is that for any ordinal α, there is a set of rank α where any physical object is (Maddy 1990, p.59).

When it comes to data then, there are two options. If, for whatever reason, we can only see sets of low rank, then it is hard to see how perception would fare any better than objectual intuition did. On the other hand, if we can see any set in our visual field, then there is no need for large cardinal axioms to *account for* the data, since, for any true large cardinal axiom, we'd just be able to see sets of any rank necessary to validate the axiom. I take it that the existence of large cardinals shouldn't be considered part of the data that large cardinal axioms account for, so it seems that the science–mathematics analogy cannot support large large cardinal axioms if we take the data to be given by either singular intuition or perception of sets.

Although a proper discussion would take us too far afield, I want to make it clear that I think objectual intuition is insufficient as a source of data for which large large cardinal axioms *specifically* are required to account. Perhaps objectual intuition is well-suited to providing regressive support for much weaker axioms; but according to our Gödelian account, such weaker theories can be validated by propositional intuition, hence there is no need for such a discussion here. We'll also discuss below the possibility that objectual intuition can still contribute to the data, even if it is insufficient to provide all of it.

Ordinary Mathematics

If the analogy between mathematics and natural science is to validate a large large cardinal axiom, we need a collection of data more expansive than what is given in intuition that does not include statements about sets of arbitrary rank. An initially promising class that appears to fall between these two extremes would be 'ordinary' mathematics. In emphasising the foundational role of set theory, we might examine

the theorems accepted by mathematicians in other areas, and see how strong set theory is required to be to account for these theorems.[9]

However, as a source of data, ordinary mathematics suffers the same defect as objectual intuition, because very little set theory is strictly required to account for ordinary mathematics. Gödel's own view in 1964 was that the lack of observable consequences in other fields was such that 'it is not possible to make the truth of any set-theoretical axiom reasonably probable in this manner' (1964, p.269). We find similar views decades later in the work of Quine (1990, pp.94–95), asserting that the higher reaches of set theory should indeed be *pruned* on account of their irrelevance (although Quine was of course concerned with their relevance only to *applied* mathematics). And decades after Quine, it is still difficult to find an example of a mathematical result from outside set theory which requires a large cardinal axiom for its verification. As Potter puts it 'the overwhelming majority of 20th century mathematics is straightforwardly representable by sets of fairly low infinite rank, certainly less than $\omega + 20$' (2004, p.220). So far, the 21st century shows no sign of being any different.

This is not to say that large cardinal axioms don't have *any* consequences that are of significant interest to mathematicians working outside set theory; large cardinal axioms all have number-theoretic consequences, and can at times be used to solve open mathematical problems (see §5 below). The point to note for now, however, is that such consequences are not regarded as true *in advance* of positing a large cardinal axiom, so cannot be seen as *data* for which such an axiom might account. They could, perhaps, be seen as analogous to the additional consequences that scientific theories have which are not themselves data. That is, perhaps such propositions are analogous to scientific *discoveries*. But such a proposition is not to be believed unless we believe the theory of which it is a consequence, and hence it cannot play the same role as the data do in establishing the truth of the theory. Outside of the context of large cardinals, Russell writes that 'the logical premises have, as a rule, many more consequences than the empirical premises, and thus lead to the discovery of many things which could not otherwise be known' (1907, p.574).

[9]In proposing this conception of data, is it convenient also to relegate category theory to the realm of extra-ordinary mathematics along with set theory, for the same kinds of reason in each case.

Large cardinal axioms certainly have many such consequences, but the point is that they cannot properly be considered logical premises for mathematics at large, since ordinary mathematics can do perfectly well without them.

As with the case of objectual intuition, it may well be that a study of ordinary mathematics would provide strong regressive support for certain set-theoretic axioms. However, those axioms would be substantially weaker than the large large cardinal axioms which are our present concern.[10]

Arithmetical Data: Primary and Secondary

Gödel's own suggestion is that the data should come from arithmetic, 'the domain of the kind of elementary indisputable evidence that may be most fittingly compared with sense perception' (1944, p.121).[11] This is a distinct proposal from the one just discussed. After all, not all verified propositions in ordinary mathematics are arithmetical; conversely not all verified arithmetical propositions are found in ordinary mathematics, since many of them are distinctly meta-mathematical.

On the face of it, arithmetic is a much more promising source of data than singular intuition: if the data are sufficiently rich that an incomplete theory is required to account for them, then this opens up the possibility of formulating a sequence of increasingly powerful theories accounting for more and more of the data, with no limit to the strengthening process. This is exactly what the large cardinal hierarchy promises to provide.[12]

[10]It's perhaps worth emphasising that here I'm only discussing the *strictly regressive* justification of large large cardinal axioms, not the quasi-scientific justification of them overall. At this stage, I do not consider myself to have said anything against the view that large large cardinal axioms can be verified by evaluating their 'theoretical virtues', which might include the ability of these axioms to solve open problems. This issue will be taken up below in §5.

[11]It is worth clarifying that Gödel is here using the term 'sense perception' is a specialized way, in the context of discussing Russell's regressive method. Hence 'sense perception' should here be taken to mean a proposition functioning as data that should be deducible from the wider explanatory theory.

[12]It may at first sight appear that the growth of the large cardinal hierarchy *does* have a limit. We know, for example, that Reinhardt cardinals are too large to exist if the axiom of choice is true. This is an easy consequence of Kunen's theorem that there is no non-trivial elementary embedding from the universe into itself, since a Reinhardt cardinal is the critical point of just such an embedding (Kanamori 2009, pp.318–319). While such cardinals *are* too large, I can't see how this differs from the requirement that the large cardinal axioms must be consistent. The formulation of cardinal axioms too strong for **ZFC** does not imply that *within* the class of large

Since there are a great many arithmetical truths yet to be formulated, let alone believed, not all arithmetical truths can function as data. In identifying a select few of these truths as data, a promising suggestion would be those arithmetical truths expressed by a Π^0_1 sentence. On the one hand, such sentences form a natural class of arithmetical sentences which can be considered suitably elementary, given that they are of the form of universal generalizations over the numbers. Secondly, Gödel's theorems imply that any recursively axiomatized consistent set theory will be Π^0_1-incomplete, guaranteeing that the data are of a suitably inexhaustible kind. Finally, all large cardinal axioms have Π^0_1 arithmetical consequences, meaning that the positing of increasingly strong axioms is guaranteed to have relevance to the data. Hence identifying the data with this class seems most likely to justify the kind of maximalism about the height of the hierarchy that Gödel and others hope to found in terms of the science–mathematics analogy.

Of course, Δ^0_0 and Σ^0_1 arithmetical sentences are just as elementary, and may well merit consideration as data. But for the purposes of large large cardinal axioms, such sentences won't matter much. The Δ^0_0 arithmetical sentences are all equivalent to Σ^0_1 arithmetical sentences (by prefixing redundant quantifiers), and **PA** is complete with respect to this latter class, so we know in advance that no consistent large cardinal axioms will settle any of these sentences that we could not have settled without their help.

An identification of the data along these lines is made by Koellner (2009a, p.98). He distinguishes the 'primary' data, which are previously verified Δ^0_1 sentences, and the 'secondary' data, which are the Π^0_1 universal generalisations of these. Koellner's motivation for this selection is that verified Δ^0_1 sentences are analogous to observation sentences in the sciences, and hence their Π^0_1 generalisations are analogous to observational generalisations in the sciences. He claims further that 'in mathematics the secondary data can be definitely refuted but never definitely verified' (2009a, p.98). This claim is made on the basis of the analogy with physics, yet it is deeply implausible in the case of mathematics. After all, any Δ^0_0 sentence is also Δ^0_1 (since it is logically equivalent to itself prefixed with redundant quantifiers), and we seem to able to verify all sorts of Π^0_1 sentences which are universal generalisations of Δ^0_0

cardinal axioms consistent with **ZFC**, there is a limit to the process of strengthening **ZFC** by successive large cardinal posits.

formulae (e.g. that every prime is odd or is identical to 2). Given this implausibility, it is perhaps tempting to think that the data are meant to be some restricted class of Π^0_1 sentences which generalise verified Δ^0_1 sentences. But Koellner's remarks tell against this. For example, he claims that '[t]wo theories are mutually interpretable if and only if they prove the same Π^0_1-sentences, that is, if and only if they agree on the secondary data' (2009a, p.98). This means that there is no room for theories to agree on the secondary data and disagree on the full class of Π^0_1 sentences.

A further problem with Koellner's suggestion, and indeed with the more general identification of the data with the Π^0_1 arithmetical data, is that we are no more persuaded of the truth of every true Π^0_1 arithmetical sentence in advance than we are of the truth of every true arithmetical sentence. Take, for example, the even perfect number conjecture (EPN). This states that all numbers which are perfect (i.e. are the sum of their proper positive divisors) are even. This conjecture is clearly Π^0_1, however it is well-documented that mathematicians are (at least collectively) ambivalent regarding the truth of EPN (Baker 2007, p.63). So, if the conjecture is true and a theory proves it, we should regard the proof as analogous to a surprising scientific discovery, and *not* as an account of any data.

Given the faults in Koellner's primary/secondary classification, the important question now is how to delimit in advance which true arithmetical sentences at most as complex as Π^0_1 are to be considered data. It's entirely possible that no sharp delimitation is possible; indeed the closer the analogy between mathematics and the sciences, the less we should expect a sharp delimitation to be possible. Nonetheless, we can give a preliminary taxonomy of the kinds of arithmetical proposition which should pass muster.

Arithmetical Data: Hard and Soft

Firstly, there are those sentences originally discussed by Russell, which gain their status as data for broadly Millian reasons, the original example being that two sheep and two sheep are always observed by shepherds to yield four sheep. Since '$2+2=4$' is Δ^0_0, it is also Π^0_1, and hence is of the right shape for our data class. Moreover, we believe it to be true pre-theoretically, and indeed with more certainty than the axioms themselves. Hence these *Russellian data* should be admissible for Gödel's

analogy too.

We previously saw in the exposition of conceptual platonism that Gödel takes a perception-like relation to hold between us and mathematical objects. I argued that this element of Gödel's thought is of little significance, since whatever perception-like relation holds, it holds in virtue of our grasp of the truth of axioms quantifying over the objects concerned. Nonetheless, such a perception-like relation is most plausibly construed along Kantian or Hilbertian lines as the singular representation of an object to a thinking subject. If this has any significant role to play in Gödel's philosophy, it is in providing *some* of the mathematical data (though for reasons discussed above, it cannot provide all the data).

Such intuitive data, like the basic Russellian data, will no doubt be restricted to propositions of a fairly simple sort. Since intuitive representation is supposed to be singular, even without a thorough account of how such intuition is supposed to work, we can tentatively say that arithmetical propositions which we can verify on the basis of this faculty should be equivalent to to a Π^0_1 or Σ^0_1 sentence, as required here. I won't dwell on this issue, since it's fairly clear that Gödel's scant remarks on singular intuition radically underdetermine the theory required to account for them. I mention the issue only because Kantian intuition is a plausible source of data going slightly beyond the Russellian. There might, for example, be simple additions such as $51,000,000,000,000 + 1 = 51,000,000,000,001$ which could plausibly be verified in intuition, but of which no plausible Millian account can be offered. It would be a wealthy shepherd indeed whose work necessitated repeated exposure to concrete instances of the addition above!

In the non-mathematical context, Russell distinguishes between *hard* and *soft* data (1914, lecture III). The distinction is broadly psychological in nature, and is not supposed to be exhaustive or exclusive. But as a heuristic it is still helpful for our purposes to consider the degree to which data can be classified as hard or soft. The paradigmatic hard data for Russell are the laws of logic, Russellian mathematical data in the above sense, and facts about one's own sense data. We can also, for the sake of thoroughness, include propositions verifiable in intuition here. The common characteristic is that Cartesian reflection on propositions of this kind do not induce doubt in us as regards their truth. Soft data are, by contrast, those which are open to at least philosophical doubt, such as the existence of material objects or other

minds. Though he does not put it in quite these terms, in the Gibbs Lecture Gödel suggests that soft data in mathematics are admissible for quasi-scientific purposes.

In particular, he argues that a platonist should feel comfortable with the verification of number-theoretic claims by enumerative induction (i.e. verification of universal number-theoretic claims by verification of instances up to large integer values). He writes:

> I admit that every mathematician has an inborn abhorrence to giving more than heuristic significance to such inductive arguments. I think, however, that this is due to the very prejudice that mathematical objects somehow have no real existence. If mathematics describes an objective world just like physics, there is no reason why inductive methods should not be applied in mathematics just the same as in physics (Gödel 1951, p.313).

In trying to reconstruct a Gödelian conception of mathematical data then, it is reasonable to suppose that Π^0_1 arithmetical sentences verified in a sufficiently large number of instances should be considered as soft data. Hence if, for example, Goldbach's conjecture were derivable from a large cardinal axiom, we should count that as regressive support for the axiom candidate.[13] This is because most mathematicians believe the conjecture to be true despite lack of a proof (Echeverria 1996, p.42). The obvious explanation as to why is that the conjecture has been verified in an enormous number of instances.[14]

Of course, there are a number of philosophical issues with soft data of this kind. For one thing, it isn't even known whether Goldbach's conjecture is independent of **PA** (and indeed, if it is false, its negation is provable in **PA**). If it *is* provable in **PA**, then it's possible that the proof is so complex that nobody could feasibly carry it out. If this is the case, then a simpler proof from a large cardinal axiom would provide justification for that axiom merely in terms enhancement of theoretical virtue, rather than a more compelling strictly regressive justification (we'll return to issues around the 'speed-up' of proofs in §5).

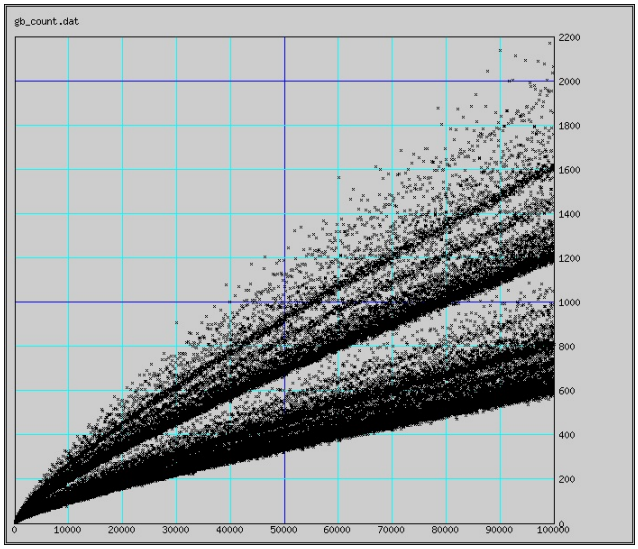
[13] In modern form, Goldbach's conjecture states that any even number greater than 2 is the sum of two primes.

[14] At least 2×10^{17} instances have been checked. Up to date information is available at Tomás Oliveira e Silva's website at 'http://sweet.ua.pt/tos/goldbach.html'. Accessed 05/04/2019.

Secondly, Gödel's statement that 'there is is no reason why inductive methods should not be applied in mathematics' is false. A very good such reason was offered by Frege, namely that induction in physics is lent plausibility by the fact that *ceteris paribus* any region of space and time can be supposed similar in the relevant physical respects. However the same is not true of the numbers, since the position they occupy in the number series makes a great deal of difference to their arithmetical properties, such as their divisors, primality, and so on (Frege 1884, pp.14–15).

I think that as far as reconstruction of Gödel goes, the results of sufficiently extensive enumerative induction should be admitted as soft mathematical data. Philosophically, however, I think this is mistaken, a view which appears to accord with mathematical practice. Although much work *has* gone into verifying large numbers of its instances, Baker (2007, pp.69–70) makes a compelling case that enumerative induction is not the source of widespread belief in Goldbach's conjecture.

Baker argues instead that the belief has its origins in Cantor's partition function. With a given even number ≥ 4 as its argument, the partition function takes as its value the number of ways it can be decomposed into the sum of two primes. Though this function does not increase monotonically, its graph, displayed below for even arguments from 4 to 100,000, is certainly suggestive.[15]



[15]The graph is taken from Mark Herkommer's Goldbach research site at 'http://www.herkommer.org/goldbach/goldbach.htm'. Accessed 07/09/2018. © Copyright, 1998-2014 Mark Herkommer. Permission to reproduce this graph has been kindly granted by the copyright holder.

Baker argues that the increase in mathematicians' confidence of the truth of Goldbach's conjecture coincides with investigation of the partition function, and hence that this confidence is not based on enumerative induction alone (which Baker takes to be illegitimate for the Fregean reason above). Rather, the confidence comes from the apparently increasing cone-like pattern exhibited by the graph of the partition function.

At this point, one might be tempted to think that enumerative induction is after all the source of the mathematicians' beliefs here, with a small change in perspective: the induction is that for many even arguments from 4 onward, the value of Cantor's function isn't 0, therefore Goldbach's conjecture is true. But this would be too quick: as Baker argues, there is more than enumerative induction going on here. Given the apparently-increasing pattern of the graph, the 'hard' cases for Goldbach's conjecture should be amongst very small numbers already tested manually. In other words, the sample cases observed are biased *against* Goldbach's conjecture, and if it were false, we should have found the counterexample amongst the previously studied instances. So mathematicians don't need to be seen as accepting the result of simple enumerative induction here, but rather as accepting the result of enumerative induction over a sample biased against the conjecture. Baker takes this to be a distinct kind of non-enumerative inductive evidence for the conjecture (2007, p.71). Therefore, even if we do not wish to countenance soft data of the kind envisaged by Gödel, we may be inclined to think that some Π^0_1 arithmetical sentences should be admitted into our class of data on the basis of such non-deductive plausibility considerations.

Lastly, there is a kind of data with which a number of previous chapters have been preoccupied, namely the Π^0_1 arithmetical sentences constructed in the proof of Gödel's theorems, like Gödel sentences, canonical consistency sentences, and Diophantine sentences. When sentences of these kinds are constructed effectively from an axiomatic system which we recognize to be sound, it follows immediately that they are true, *and* that they are not 'accounted for' by the axiom system in question. However, given the previous argument that certain such sentences are absolutely undecidable, the data will not include all true consistency sentences, all true Diophantine sentences, etc.

Moreover, such sentences constructed from the axioms of a system which we do

not believe in advance to be sound will not pass muster either. Since the justification of such propositions is parasitic on the axiomatic system from which they are obtained, data of this kind will be harder the higher our degree of confidence in the soundness of the relevant axiom system. A proposition such as $Con_{\mathbf{PA}}$ should be regarded as data of the hardest kind, with $Con_{\mathbf{ZFC}}$ as perhaps somewhat softer. Something like $Con_{\mathbf{ZF}+\exists x\ x \text{ is Reinhardt}}$ should not be considered data at all.[16]

In summary, 'ordinary' mathematics, mathematical perception, and singular intuition cannot supply a collection of data by which the analogy between mathematics and natural science could justify the positing of large large cardinal axioms. The elementary part of mathematics which most plausibly can behave as data is number theory. A restriction on which number-theoretic sentences can be considered data is nonetheless required. Although we cannot determine which statements precisely are data, a natural class consists of Russellian data, together with the Π^0_1 sentences generated by Gödelian incompleteness that we have reason to believe are true. Gödel's writings suggest that he thought the results of certain enumerative inductions should be considered as well; although I've expressed scepticism on the matter, it is plausible that some Π^0_1 arithmetical sentences should be considered data even in the absence of (formal or informal) proof, namely where there are strong heuristic reasons to suspect they are true. In the next sections, we'll examine the sense in which large large cardinals might be thought to account for such data.

5.4 The Laws of Nature Analogy

Since large cardinals themselves are not analogous to material bodies as Gödel initially suggested, the alternative is that the axioms which assert the existence of such sets are analogous to 'laws of nature' or other theoretical posits. One respect in which an axiom playing the role of a law can be successful is *strictly regressive*, if it allows for the deduction of data which could not be obtained by weaker principles.

[16] Although the existence of Reinhardt cardinals is known to be inconsistent with **ZFC**, it is unknown whether they are consistent with **ZF**. A recent attempt by Rupert McCallum to prove Kunen's inconsistency theorem without the axiom of choice (which would settle the question negatively) almost succeeded, though not quite. The saga is documented by Joel Hamkins on his website at 'http://jdh.hamkins.org/tag/rupert-mccallum/'. Accessed 05/04/2019.

The other possibility is that large cardinal axioms function as laws of nature the positing of which enhances the theoretical virtues of set theory. This section will be concerned with the strictly regressive justification of large large cardinal axioms, by analogy to laws of nature. Our key question is whether large cardinal axioms can account for any data that cannot be obtained without them, according to the delimitation of the data given in the last section. If so, that would give strong regressive support to the large cardinals project, since in the sciences we certainly do accept natural laws which are posited for such reasons. However, I'll argue that Gödel's analogy cannot be sustained in this case.

Whether the adoption of large large cardinal axioms can account for data not accountable for without them (or alternative axioms of similar strength) is of course tremendously sensitive to what we take the data to consist of. The outline of the data just given is neither sharply delimited, nor precisely defined. But critically, the sentences expressing such propositions are all provable in **PA**, or else are of a restricted kind of Π^0_1 arithmetical sentences independent of **PA**. So we can say something relatively precise about the sentences expressing the data (though admittedly not as precise as in Koellner's account), and hence can say something quite definite about the role large cardinals might play in accounting for them. As one would expect in advance, the data are (at least in one respect) not very complicated sentences, and their simple form might give us good reason to suppose that large cardinal axioms do allow us to account for data which cannot be accounted for in their absence. This is because the addition of any large cardinal axiom to **ZFC** will reduce the degree to which it is Π^0_1-incomplete, as can be seen from the arrangement of large cardinal axioms in a hierarchy of consistency strength.[17]

Indeed, we may even think that the inclusion of sentences constructed by Gödelian methods in our selection of data *guarantees* that positing a large cardinal axiom will relevantly account for some of the data, in the following way: suppose you are persuaded that intuitive considerations justify the belief that the theory $\mathbf{ZFC} + \exists x\ x = \kappa(\omega)$ is sound, as discussed in the previous chapter. You'll then certainly believe that the theory is consistent, but by Gödel's theorem $\mathbf{ZFC} + \exists x\ x = \kappa(\omega) \nvdash Con_{\mathbf{ZFC}+\exists x\ x=\kappa(\omega)}$. So $Con_{\mathbf{ZFC}+\exists x\ x=\kappa(\omega)}$ is a relevant piece

[17]As in fn.1 (above), it simply *appears* that the large cardinal axioms are so arranged. There is no theorem to this effect.

of data, namely a Π^0_1 arithmetical sentence that we take to be true and subject to Gödelian incompleteness. The adoption of a large cardinal axiom stronger than one asserting the existence of $\kappa(\omega)$ will allow you to prove the consistency sentence, and hence account for more data than the unsupplemented theory.

Since our newly supplemented theory accounts for more of the data, the analogy between science and set theory justifies (albeit not with certainty) a belief that it is sound, and hence that it is consistent. And the whole process starts again, justifying a set theory of ever increasing strength by extension with stronger and stronger large cardinal axioms.[18] The fact that a large cardinal axiom 'accounts for the data' gives us only probable reason to believe that it's true, so we have the expected gradual loss of certainty as we move up the hierarchy of large cardinal axioms as well.

In my view, this is the most persuasive quasi-scientific argument for large cardinal axioms in which those axioms are afforded strictly regressive support (as opposed to being merely virtue-enhancing), and as a point of interpretation it fits well with Gödel's general remarks on the issue. Firstly, the initial step of the argument requires the use of intuition to found the truth of the axioms of some strong set theory.[19] Secondly, the incompleteness theorems play a crucial role in the argument, since they are required to establish the need for a *series* of extensions via large cardinals, as outlined at (1964, pp.260–261). Thirdly, there is a clear sense in which the large cardinals 'account' for the data, since they do so directly via increasing the deductive strength of the base theory. Finally, the picture accords well with Gödel's twin claims that such axioms need no intuitive justification, and thereby introduce axioms and theorems into mathematics the truth of which can only be maintained as probable.

Compelling though it may be, this argument suffers from a serious philosophical flaw, as well as creating a secondary issue for Gödel in particular. We saw in chapter 2 that Gödel's rationalistic optimism required him to accept that, with respect to a recursive reflection progression of **PA**, we could select a path Δ within $\mathcal{O}$ such that

[18]The argument here is inspired by Gödel's remarks (1964, p.269), though the idea there is actually about intuitive justification, and does not mention large cardinals in particular.

[19]Even in papers like (1964), where the quasi-scientific programme is well underway, Gödel maintains that intuition has an important role in founding the general platonist interpretation of the axioms, suggesting that he does not think quasi-scientific justification is alone sufficient for developing such a picture.

$\bigcup_{n\in\Delta:|n|<\omega^{\omega^2+1}} \mathbf{T}_n$ is arithmetically complete. I argued that the possession of such an ability by even idealised mathematicians is a deeply implausible hypothesis, and I won't rehearse the argument here. The main point, however, is that for Gödel, large large cardinal axioms shouldn't have any strictly regressive justification. This is because the data are arithmetical sentences, and so according to the rationalistic optimist, can be accounted for by reflection on arithmetic. Even if it is hard to imagine an actual mathematician doing so, rationalistic optimism entails that, if it is consistent to add a measurable cardinal, for instance, to **ZFC**, this can be determined merely by reflecting on the soundness of **PA** in the right way.

Some set theory is required for this, of course, since constructing the required reflection sequence relies crucially on the representation of recursive ordinals via a subset of the natural numbers. But for the optimist, regressive justification shouldn't extend beyond the segment of set theory required to validate the existence of the recursive ordinals, which constitute merely a proper initial segment of the countable ordinals. Unless an argument could be provided that large large cardinal axioms were *essential* to recognising suitable notations, nothing even approaching such an axiom is required for the relevant construction. We have some tension, therefore, between two key Gödelian thoughts: if rationalistic optimism is correct, it is unclear how large cardinal axioms could have any strictly regressive justification (though justification by virtue-enhancement is still a live possibility for such axioms).

That problem, however, only applies to the rationalistic optimist. For those who share my scepticism, it may seem as if the selection of data made affords powerful regressive support for the large cardinals programme. According to the picture sketched above, it looks as if we might come close to a hierarchy of regressively justified large cardinal axioms constrained only by consistency. A potential problem is that the progressive decrease in the certainty of our axioms could perhaps lead to a decrease in regressive support such that we stop being justified in positing new cardinals rather early in this process. But that is merely a possibility. In reality, there are more substantial issues in the vicinity.

A first point to note is that, as we've seen, much of the mathematical data will be consistency sentences, or sentences which are equivalent to consistency sentences. An immediate problem that raises doubts about the need for large cardinal hypotheses with respect to such data is related to ordinal analysis. If the consistency

sentences we can take to be data are those of sound recursive theories, then the consistency sentence should be provable via Gentzen's method of transfinite induction up to the theory's proof-theoretic ordinal. Since the proof-theoretic ordinal of a theory $\mathbf{T}$ is the supremum of ordinals for which there is a notation in $\mathcal{O}$ which $\mathbf{T}$ verifies is a notation, it follows that any proof-theoretic ordinal is $< \omega_1^{CK}$. The initial worry then, is that for the purposes of verifying elementary data, large cardinals are *excessive*; much of the work could be done using much more conservative resources in the large countable ordinals. That said, identifying the proof-theoretic ordinal of a theory is often far from straightforward. And moreover, we admitted that other Π_1^0 arithmetical sentences besides consistency sentences might pass muster as data, so the picture outlined above remains intact. There is, however, a much more severe problem with the proposal, to the effect that large cardinal axioms cannot be *required* to account for the data as construed.

In particular, the problem is that when a large large cardinal axiom 'accounts' for some otherwise unaccounted for piece of data, that gives us no reason to believe that the axiom is *true*. The key reasons are that **PA** is sound, and is complete with respect to Σ_1^0 arithmetical sentences, although this requires a little explanation. Suppose that δ is some large cardinal axiom consistent with **ZFC**, and that ϕ is a Π_1^0 datum such that $\mathbf{ZFC} \nvdash \phi$ and $\mathbf{ZFC} + \delta \vdash \phi$. Suppose ϕ is false; in that case $\neg\phi$ is equivalent to a true Σ_1^0 arithmetical sentence. Since **PA** proves all true Σ_1^0 arithmetical sentences, $\mathbf{PA} \vdash \neg\phi$. However, since **ZFC** extends **PA**, $\mathbf{ZFC} \vdash \neg\phi$. This contradicts our assumption that δ is consistent relative to **ZFC**, since $\mathbf{ZFC} + \delta \vdash (\phi \wedge \neg\phi)$. So (assuming **ZFC** is consistent), ϕ is true. Hence, we have accounted for a new piece of data, namely ϕ, by proving that it is true. Crucially, however, *at no point* was the truth of δ required. All that was used in the argument was the assumption that δ was consistent with **ZFC**.

To see this, suppose that γ is some axiom candidate consistent with **ZFC** such that $\mathbf{ZFC}+\gamma+\delta \vdash 0 = 1$. Suppose further that ψ is a Π_1^0 datum such that $\mathbf{ZFC} \nvdash \psi$ and $\mathbf{ZFC} + \gamma \vdash \psi$. The same argument as before suffices to show that ψ is true: if it is false, $\neg\psi$ is equivalent to a true Σ_1^0 arithmetical sentence. Hence, $\mathbf{PA} \vdash \neg\psi$, so $\mathbf{ZFC} \vdash \neg\psi$. This contradicts our assumptions, hence ψ is indeed true. Now γ and δ, by construction, are not *both* true. Yet the data for which these axioms were supposed to account are both true regardless. So the deduction of data which

are not derivable in **ZFC** by large cardinals axioms provides no regressive support for the truth of these axioms; rather it at best supports the assumption of their consistency relative to **ZFC**.[20]

It seems that this observation should be of some interest to both Gödelians and formalists. If, as I have argued we must, we restrict mathematical data to a special class of Π^0_1 arithmetical sentences, then as far as these data are concerned, we seem to be in a Hilbertian scenario with respect to large cardinal axioms. In other words, as far as large cardinal axioms go, their consistency is as good as their truth. This is not, of course, true in general as regards their distinctively *set-theoretic* consequences.[21] But such set-theoretic consequences are liable to be interpreted as merely ideal by a Hilbertian formalist anyhow.

The interest here is limited, however, by the fact that for a truly Hilbertian formalist, basic **ZFC** would likely be regarded as itself merely ideal, so it's not as if this argument will persuade them that this theory extended by a relative consistency statement is the correct formulation of the foundations of mathematics. Nevertheless, the formalist could take some comfort from the fact that the Gödelian analogy cannot establish the truth of wildly infinitary large cardinal axioms.

Moreover, I think that for the platonist with a substantial notion of mathematical truth going beyond mere consistency, the argument presented should be troubling. If we want to put the large cardinals programme on solid philosophical ground, it won't do to think that the open series of extensions of **ZFC** that we ought to believe consists simply of **ZFC** extended by propositions asserting the consistency of **ZFC** with certain large cardinal statements. But that picture is all that is regressively supported by the data, so the idea that the methods of theory choice in science can be applied in set theory is placed under considerable strain.

To put it the other way, recall that the analogy between mathematics and natural science was founded on the view that the subject matter of mathematics was analogous to the subject matter of natural science, such that some version of the methods of the latter was thereby admissible in the former. Given that the theories

[20]Technical details relevant to this point are explained further in (Potter 2004, pp.217–218).

[21]There are, however, scattered examples to be found, e.g. Solovay's theorem that **ZFC** doesn't refute the proposition that all uncountable Π^1_2 sets have a perfect subset. The proof relies on the assumption that an axiom asserting the existence of an inaccessible is consistent; the existence of such a cardinal is irrelevant (Maddy 1990, p.127 fn.60).

of natural science are certainly not confirmed by mere consistency, the platonist's case for large large cardinals is substantially undermined. The science–mathematics analogy prevents the Gödelian from taking up a Hilbertian conception of axioms, according to which consistency and truth coincide. The argument that I've presented shows that elementary mathematical data will *at best* support the view that large cardinal axioms are consistent, and not that they are true. So either the laws of nature analogy must be abandoned, or else the platonist must admit that it fails to justify a belief in large large cardinals.

None of this is to say that the Gödelian *cannot* justify the truth of large large cardinal axioms by other quasi-scientific methods; after all they can still argue that the truth of such an axiom can enhance the theoretical virtues of set theory to a greater extent than can the corresponding consistency sentence. But this is a much weaker kind of support far more open to doubt. Indeed, the situation for quasi-scientific justification keeps getting worse: the strongest form of regressive support that has been offered for large cardinal axioms was that the existence of large cardinals was only as open to doubt as the existence of medium-sized dry goods. But as we saw, that analogy could not be sustained. Now we have seen that large cardinal axioms do not even contribute to the adequacy of set theory with respect to its data, since the statement that they are consistent relative to **ZFC** will do that just as well. Making a case that the truth of large cardinal axioms is substantially more virtue-enhancing than their mere consistency relative to **ZFC** is the only option left for the platonist who takes Gödel's analogy seriously.

On the other hand, the argument above is unlikely to trouble platonists who don't share Gödel's view that theory confirmation within mathematics is analogous to theory confirmation in physics. You could of course be both a platonist *and* a maximalist of a less naturalistic persuasion; for instance, if you thought that the concept *set* mandated the adoption of any consistent maximising principle, large cardinal axioms included, the argument above would be of little consequence.[22] It's

[22]Certain remarks of Gödel's do at times suggest that he is tempted by such a position. For example, footnote 23 in the 1964 version of the continuum paper cautiously suggests that the concept *set* dictates a maximality principle inconsistent with $V = L$. If this is correct, then mathematical intuition would verify a principle considerably stronger than the reflection principles previously examined; this is because the existence of $\kappa(\omega)$ is consistent with $V = L$ (Jech 2003, p.304). The corresponding footnote in the 1947 version is number 22, which contains no such

clear however, that such a position can make no room for a substantial analogy between mathematics and natural science, since in the natural sciences there is no sense in which the consistency of a theory amounts to its truth. In summary, the platonist who takes the analogy between mathematics and natural science seriously cannot maintain that large cardinal axioms function analogously to laws of nature which are necessary to account for the data.

It is worth noting that the argument offered here is of some relevance beyond the narrow confines of the large cardinals debate. The only plausible candidates for mathematical data are arithmetical sentences of at most Π^0_1 complexity that we have prior reason to believe are true. If I'm correct about this, then the completeness of **PA** with respect to Σ^0_1 arithmetical sentences places severe constraints on the regressive justification of any axioms which are stronger with respect to arithmetical sentences than the axioms of **PA** themselves. Therefore any regressive epistemology in the vicinity should have at most modest aspirations. The problem then, is that weaker axioms are more likely to be persuasive candidates for self-evidence, and therefore the significance of the regressive project as a whole is put into question by the arguments of this section.

This point certainly has some relevance for Koellner's account of axiom selection in set theory (2009a, p.98). As mentioned above, he is an advocate of a Π^0_1 account of the data. Invoking a theorem due to Guaspari and Lindström, Koellner highlights that extensions of **ZFC** by new axiom candidates will prove exactly the same Π^0_1 data iff they are mutually interpretable.[23] For Koellner, the correct method of selecting axiom candidates in set theory is thus to partition the extensions of **ZFC** into equivalence classes under sameness of Π^0_1-consequences and to select, from each class where these consequences are independently verified, the axiom candidate which possesses the greatest theoretical virtues. What the argument of the present section shows is that none of the axioms so considered are actually required to account for this data at all.

I make no claim that Koellner's project as a whole collapses under the weight of this observation. A central part of his project is to argue that certain axiom

suggestion, possibly indicating a shift in Gödel's view over the intervening years.

[23]See Button and Walsh 2018, pp.113-114 for the definition of mutual interpretability, and p.124 for full details of how the Gauspari-Lindstöm theorem fits into Koellner's programme.

candidates are much more strongly supported by 'theoretical reason' than others. Theoretical reason is left largely undefined by Koellner, but he explicitly contrasts considerations of theoretical reason with those of expediency or convenience, and it is clear that theoretical reason in mathematics is supposed to play a similar role to the standard principles of theory choice in the sciences. In the case of one particular axiom candidate,[24] he provides a detailed account of eight theoretical virtues possessed by the axiom candidate over its mutually interpretatable rivals (2009a, pp.101-102). According to him, these virtues make it clear that the axiom is highly favoured by theoretical reason, at the expense of rival principles with incompatible set-theoretic consequences.

None of what I have said contradicts any of Koellner's claims about the virtues of specific axiom candidates. However, my argument above serves to show just how much work is being done in his account of mathematical theory choice by the theoretical virtues. If indeed we should adopt any large large cardinal axioms, it will not be because they must be adopted to account for any of the data, since that data can be equally accounted for by much weaker consistency sentences. This is significant because, *prima facie*, the ability of a theory to account for the relevant data in the sciences offers a strong reason for accepting it. But no such justification is available to large large cardinal axioms if we have a conception of the data similar to Koellner's. This not only weakens the case for adopting such axiom candidates, but also places a good deal of strain on the overall analogy that he wishes to draw between mathematics and the natural sciences.

5.5 Theoretical Virtues

Things are not looking promising for the analogy between mathematics and natural science as a means of justifying the large cardinals programme in full generality. We've seen that, although a plausible delineation of the data is possible, there is no sense in which accounting for this data can give large large cardinal hypotheses strong regressive support. On the one hand, such cardinals themselves are not

[24]The axiom in question is the determinacy principle $\mathrm{AD}^{L(\mathbb{R})}$. Determinacy principles are discussed in greater detail below.

required for the prevention of a philosophical miracle, as material bodies plausibly are. This is because no particular cardinal plays the right explanatory role; all that is required is the truth of some existential generalisation of a certain consistency strength or greater. On the other hand, large large cardinal *axioms* do not receive strictly regressive support by accounting for the data in the relevant way, since demonstrably only their consistency is required for this.

There is another respect in which large large cardinal axioms can be quasi-scientifically successful, namely by enhancing the *theoretical virtues* of set theories to which they are added. This aspect of the analogy is by far the most often discussed in the literature, and appears to be the central justification for large large cardinal axioms, as far as several mathematicians and philosophers are concerned. The classic exposition of the view, unsurprisingly, comes from Gödel:

> Success here means fruitfulness in consequences, in particular in "verifiable" consequences, i.e., consequences demonstrable without the new axiom, whose proofs with the help of the new axiom, however, are considerably simpler and easier to discover, and make it possible to contract into one proof many different proofs... A much higher degree of verification than that, however, is conceivable. There might exist axioms so abundant in their verifiable consequences, shedding so much light on a whole field, and yielding such powerful methods for solving problems (and even solving them constructively, as far as that is possible) that, no matter whether or not they are intrinsically necessary, they would have to be accepted at least in the same sense as any well-established physical theory (1964, p.261).

Although Gödel thought that the verification of large cardinal axioms by such means could only ever be probable, and that at the time of writing no proposition had been so verified, the core idea of this passage has been remarkably influential in the philosophy of mathematics. Quine (1990, pp.94–95), Maddy (1997, p.233), Koellner (2010, p.190) and others have all adopted the idea that a decision on the truth of at least some axiom candidates can be reached on the basis of analysing the extent to which these axioms enhance the *theoretical virtues* of **ZFC** when they are added to it.

As is the case with respect to the natural sciences, it isn't clear exactly what properties are to count as theoretical virtues, and there are difficult questions in the vicinity about how such virtues are to be weighted, and how virtues collectively should fare against other criteria for theory choice. Nonetheless, there are canonical examples of theoretical virtues in mathematics that should prove sufficient for our discussion. In the passage above, Gödel focuses on the contraction of existing proofs, the proof of new theorems, and the solution of open problems. Other virtues discussed include the 'naturalness' of an axiom candidate (Gödel 1938, p.27), the naturalness or expectedness of its deductive consequences (Moschovakis 1980, p.610), and the 'effective completeness' of the supplemented theory (Koellner 2010, p.204).[25] Other mathematical theoretical virtues which have been proposed include the speed-up of proofs, maximisation of interpretative power, and many more besides. Far too many virtues have been proposed in the literature to canvass here, however the most important examples will be explored in some detail. The ability to solve open problems is the most prominent theoretical virtue discussed in connection with mathematics, and likewise parsimony takes center-stage in discussions of scientific theories. I'll argue that with respect to these theoretical virtues (and also the speed-up of proofs), large large cardinal axioms should not expect a very positive evaluation.

Open Problems

The solution of open problems is a virtue of strong set theories that has received an enormous amount of attention, from Gödel onwards. The initial axiom candidate lauded with this virtue was $V = L$. Although the virtues of several large cardinal axioms inconsistent with this principle are now well-regarded in the literature, $V = L$ does indeed have the virtue of solving many open problems, both within set theory and without.[26] Within set theory, it solves GCH affirmatively, as proved by

[25]This is perhaps a theoretical virtue that has no strong analogue in the natural sciences. A theory is said to be 'effectively complete' with respect to a given class of statements if it decides every statement in the class *except* the undecidable statements generated by Gödelian incompleteness.

[26]Large large cardinal axioms become inconsistent with $V = L$ quite rapidly. If the Erdős cardinal $\kappa(\omega_1)$ exists (a rather small large large cardinal), then so does $0^\#$, the set coding true statements about indiscernibles in L (Kanamori 2009, p.107). A theorem of Kunen shows that if $0^\#$ exists, then there is a non-trivial elementary embedding $j : L \to L$ (Kanamori 2009, p.XX).

Gödel. Perhaps slightly less well-known is that $V = L$ implies that every Whitehead group is free, solving a famous open conjecture in algebra (Shelah 1974).

Nowadays, the focus is on the ability of large cardinal hypotheses to solve open problems in descriptive set theory. A proliferation of results exist using large cardinals to prove that sets of reals have various separability and measurability properties, and that particular games on sets of reals are determinate. The most famous such result is probably Martin and Steel's proof (Martin and Steel 1989) of the projective determinacy axiom,[27] which follows from the existence of infinitely many Woodin cardinals. There are many other well-known examples of open problems solved by large cardinal hypotheses, and there is no need for my purposes to report them all. It is important to note that even where large cardinals have consequences for more concrete areas of mathematics, the decision of open problems can only be taken to enhance the theoretical virtue of set theory including such an axiom. We cannot take the extension of set theory by such an axiom as having strict regressive support on this basis, since the solved problems are viewed as being genuinely open in advance of positing the large cardinal axioms which facilitate their solution. As remarked above, the solution to open problems should be viewed as analogous to the making of a novel scientific discovery: the discovery is to be trusted only if the theory from which it follows is already believed to be sound.

This is not, of course, to say that the ability of an axiom candidate to provide solutions to open problems should not be considered highly virtuous. But there are a number of considerations which should make us regard this kind of justification with some caution. In the first instance, the strength of this kind of support is sensitive to whether the problem solved is one about which mathematicians have a strong prior view. For example, it may be that a large cardinal axiom which proved Goldbach's conjecture would be very virtuous indeed due to the widespread belief in the truth of that conjecture. But the solutions to open problems that we see in reality are by no means so persuasive.

For example, Maddy favourably cites the result that if there is a measurable

Another theorem of Kunen shows that there is no such embedding $j : V \to V$ in models of **ZFC** (Kanamori 2009, pp.318–319).

[27]This axiom, PD, states that in every two-player game of length ω with perfect information on a projective set of reals, one of the players has a winning strategy (i.e. the game is determined). See (Moschovakis 1980, ch.6) for the relevant details.

cardinal, then there is no projective well-ordering of the reals (Maddy 1990, p.138). But it is not as if the mathematical community at large suspected in advance there should be no such well-order. She cites Martin as expressing the view that this result, and others like it, are 'pleasing'. But hypotheses in science are not, to my knowledge, accepted on a regular basis for having pleasing consequences. Indeed what we find pleasing is highly contingent of the history on the discipline, not to mention personal taste: perhaps to some, the implication from $V = L$ that there is a relatively simple Δ^1_2 well-ordering of the reals would be pleasing. Indeed $V = L$ was described by Gödel as being a very natural principle at the time of his relative consistency proofs. Of course many open problems are such that the mathematical community is overall undecided with respect to their solution. The continuum hypothesis is a good example of such a problem. The collective ambivalence of mathematicians as regards it partly explains how it is possible that the most popular axiom candidates leave it open, where it is settled positively by the unpopular $V = L$.

Secondly, it is clear that the strength of support lent to an axiom by the solution of an open problem is related to the urgency within the mathematical community of solving the problem in question. And this matter is clearly relative to the interests of the community under consideration. As Potter highlights, (2004, p.221), the open problems solved by large cardinals are typically set-theoretic in nature, and not part of 'ordinary mathematics'. Examples such as $V = L$ solving the Whitehead conjecture are not easy to come by; in most cases, large cardinal axioms are typically used to solve problems and conjectures raised by set theorists themselves, rather than by practitioners in more mainstream areas of mathematics. If a large cardinal axiom could be used to solve a live conjecture posed by a number theorist, that should count as a greater theoretical virtue than the ability to solve a set-theoretical problem. As of yet, no example of such a conjecture has been found. The closest example of a genuinely mathematical open problem solved using large cardinals is that of Borel determinacy, proved by Martin (1970), under the assumption of a measurable cardinal. However, Martin subsequently proved the Borel determinacy axiom in unaugmented **ZFC** (1975), so the example should not inspire us with confidence that large cardinal axioms are useful in the solution of problems outside of set theory. So, for now at least, we should not in general place too great an emphasis

on the solution of open problems as a theoretical virtue of large cardinal axioms.

Moreover, there is a much deeper problem for a platonist with the idea that the solution of open problems can confirm a large cardinal hypothesis. The reason is that, as mentioned above, this kind of support is interest-relative, and hence the solution of open problems provides us only with interest-relative justification for those axioms. But for a platonist, such interest-relative justification cannot be considered justification proper, since the hierarchy is surely indifferent to the questions that interest us mathematically.

For example, it is quite possible that there be some mathematical community, exactly similar to ours except with respect to their interests, who place overwhelming value on the kinds of determinacy problems that appear in descriptive set theory. Suppose such people regard the solution of determinacy problems as the proper goal of all mathematical enquiry. To such a community, the full Axiom of Determinacy, AD, would have overwhelming theoretical virtue in respect of solving open problems.[28]

Actual mathematicians don't typically consider AD to be viable, since it contradicts the axiom of choice (Kanamori 2009, p.368). And if faced with such a community, we could certainly try to dissuade them by highlighting the merits of the axiom of choice, both intuitive and quasi-scientific. Indeed, for the platonist, this would be the only honest course of action. A pluralist might think that the imagined community have a perfectly good justification for studying set theory with AD, and that the actual community of today is quite right to ignore it. But for a Gödelian platonist, AD is simply a blatant falsehood. It is not *merely* that for the imagined community, AD has many virtues which (according to a choice-favouring platonist) ought to be out-weighed by other considerations. It is rather that the imagined community has *misleading* interests, in that the pursuit of such interests is counter-productive to uncovering the truth about sets. So it is hard to see how a platonist could make sense of the idea that mathematicians in another community have *any* reason to believe that AD is true, *merely* in light of their bizarre mathematical interests. After all, their position is assumed to be epistemically similar to ours in all ways other than with respect to their interests.

[28]This axiom is a generalisation of PD, and states that that every two-player game of length ω with perfect information on *any* set of reals is determined.

Moreover, when we get past the axiom candidates which possess intuitive support, and consider large large cardinal axioms with *only* quasi-scientific justification, it is hard to know how we could verify whether or not our own interests are misleading in this way. So I think that the platonist in particular should not take the solution of open problems too seriously when it comes to justifying axiom candidates, especially given the narrowly focused achievements of large large cardinal axioms in this regard to date.

Speed-Up Results

One theoretical virtue, mentioned by Gödel above, is the ability of more powerful systems to speed up the proof of theorems in weaker theories. While this virtue is discussed much less often that either the solution of open problems (above) or parsimony (below), I have chosen to include a discussion of it because the ability of an axiom candidate to speed-up proofs can do much more than simply make a theory more virtuous. In the right circumstances, the effect that speed-up has on a system can close a genuine and pressing explanatory gap.

A classic presentation of the issues at stake can be found in (Boolos 1987). In that paper, Boolos presents an argument which is essentially a Sorites-paradox-style inference appended with a definition of a very fast-growing function. The number of steps of the shortest proof of this result in first-order logic is given by an exponential stack of 64 '2's, far greater than the number of particles in the universe. Yet the proof is not difficult in second-order logic; indeed Boolos provides this in a short appendix to the paper. Moreover, the reasoning is obviously valid, as can be seen from its appropriate arithmetical interpretation. The moral of the story is that, since we *should* be able to prove the conclusion of the argument, given that it obviously follows from the premises, the fact that we *can't* (in the relevant sense) give a first-order proof of it is evidence that second-order logic is logic. After all, second-order resources are required for a feasible proof, which we seem perfectly able to provide.

We might hope to find similar support for a large large cardinal axiom. For this, we would need to find a genuine mathematical example of an agreed-upon theorem, such that the formal proof is unfeasibly long without the axiom, but is completely

feasible when the axiom is used. This would offer some powerful support for thinking that the axiom was true. If the informal justification for the theorem is obviously valid, this demands an explanation. In particular, if the formal proof of the theorem in our unsupplemented set theory is so long that it would take more than a human lifetime to complete, then the ability to follow the reasoning of that proof cannot explain our recognition of the theorem's validity. If a simple, feasible proof relies critically on a large cardinal assumption, then the large cardinal assumption gains a good degree of support from the fact that its truth is required to explain why a piece of reasoning which we all recognize to be valid and appear to be able to follow has these properties.

There are, however, several reasons to think that such an example will be extremely difficult to come across for our current purposes. In the first instance, given that we have an intuitive basis for believing in small large cardinals (according to the conceptual platonist, at least), the example would have to be one where a natural mathematical theorem from outside set theory was plainly valid, had an unfeasibly long proof without assuming the existence of a large large cardinal, and had a feasible proof *with* the assumption of such a large large cardinal. I certainly know of no example of a theorem meeting such specific constraints.[29]

Another limitation on the use of speed-up results to verify the existence of large large cardinals is that, according to the Gödelian conception of set theory being considered, we have *already* benefited from a huge amount of speed-up by using a second-order theory. Combining results from Gödel (1936) and Buss (1994), we get the following theorem:

> **Speed-Up Theorem:** For any function f, there are infinitely many formulae such that for any one of them, ϕ, $\mathbf{PA} \vdash \phi$ and $\mathbf{PA}_2 \vdash \phi$ and

[29]Tim Button has pointed out to me that in other circumstances, speed-up provided by a large large cardinal axiom could perhaps enhance the virtue of a theory even where the sped-up proof was not of a theorem the truth of which we were convinced of in advance. Namely, if we could show that the large large cardinal axiom was conservative over **ZFC** with respect to some class of statements, a proof of feasible length of some theorem T belonging to this class using the large cardinal axiom should convince us that T is true. If no feasible proof in **ZFC** of T can be found, then the speed-up of the proof of T could then be counted in the axiom's favour, as normal. This strikes me as correct, though I am pessimistic about the prospects of finding a concrete example of the phenomenon; especially since large large cardinal axioms tend to be radically non-conservative even over very simple classes of statements.

where n is the number of lines of the **PA** proof, and m is the number of lines of the $\mathbf{PA}_2$ proof, $n > f(m)$.

Hence the move to a second-order theory has already vastly increased our proof speed, at least with respect to arithmetical sentences. So, more specifically, verifying a large cardinal axiom θ via speed-up would require a formula ϕ which mathematicians regard as being informally valid, has an unfeasibly long proof in $\mathbf{ZFC}_2$ supplemented by any intuitively verifiable large cardinal axioms, and has a feasibly long proof in $\mathbf{ZFC}_2 + \theta$. This is a tall order indeed, but it is not impossible that such an example could be found. To my mind, finding such an example would offer the strongest quasi-scientific justification available for a large large cardinal axiom. Sadly, most of the available research on speed-up results relates to the order of the logical apparatus of a theory, rather than the large cardinal axioms it includes, so it is difficult to say anything conclusive on the subject of such axioms specifically.[30] For now, at least, we have good reason to believe that large large cardinal axioms are not substantially supported by the speed-up of proofs they provide, and that such support could be earned only in very exacting circumstances.

Parsimony

Another central theoretical virtue, more often discussed in connection with the sciences than with mathematics, is parsimony, or simplicity, of both ontology and ideology.[31] As I noted above, there is a difficult question about how to weight the virtues against each other, but I'll argue for a form of *pessimism* about the quasi-scientific justification of large large cardinal axioms, on the basis of parsimony considerations. To be clear, I'm not going to argue against the existence of such cardinals *tout court*; rather I'm going to argue that we have prior reason to believe

[30] In the more general area, Potter (2004, p.235) gives a very nice example of speed-up at work in enhancing the virtues of set-theoretic axioms: for large values of m, that the Goodstein sequence with m as its starting value terminates is provable in **PA**, but the proof is unfeasibly long. Since every Goodstein sequence terminates, it is obvious that the sequence which starts with m terminates. Replacement-free first-order set theory gives a feasible proof of Goodstein's theorem, and a proof of it for m by universal quantifier elimination. In this case, speed-up certainly supports the belief that certain first-order set theoretic axioms are true.

[31] Although it is not often discussed in connection with large large cardinals, the idea has been in circulation since at least (Quine 1951a, p.45).

that axioms positing them will score poorly on the front of theoretical virtues, as construed by the analogy with the natural sciences.[32]

The first ingredient in my argument is merely the observation that considerations of both ontological and ideological parsimony play an important role in the justification of theories in natural science. The two principles under consideration are:

1. **Ockham's Razor:** Entities are not to be multiplied beyond necessity.

2. **Kant's Razor:** Principles are not to be multiplied beyond necessity.

Both of these principles are no doubt familiar, and widely deployed within philosophy and elsewhere.[33] Similar principles have been endorsed by philosophers at least since Aristotle, but much more significantly they have been strongly endorsed within the natural sciences themselves. Galileo, in his critique of the Ptolemaic system, deployed the principle that 'Nature does not multiply things unnecessarily; that she makes use of the easiest and simplest means for producing her effects; that she does nothing in vain, and the like' (Galileo 1632, p.397). A similar principle appears under the heading of 'Rule I' in Newton's *Principia* (1687, p.320). More recently, the sentiment was echoed by Einstein:

> [T]he grand aim of all science... is to cover the greatest possible number of empirical facts by logical deductions from the smallest possible number of hypotheses or axioms (Einstein, in (Nash 1963, p. 173)).

These examples are all taken from physicists, since physics is the science to which Gödel thought mathematics most analogous. There are examples to be found from across the range of the sciences, however.[34] While a full sociological or historical

[32]There is a distinct view, proposed by Maddy (1997), that mathematics has its own *autonomous* theoretical virtues, and that at least certain large large cardinal axioms score very well on this front. My argument will have nothing to say for or against such a view; here I am just focusing on the theoretical virtues that drop out of the analogy with the natural sciences.

[33]Of course, Kant didn't invent the principle that the non-ontological aspect of a theory should be a simple as possible. But then again, Ockham didn't invent the corresponding ontological principle. The formulation of Kant's razor here is taken from remarks at A652/B680 of the *Critique of Pure Reason* (1787, p.595).

[34]Baker's article (2016) contains a veritable trove of such examples, from many sub-fields of both philosophy and science.

investigation is out of the question here, it is sufficient for my purposes merely that theoretical and ontological simplicity are important virtues in the natural sciences. Since that is a rather unremarkable claim, I'll proceed with the argument that large large cardinal axioms should automatically score poorly when evaluated with respect to parsimony (of both relevant kinds).

Firstly, the adoption of large cardinal axioms will substantially bloat the ontology of mathematics in a fairly straightforward way: such axioms tell us that there are more sets than were previously thought. Indeed, large large cardinal axioms often tell us that there will be *drastically* many more levels in the hierarchy than previously thought, since a relatively common feature of such axioms is that they imply the existence of an unbounded class of cardinals satisfying weaker large cardinal hypotheses.

On a straightforward reading of Ockham's principle, this observation is sufficient to show that large large cardinal axioms will score poorly on the front of ontological parsimony. Adding a large large cardinal axiom to **ZFC** involves massively bloating the size of the ontology of the theory, and these entities will, in a strict sense, have been multiplied 'beyond necessity'. After all, the arithmetical data accounted for by a large cardinal principle will equally be accounted for by a corresponding consistency sentence. In a more general sense, **ZFC** is already powerful enough to reproduce all of 'ordinary' (i.e. non-foundational) classical mathematics; so even if the multiplication of entities brought about by a large cardinal axiom is in some way desirable, or virtuous, it is certainly beyond necessity.

An immediate objection would be that Ockham's razor, as a general principle, is not supposed to count against theories which posit *more* entities (all else being equal), rather it is supposed to count against theories which posit *more kinds* of entities (all else being equal).[35] An objector might claim then, that large large cardinal axioms do not at all imply the existence of new kinds of entities; rather they imply the existence of (many, many) more entities of the same kind, namely sets. It would not count against a theory in physics, the objector might say, if it entailed that there are more entities than previously supposed of a kind we already countenance, such as electrons. So why should positing more sets count against large cardinal

[35]This qualitative version of Ockham's Razor was famously championed by Lewis (1973, p.87).

hypotheses in set theory, since set theory is itself analogous to a natural science?

The problem with this suggestion is that Ockham's razor can be rendered *trivial* by permitting sufficiently wide kinds. There is clearly a good deal of slack in the notion of a kind of entity, at least for the purposes of considering parsimony principles, but the delineation of kinds for such purposes appears constrained, at least in practice. Violation of Ockham's razor played an important role in Lavoisier's critique of phlogiston theory, for instance (Baker 2016, §1). It would have been no defence to claim that phlogiston is of a kind we already accept, since it is a *physical substance* or similar. In scientific cases such as this, Ockham's razor is applied non-trivially, and so if the science–mathematics analogy is appropriate, as Gödel argues, some parallel restriction should also be in place when considering mathematics.

While I don't want to make any grand claims about what kinds of set there are, it is clear that *set* is too general a kind for the meaningful use of Ockham's razor as a principle of theory choice within mathematics. Indeed, since the universe is class-free, according to the Gödelian platonist, the ontology of set theory is exclusively given in terms of sets (or perhaps sets with numbers or integers as urelements). Hence the admissibility of *set* as a kind for the purposes of Ockham's razor would render that principle trivial within the domain of mathematics. That should not be an acceptable conclusion for the advocate of the science–mathematics analogy, since the application of the principle in the former domain is highly non-trivial.

I certainly don't want to claim that Nature Herself divides sets up into certain kinds, which are or are not subject to Ockham's razor. It is more plausible that the appropriate evaluative kinds should be based on *salience* for mathematical purposes, and hence sensitive to the investigative context. But sets come in many mathematically salient kinds. Some are ordinals, some are cardinals, some are arithmetical, some analytical, some Borel, some projective, and so on. The fact that they are all sets certainly does not mean that a large cardinal with hitherto uninstantiated properties does not constitute a new kind of entity. Indeed, the significant increase in the strength of set theory that is offered by large cardinal hypotheses render it all but certain that in any investigative context, the adoption of large large cardinal axioms will bloat the ontology of set theory beyond necessity, even if we envisage the evaluation as being about the number of kinds of entities in the ontology.

Similar points can be made about the increase in theoretical complexity incurred

by the addition of large large cardinal axioms. As we've seen, such posits are not *necessary* to account for the data, since it is sufficient that they are consistent relative to **ZFC**. So the addition of complexity to the theory is not automatically legitimate. And similarly to the case of Ockham's razor, Kant's razor tells quite strongly against the addition of large large cardinal axioms to set theory. On a straightforward understanding of the virtue of ideological simplicity, **ZFC** will fare better than its extension by any large cardinal principle, since those extra principles go beyond what is necessary to account for the mathematical data.

However, it is not entirely clear when one theory is more ideologically complex than another. Instead of looking just at the number of axioms or schemata in a theory (as on the straightforward understanding of this virtue), Quine (1951, p.14) considers the 'ideology' of a theory to be the range of ideas expressible within the theory. This corresponds roughly to the kinds-based understanding of Ockham's razor, since on this understanding one theory can contain more principles than other without having a more bloated ideology, as long as no further 'ideas' are expressible in the more verbose theory. Quine's notion of ideology is a primarily linguistic matter, the formulation in terms of ideas being (hopefully) eliminable (Quine 1951, p.15). Nonetheless, there is an ambiguity here. Are we to understand parsimony as favouring overall less expressively powerful theories, or merely theories with a smaller number of primitive expressions?

If we understand the 'expressible ideas' of a theory in terms of its primitive vocabulary, large large cardinal axioms will score neutrally with respect to ideological parsimony. This is because the addition of a large cardinal principle to set theory leaves the undefined primitives of the theory (logical vocabulary, '$\in$', and possibly a symbol to distinguish sets from urelements) undisturbed.[36]

Things are very different, however, on the former disambiguation, according to which the extent of the ideology of a theory corresponds to its general expressive power. On this understanding, large large cardinal axioms will score poorly with re-

[36]If our set theory contains urelements *and* the empty set, the inclusion of a further non-logical primitive is required to distinguish sets from urelements. There are several means by which this can be achieved: a distinguished predicate for sets, a distinguished predicate for urelements, or a singular term for the empty set. This works straightforwardly in the first two cases. In the third case, one can formally define the predicates S ('is a set') and U ('is an urelement') stipulating that $\forall x(Sx \leftrightarrow \exists y\ y \in x \vee x = \emptyset)$ and $\forall x(Ux \leftrightarrow \neg\exists y\ y \in x \wedge x \neq \emptyset)$.

spect to ideological simplicity. The addition of a large cardinal principle to set theory increases the range of definable sets lower down in the hierarchy, and correspondingly the theory will be able to express, and prove, many more 'ideas' about sets than its unsupplemented counterpart. The very reason that large cardinal axioms have Π^0_1 arithmetical consequences which are independent of **ZFC** is that more and more subsets of ω are definable under stronger and stronger large cardinal assumptions. If Quine's notion of an 'expressible idea' is cashed out in terms of propositions about definabile sets, then large cardinal axioms will be largely uneconomical. It seems likely to me that this disambiguation corresponds closely to Quine's intentions, since he claims that the classical theory of the reals has a denumerable ideology, and claims that investigation of primitive ideology is a 'subdivision' of the overall investigation (1951a, p.14). Both comments would be misleading if his intention had been to refer to the *finite* number of analytical primitives, and considered the investigation of primitive ideology to be an improper subdivision of overall ideological investigation!

Given the lack of clarity in Quine's suggestion, perhaps a broader notion of an ideologically parsimonious theory is required. However, on any reasonable construal of a theory's ideology (other than as consisting of its primitive vocabulary), large cardinals will bloat it: more properties of sets will be instantiated, new embeddings between the universe and transitive classes appear, new sets are definable, and so on. Given that large large cardinals can be used to solve set-theoretic problems not solvable in **ZFC** alone, a large large cardinal axiom will *always* give a richer picture of the hierarchy than is strictly required to explain the data, since the data is accounted for by only positing the consistency of large cardinal axioms relative to **ZFC**, without the need for any further increase in the power of the theory.

However exactly you construe the notions of ontology and ideology, a very basic problem for the large cardinal advocate appears: large cardinals axioms posit new sets (bloating its ontology), with new and remarkable properties (bloating its ideology). So, large large cardinal axioms are always purchased at the expense of two theoretical virtues which have historically played a central role in scientific theory choice. A conception of set theory which places overwhelming value on maximal ontology and richness of the structure of the hierarchy needn't be concerned by such issues at all. But on such a picture, the theoretical virtues of set theory become more distant from those of natural science.

I don't want to claim that there is *no* analogy between set theory and natural science; any things are analogous if you squint hard enough. But nor do I want to claim that set theory and natural science are strongly analogous. Rather, I'm claiming that the more analogous you think the two disciplines are, the more difficult the justification of large large cardinal axioms becomes. And *that* point spells doom for Gödel's attempt to find the justification for large large cardinal axioms in any kind of an analogy between set theory and natural science.

It's important to note, however, that parsimony considerations come with a *ceteris paribus* clause: the bloating of ontology and ideology only tells against a theory if the other virtues of the theory don't compensate. I expect that it is *possible* that the virtues of large cardinal axioms could be so overwhelming as to compensate for a bad score in both kinds of parsimony. An example of the right kind of speed-up result discussed earlier would be a possible example of this. But in actuality there is room for much scepticism of the theoretical virtues which are commonly ascribed to large cardinal axioms. Most importantly, the solution of open problems is not to be valued for its own sake when the solution provided for by an axiom does not itself enjoy extensive support from elsewhere.

I think it's worth distinguishing the argument offered here from that presented by Quine (1990, pp.94–95). The argument there is somewhat ambiguous. On the one hand, Quine suggests that 'higher' set theory is *meaningless*, because, whatever the axioms that constitute higher set theory are supposed to be, they never have any implications for natural science. They are treated by us as meaningful only because to do otherwise would constitute an 'unnatural gerrymandering of grammar'. Another argument offered, however, is that the questions of higher set theory are, at least in part, settled by parsimony considerations. In particular, Quine argues that considerations of simplicity, economy, and naturalness compel us to adopt $V = L$ as a new axiom, since it 'inactivates the more gratuitous flights of higher set theory'. I take it that this is a reference to the inconsistency of $V = L$ with most large large cardinal axioms, though Quine is not specific. I'm unsure how to reconcile these two arguments, since a decision of the kind Quine envisages seemingly involves regarding the relevant axiom candidates as meaningful. After all, it is hard to see how one uninterpreted string of symbols can give a more natural or economical picture of the hierarchy than another. In any case, Quine claims that considerations of theoretical

virtue tell decisively against large large cardinal axioms.

My argument, on the contrary, involves no such claim, and approaches the problem from an entirely different perspective. In the first instance, Quine subjects set theory to evaluation in terms of the theoretical virtues of natural science and with respect to scientific applications, since presumably his holistic naturalism implies that this is the only appropriate set of virtues to figure in *any* theory choice, regardless of the subject matter. Gödel's analogy, on the other hand, requires only that the means of theory choice be analogous between mathematics and science, and does not require that the theoretical virtues of a putative axiom of set theory be considered in relation to its application in natural science. Since I'm here in the business of assessing Gödel's position, my argument does not consider the virtues of set theory as they relate to scientific applications.

Secondly, and more significantly, I have not attempted to offer an argument that $V = L$ is true, nor have I even offered an argument that we shouldn't accept large large cardinal axioms in general. Rather, I have argued for the much weaker claim that, however such axioms are justified (if at all), it does not look much like how theoretical posits are justified in natural science. The argument presented here therefore should certainly not be confused with the one offered by Quine, despite the central role played by considerations of economy in each.

Conclusion

I've argued that whatever analogy the conceptual platonist might see between mathematics and the natural sciences, it cannot serve as a justification for large large cardinal axioms. Three attempts were offered to provide these axioms with a viable quasi-scientific justification, inspired by remarks made by Gödel. None of them proved to be successful. Although this is not to say that such arguments cannot justify the existence of *any* sets, such an account would be redundant for the conceptual platonist who thinks that the weaker axioms follow directly from the iterative conception.

Firstly, we saw that large large cardinal axioms cannot have roughly the status of propositions asserting the existence of ordinary material bodies. This would afford

us an enormous degree of confidence in the existence of larger cardinals, but the account is not viable. In particular, the large cardinals cannot play the same kind of explanatory role that posited material bodies do.

More promisingly, we investigated the idea that large cardinal axioms could play the role within mathematics played by laws of nature in science, as pioneered by Russell. The statements of scientific laws are strongly supported because they allow us to predict the initial data, regardless of the degree of intuitive appeal such principles may have. I argued, however, that large cardinal axioms cannot enjoy this same degree of regressive support.

Various candidates for the mathematical data were considered. The only viable conception of the data on offer is that propositions acting as data are expressed by Π^0_1 arithmetical sentences and are either hard data in Russell's sense, are generated by the Gödelian incompleteness of a theory we believe to be sound, or perhaps are such that a strong heuristic justification can be offered for their truth, as with Goldbach's conjecture.

The problem for the Gödelian is that accounting for such data affords justification only to the propositions that such large cardinal axioms are consistent relative to **ZFC**, and not to the truth of the axioms themselves. The trouble is not that we don't know whether such axioms are consistent; the independence results that proliferate in modern set theory should teach us to be less ambitious than that. Rather it is that the consistency of a large cardinal axiom is a strictly weaker proposition than its truth, and is alone sufficient to account for any data in the relevant sense.

Quite aside from considerations of data, I've argued further that the justification of large large cardinal axioms does not look much like the justification of virtue-enhancing principles in science, since adding a large cardinal axiom to a theory always causes significant bloating to the ontology and ideology of a theory, a practice which is anathema to the modes of theory choice in natural science where simplicity and parsimony are highly respected arbitrators between competing empirically equivalent theories.

The problem here can be put in quite simple terms: between empirically equivalent theories, the mode of theory choice in natural science is minimising and conservative with respect to ontology and ideology. Since large large cardinal axioms are *maximising* with respect to ontology and ideology, it follows that either the modes

of theory choice in mathematics are not much like their scientific counterparts, or that large large cardinal axioms fail to be justified. Unlike Quine, I don't wish to take a side on this matter; the disjunction is sufficient to make my point, which is that this well-received aspect of Gödel's thought is ultimately not fit for purpose as regards large large cardinal axioms.

Unlike the case with Gödelian incompleteness, we do not here have a strong reason to believe that set-theoretic incompleteness is absolutely ineliminable. Rather, we simply have good cause to think that the means of alleviating it proposed by Gödel have met with at most a modest degree of success. The need to find a compelling justification for the larger large cardinal axioms is of some urgency, however. The general programme of formulating large cardinal axioms and investigating the consequences of their assumption is one of the central research areas in the foundations of mathematics. It would be a philosophical scandal if we could say no more than that this programme involved its practitioners in mere 'if-then-ist' thumb-twiddling. The mathematical significance of the large cardinals enterprise demands philosophical explanation, preferably one which justifies a contentful interpretation of the consistency-constrained maximalism at work in current set-theoretic practice. Unfortunately, this explanation cannot be given by means of an analogy between mathematics and the sciences.

Concluding Remarks

Ineliminable Incompleteness

The overarching question of this thesis has been: how can we strengthen our axiomatic mathematical theories so as to reduce the degree to which they are incomplete? I've argued that there are two distinct kinds of incompleteness; Gödelian and set-theoretic, and addressed various attempts to eliminate or reduce both kinds. I've approached the question via close engagement with four themes found in the writings of Kurt Gödel: anti-mechanism, rationalistic optimist, conceptual platonism, and quasi-scientific methods. At times, the views with which I've engaged have been, of necessity, only tentative reconstructions of Gödel's views based on the scarcity of available material. Nevertheless such views are recognizably Gödelian. Indeed, if I've convinced you only that Gödel's philosophical views are richer and more deserving of attention than you thought they were before, I can count my efforts as a success. But nonetheless, I hope to have convinced you that the prospects for eliminating either kind of incompleteness look bleak, and our attempts at justifiably strengthening the axioms of arithmetic and set theory by any of these methods can hope only for a modest degree of success. That is my answer to the central question of this discussion. A good deal more than that was argued for, however:

Anti-Mechanism

Mechanism is most neutrally described as the view that our mathematical output is, under idealisation, coextensive with the output of a Turing machine. Anti-mechanism, most neutrally stated, is the view that our idealised mathematical output is more extensive than that of any Turing machine. It is a view to which Gödel was inclined, but one for which he offered no comprehensive philosophical argument. We are entitled, however, not just to an argument for the view, but also to an explanation of *how* the idealised mind proves things in a way that could not be executed by any Turing machine.

Since Gödel believed that *any* arithmetical proposition could be proved by the

human mind, under suitable idealisation, I suggested that a natural model of our in-principle arithmetical capacities on such a view would be a reflection sequence obtained by the transfinite iteration of a soundness reflection principle. Feferman's completeness theorem guarantees that, by reflecting on the soundness of **PA**, any arithmetical truth can be proved, making this a plausible fleshing-out of Gödel's hopes that our arithmetical capacities outstrip those of any Turing machine.

I argued however, that thinking of our arithmetical abilities in this way is inconsistent with any plausible epistemology of arithmetic, since it presupposes that the idealised mathematician already possesses (some of) the arithmetical knowledge that the account was supposed to explain. Gödelian anti-mechanism was therefore rejected, despite the significant (indeed, total) reduction in Gödelian incompleteness the theory promised. I argued further that similar criticisms could be levelled at much weaker forms of anti-mechanism, based on a study of Turing's completeness theorem.

Absolute Undecidability

There may be more palatable forms of anti-mechanism in the vicinity, so the argument of the previous chapter doesn't establish anything definitive about the nature of the human mind. However, it does allow us to build up a positive argument for the absolute undecidability of certain arithmetical propositions.

We saw that Gödel's two main arguments against the existence of absolutely undecidable arithmetical propositions were very persuasive in a limited range of cases, but failed to establish their conclusion in sufficient generality. In particular, we saw that they relied on the assumption that we possess an ability to select ordinal notations in a very particular way, or else an ability equivalent to this. However, there is no reason to suppose that we possess this ability (and no argument for it has been offered to my knowledge), as well as several reasons to suppose we don't: the ability provably corresponds to no recursive procedure, and any epistemology of arithmetic according to which we possess such an ability would suffer from the same problems as Gödelian anti-mechanism.

If we lack this recursive ordinal selection ability, it follows that certain arithmetical propositions are absolutely undecidable, and hence that Gödelian incompleteness

is ineliminable. It does *not* follow, as Gödel feared, that human reason is 'irrational' or 'inconsistent', since the undecidable propositions are not ones that we could ever recognize to be such.

Feferman's theorem, in addition to playing a central role in my arguments for absolute undecidability and anti-anti-mechanism, can also be used to refute Dummett's argument for the vagueness of the concept *natural number*. The ordinal bound required for an arithmetically complete theory is so low that it is intuitionistically acceptable. Coupled with an argument that adding Feferman's reflection principle to a theory is intuitionistically justified by the soundness of the theory, I argued that even by Dummett's own lights, the concept *grounds for asserting something of all natural numbers* is not indefinitely extensible, and that therefore the concept *natural number* is not vague.

I finally argued that the ineliminability of Gödelian incompleteness should prompt us to accept a kind of quietism about the limits of our arithmetical knowledge.

Conceptual Platonism

The remainder of the thesis was dedicated to questions of reducing set-theoretic incompleteness. Gödel's approach to the problem makes sense only in terms of his platonist interpretation of mathematics: his notion of intrinsic justification is bound up with his view of mathematical intuition, and his use of extrinsic justification is grounded in realism about sets. I argued that, although Gödel's position shifted substantially over time, we can isolate a Gödelian position called *conceptual platonism*. According to this view, certain mathematical *concepts* are such that we can gain *intuitive* (i.e. non-deductive and non-empirical) knowledge of axioms quantifying over the objects falling under such concepts by reflecting on what the concepts commit us to. Hence, Gödel's platonism about concepts is prior to his platonism about objects, and intuitive knowledge plays a crucial role in determining how our mathematical theories are to be axiomatized, since it is in axiomatic theories that our grasp of mathematical concepts is expressed.

Although I defended this theory against charges of mysticism and theology, I argued that it is insufficiently developed to merit endorsement since, although we

are given examples by Gödel of concepts which have the required kind of objective content, it is unclear what the having of such content amounts to. Gödel claims that concepts of the relevant kind are axiomatized in a non-arbitrary way, such that the system strikes us as clearly sound; while this is somewhat helpful, the criterion will not allow us to precisely distinguishing concepts with objective content from those which lack it. However, conceptual platonism is clear enough to examine Gödel's attempts to reduce set-theoretic incompleteness by means which are closely related to it.

Small Large Cardinals

Distinctively set-theoretic incompleteness perhaps cannot be isolated from a first-order perspective. But from a second-order perspective we can see that the inability of set-theoretic axioms to determine the height of the hierarchy is not of a piece with the inability of **PA** to prove that it is consistent. Given that Gödel thought some set-theoretic axioms can be known on the basis of intuition, a natural attempt to reduce set-theoretic incompleteness is by certifying certain large cardinal axioms as following the concept *set*.

I argued, against Potter, that the conceptual platonist can justifiably regard second-order reflection principles as following from the concept *set*, but only if they are willing to abandon platonism about values of any variables of third-order and higher. We saw, however, that Koellner's theorems make it extremely unlikely that any large cardinal axiom stronger than the existence of the Erdős cardinal $\kappa(\omega)$ could be justified by these means. Hence the prospects for reducing incompleteness substantially by mathematical intuition is somewhat bleak, since by modern standards the cardinals that could perhaps be known to exist intuitively would be rather small.

Having clarified the shape that Gödelian platonism takes in the case of set theory, I argued that Gödel's claim in the Gibbs lecture that such platonism is supported by the incompleteness theorems is vindicated. However, the degree of support which the incompleteness theorems lend to platonism is very limited, as the constraints they impose on anti-platonist theories are consistent with most of platonism's main rivals. Moreover, the incompleteness of a theory should only support a platonist

interpretation of it in cases where, by Gödel's own lights, we already have good cause to be a platonist.

Large Large Cardinals

According to Gödel, our means of justifying set-theoretic axioms is by no means exhausted by our faculty of intuition. Given that the entities of mathematics are ontologically on a par with those of the natural sciences, Gödel supposes that the kinds of reasoning that license belief in the latter can be also applied in the mathematical case. If such methods could be used to justify large cardinal axioms going beyond what is justifiable using intuition, then a substantial reduction in incompleteness could be effected.

We saw that the very strongest justification that we have for belief in material bodies is inapplicable in the case of large large cardinals. I also argued that the regressive method of justifying laws of nature in science is inapplicable in mathematics, at least as far as such axioms are concerned. This is because mathematical propositions which might plausibly function as data are expressible by Π^0_1 arithmetical sentences, and in this restricted arena, the consistency of a large cardinal axiom relative to **ZFC** is as good as its truth.

Lastly, I argued that the prospects for justifying large cardinal axioms via an analysis of their theoretical virtues were extremely limited. In the natural sciences, principles which minimize the richness of the theory and the size of its ontology are highly valued, and large large cardinal axioms are guaranteed to score poorly on this front. Hence the methods for reducing set-theoretic incompleteness suggested by Gödel are of very limited use. This gives us some cause to think that set-theoretic incompleteness might be just as much an essential part of mathematics as Gödelian incompleteness was shown to be.

The Large Cardinals Programme

I'd like to end by considering, albeit briefly, how a platonist might try to place the consistency-constrained maximalism of the large cardinals programme on a solid

philosophical footing, since that can't be achieved by means of an analogy between set theory and the sciences, nor by intuition. Throughout several of Gödel's writings, we find him comparing the axioms of set theory with those in geometry, in general finding their similarities more striking earlier in his career and less persuasive later on.

A philosophical interpretation of geometry that has become relatively well-accepted is offered by Einstein (1921), according to which each consistent geometry is true, though not perhaps true of physical space. Rather, each consistent geometry determines its own kind of space in which the axioms are true. In his lecture at Göttingen (1939, p.155), Gödel notes that the use of $V = L$ to prove the consistency of the continuum hypothesis is very similar to results obtained in recent axiomatic geometry. More significantly, he proposes that a proof of the *independence* of CH from **ZFC** would demonstrate its absolute undecidability, and thus set theory would 'bifurcate' into multiple different legitimate systems, as with Euclidean and non-Euclidean geometry, presumably at each absolutely undecidable proposition.

Of course, Gödel retreats from this view as his confidence in rationalistic optimism and the quasi-scientific justification of set-theoretic axioms increases; by 1966 his view was that the independence of CH from **ZFC** had no substantial philosophical implications (1966, p.372).[37] But the 'geometric view' entertained in the Göttingen lecture might be a means of justifying extensions of **ZFC** by large cardinal axioms without a direct intuitive justification.

The idea would be that many distinct set-theoretic universes are consistent with iterative conception, none of which can be said to privileged over another as *the* universe. One reason for this might be that many different concepts constitute an admissible precisification of the ill-defined maximality principle behind the iterative conception. All of these concepts might be such that their axiomatizations overlap considerably as regards a range of core principles, and diverge only in the strength of axioms embodying the idea that it is possible to iterate the 'set-of' operation at least

[37]Which is not to say that Gödel had ceased to find the analogy between set theory and geometry illuminating; despite having claimed in 1938 that $V = L$ was a 'natural completion of the axioms of set theory' (1938, p.27), Gödel later claim to think that $V = L$ should be rejected on the grounds that it states a *minimum* property of the hierarchy. A more favourable axiom stating a maximum principle would be a set-theoretic analogue of Hilbert's completeness axiom in geometry (1964, pp.262–263, fn.23).

a certain distance through the ordinals. A somewhat similar view has recently been proposed by Hamkins (2012), according to whom a proliferation of set-theoretic universes obtained by forcing exist. A Gödelian view inspired by the Göttingen lecture needn't be quite so extreme.

It is possible, for instance, that large cardinals be given a special place in determining the existence conditions for universes, thanks to the quasi-categoricity theorem and Gödel's second-order conception of the axioms. If everything about the iterative conception *except* for the strength of its maximality principle is determinate, then the Gödelain multiverse will be substantially smaller than that envisaged by Hamkins, because the universes will disagree only with respect to height. He denies that we have so determinate a second-order conception of property as is required to make such an idea work, but this is certainly up for debate, especially since the crucial properties are those of a well-order and natural number (according, at least, to Martin (2001)). The argument presented in chapter 2 against one kind of indeterminacy in the concept *natural number* can be expected to be of use here.

This is a significant retreat from any 'bifurcation' in set theory at an absolutely undecidable proposition; if we view set-theoretic universes from a second-order perspective there is of course no room for disagreement about CH, for example. Some pluralism is maintained, but Gödel's mature platonism is better-represented; as Hamkins puts it, multiverse views are a kind of 'higher-order platonism' (2012, p.417). Absolutely undecidable propositions of the kind discussed in chapter 2 would all have a determinate truth-value, and the smallest universes in the multiverse would be isomorphic models of the minimal set theory delivered by the iterative conception.[38] This will be no stronger that $\mathbf{ZFC}_2 + \Gamma^2_n$ reflection, though as I argued in chapter 4, Gödel might be safer with a weaker theory, abandoning these reflection principles in favour of maintaining realism about higher-order properties. Either way, it is likely that a Gödelian multiverse would be more restricted than Zermelo's unbounded series of models of set theory, given that the minimal theory considered in (Zermelo 1930) (i.e. second-order **ZF** without the axiom of infinity) is somewhat weaker than what Gödel thought could be intuitively justified.

Given the special place of large cardinal axioms on the view sketched, the large

[38]Of course, height-sensitive propositions such as GCH might not be settled here.

cardinals programme is thus given a secure philosophical basis: all large cardinal axioms consistent with the basic iterative conception can be seen as explicating a kind of maximality principle, and are true in some universes. The project of articulating a (somewhat) Gödelian multiverse view therefore strikes me as a promising one, but it is a project for another day.

Bibliography

Anselm of Canterbury, St. (1077/8). *Proslogium,* trans. Dean, S. In: The Internet Medieval Sourcebook at Fordham University. URL: `https://sourcebooks.fordham.edu/basis/anselm-proslogium.asp` (visited on 20/05/2018).

Baker, A. (2007). "Is there a problem of induction for mathematics". In: Leng, M., A. Paseau, and M. Potter (2007), pp. 59–73.

— (2016). "Simplicity". In: Zalta, E. (ed.) (2016) *The Stanford Encyclopedia of Philosophy* (Winter 2016 Edition). URL: `https://plato.stanford.edu/archives/win2016/entries/simplicity/` (visited on 06/08/2018).

Benacerraf, P. (1967). "God, the Devil, and Gödel". In: *The Monist* 51, pp. 9–32.

Bernays, P. (1961). "On the problem of schemata of infinity in axiomatic set theory". In: Müller, G. (1976), pp. 121–172.

Boolos, G. (1971). "The Iterative Conception of Set". In: *The Journal of Philosophy* 68, pp. 215–231.

— (1987). "A Curious Inference". In: *Journal of Philosophical Logic* 16, pp. 1–12.

— (1989). "Iteration Again". In: *Philosophical Topics* 17, pp. 5–21.

Bueno, O. and Ø. Linnebo, eds. (2009). *New Waves in the Philosophy of Mathematics*. Basingstoke: Palgrave Macmillan.

Buss, S. (1994). "On Gödel's Theorems on Lengths of Proofs I". In: *The Journal of Symbolic Logic* 59, pp. 737–756.

Button, T. and S. Walsh (2018). *Philosophy and Model Theory*. Oxford University Press.

Caicedo, A., J. Cummings, P. Koellner and P. Larson, eds. (2017). *Foundations of Mathematics*. Providence, RI: American Mathematical Society.

Chihara, C. (1982). "A Gödelian Thesis Regarding Mathematical Objects: Do They Exist? and Can We Percieve Them?" In: *The Philosophical Review* 91, pp. 211–227.

— (1990). *Constructibility and Mathematical Existence*. Oxford: Clarendon Press.

Davis, M. (1995). "Introductory note to *193?" In: Feferman, S. et al. (1995), pp. 156–163.

Dedekind, R. (1888). "The Nature and Meaning of Numbers". In: Paseau, A. (2017), pp. 203–245.

Dummett, M. (1963). "The Philosophical Significance of Gödel's Theorem". In: Dummett (1978), pp. 186–201.

— (1978). *Truth and Other Enigmas*. Cambridge, MA: Harvard University Press.

— (1991). *Frege: Philosophy of Mathematics*. London: Duckworth.

— (1994). "Reply to Wright". In: McGuinness, B. and G. Oliveri (1994), pp. 329–338.

— (2000). *Elements of Intuitionism* (2nd. ed.) Oxford University Press.

Echeverria, J. (1996). "Empirical methods in mathematics: A case study: Goldbach's Conjecture". In: Munévar, G. (1996), pp. 19–56.

Einstein, A. (1921). "Geometry and Experience". In: Paseau, A. (2017), pp. 246–255.

Ewald, W. (2005). *From Kant to Hilbert Volume II*. Oxford University Press.

Feferman, S. (1960). "Arithmetization of Metamathematics in a General Setting". In: *Fundamenta Mathematicae* 49, pp. 35–92.

— (1962). "Transfinite Recursive Progressions of Axiomatic Theories". In: *Journal of Symbolic Logic* 3, pp. 259–316.

— (1988). "Turing in the Land of O(z)". In: Herken, R. (1988), pp. 131–147.

— (1998). "Ordinal Logics". In: *The Routledge Encyclopedia of Philosophy*. URL: `https://www.rep.routledge.com/articles/thematic/ordinal-logics/v-1` (visited on 28/11/2017).

— (2006). "Are There Absolutely Unsolvable Diophantine Problems?" In: *Philosophia Mathematica* 14, pp. 134–152.

Feferman, S., J. Dawson, W. Goldfarb, C. Parsons and W. Sieg, eds. (2014). *Kurt Gödel: Collected Works, Volume IV*. Oxford University Press.

Feferman, S., J. Dawson, W. Goldfarb, C. Parsons and R. Solovay, eds. (1995). *Kurt Gödel: Collected Works, Volume III*. Oxford University Press.

Feferman, S., J. Dawson, S. Kleene, G. Moore, R. Solovay and J. van Heiejenoort, eds. (1986). *Kurt Gödel: Collected Works, Volume I*. Oxford University Press.

— eds. (1990). *Kurt Gödel: Collected Works, Volume II*. Oxford University Press.

Feferman, S., C. Parsons and S. Simpson, eds. (2010). *Kurt Gödel: Essays for his Centennial*. Oxford University Press.

Feferman, S. and C. Spector (1962). "Incompleteness Along Paths in Progressions of Theories". In: *The Journal of Symbolic Logic* 27, pp. 383–390.

Fraenkel, A., Y. Bar-Hillel and A. Levy (1958). *Foundations of Set Theory.* Amsterdam: North-Holland.

Franzén, T. (2004). *Inexhaustibility.* Wellesley, MA: Association for Symbolic Logic.

— (2004a). "Transfinite Progressions: A Second Look at Incompleteness". In: *Bulletin of Symbolic Logic* 10, pp. 367–389.

Frege, G. (1884). *The Foundations of Arithmetic,* trans. Austin, J. (2nd ed.) (1959). Evanston, IL: Northwestern University Press.

— (1899). "Frege to Hilbert 27.12.1899". In: Gabriel, G. (1980), pp. 34–38.

— (1900). "Frege to Hilbert 6.1.1900". In: Gabriel, G. (1980), pp. 43–48.

Gabriel, G., ed. (1980). *Gottlob Frege: Philosophical and Mathematical Correspondence.* Oxford: Blackwell.

Galileo, G. (1632). *Dialogue Concening the Two Chief World Systems,* trans. Drake, S. (1962). Berkeley, CA: University of California Press.

Gödel, K. (193?). "Undecidable diophantine propositions". In: Feferman, S. et al. (1995), pp. 164–175.

— (1933). "The present situation in the foundations of mathematics". In: Feferman, S. et al. (1995), pp. 45–53.

— (1936). "On the lengths of proofs". In: Feferman, S. et al. (1986), pp. 397–398.

— (1938). "The consistency of the axiom of choice and of the generalized continuum hypothesis". In: Feferman, S. et al. (1990), pp. 26–27.

— (1939). "Lecture at Göttingen". In: Feferman, S. et al. (1995), pp. 126–155.

— (1944). "Russell's mathematical logic". In: Feferman, S. et al. (1990), pp. 176–187.

— (1946). "Remarks before the Princeton bicentennial conference on problems in mathematics". In: Feferman, S. et al. (1990), pp. 150–153.

— (1947). "What is Cantor's continuum problem?" In: Feferman, S. et al. (1990), pp. 176–187.

— (1951). "Some basic theorems on the foundations of mathematics and their implications". In: Feferman, S. et al. (1995), pp. 304–323.

— (1953/9). "Is mathematics syntax of language? v.III and v.V". In: Feferman, S. et al. (1990), 334–356 and 356–362.

Gödel, K. (1961/?). "The modern development of the foundations of mathematics in the light of philosophy". In: Feferman, S. et al. (1995), pp. 374–387.

— (1964). "What is Cantor's continuum problem?" In: Feferman, S. et al. (1990), pp. 254–270.

— (1966). "Gödel to Church, September 29, 1966". In: Feferman, S. et al. (2014), pp. 372–373.

— (1970). "Ontological proof". In: Feferman, S. et al. (1995), pp. 403–404.

Good, I. (1967). "Human and Machine Logic". In: *The British Journal for the Philosophy of Science* 18, pp. 144–47.

— (1969). "Gödel's Theorem is a Red Herring". In: *The British Journal for the Philosophy of Science* 19, pp. 357–358.

Hamkins, J. (2012). "The Set-Theoretic Multiverse". In: *The Review of Symbolic Logic* 5, pp. 416–449.

Herkin, R., ed. (1988). *The Universal Turing Machine: A Half-Century Survey*. Oxford University Press.

Hilbert, D. (1899). "Hilbert to Frege 29.12.1899". In: Gabriel, G. (1980), pp. 38–43.

Horsten, L. and P. Welch, eds. (2016). *Gödel's Disjunction*. Oxford University Press.

— (2016a). "Reflecting on Absolute Infinity". In: *The Journal of Philosophy* 113, 89–111. Pagination cited is from the online version: URL: `https://research-information.bristol.ac.uk/en/publications/reflecting-on-absolute-infinity(c27493b5-c162-4c98-a42e-282acf9c5261).html` (visited on 16/01/2019).

Incurvati, L. (2016). "Maximality Principles in Set Theory". In: *Philosophia Mathematica* 25, pp. 159–193.

Isaacson, D. (1987). "Arithmetical Truth and Hidden Higher-Order Concepts". In: Paseau, A. (2017a), pp. 89–108.

Jech, T., ed. (1974). *Axiomatic Set Theory*. Providence, RI: American Mathematical Society.

— (2003). *Set Theory: The Third Millenium Edition*. Belin: Springer-Verlag.

Kanamori, A. (2009). *The Higher Infinite* (2nd ed.) Berlin: Springer-Verlag.

Kant, I. (1787). *Critique of Pure Reason* (2nd ed.) Trans. Guyer, P. and A. Wood (1998). Cambridge University Press.

Koellner, P. (2009). "On Reflection Principles". In: *Annals of Pure and Applied Logic* 157, pp. 206–219.

— (2009a). "Truth in Mathematics: The Question of Pluralism". In: Bueno, O. and Ø. Linnebo (2009), pp. 80–116.

— (2010). "Absolute Undecidability". In: Feferman, S. et al. (2010), pp. 189–225.

Leach-Krouse, G. (2016). "Provability, Mechanism, and the Diagonal Problem". In: Horsten, L. and P. Welch (2016), pp. 211–242.

Leng, M., A. Paseau and M. Potter, eds. (2007). *Mathematical Knowledge*. Oxford University Press.

Lewis, D. (1973). *Counterfactuals*. Oxford: Basil Blackwell.

Linnebo, Ø. (2018). *Thin Objects*. Oxford University Press.

Lucas, R. (1961). "Minds, Machines and Gödel". In: *Philosophy* 36, pp. 112–127.

— (1968). "Human and Machine Logic: A Rejoinder". In: *The British Journal for the Philosophy of Science* 19, pp. 155–156.

— (1996). "Minds, Machines and Gödel: A Retrospect". In: Millican, P. and A. Clark (1996), pp. 103–124.

McCallum, R. (2017). "Intrinsically Justified Reflection Principles". URL: `https://arxiv.org/pdf/1403.8058.pdf` (visited on 15/09/2018).

Maddy, P. (1990). *Realism in Mathematics*. Oxford: Clarendon Press.

— (1997). *Naturalism in Mathematics*. Oxford: Clarendon Press.

Marshall R., M. Victoria (1989). "Higher Order Reflection Principles". In: *The Journal of Symbolic Logic* 54, pp. 474–489.

Martin, D. (1970). "Measurable Cardinals and Analytic Games". In: *Fundamenta Mathematicae* 66, pp. 287–291.

— (1975). "Borel Determinacy". In: *Annals of Mathematics* 102, pp. 363–371.

— (2001). "Multiple Universes of Sets and Indeterminate Truth Values". In: *Topoi* 20, pp. 5–16.

— (2005). "Gödel's Conceptual Realism". In: *Bulletin of Symbolic Logic* 11, pp. 207–224.

Martin, D. and J. Steel (1989). "A Proof of Projective Determinacy". In: *Journal of the American Mathematical Society* 2, pp. 71–125.

McGuinness, B. and G. Oliveri, eds. (1994). *The Philosophy of Michael Dummett*. Dordrecht: Kluwer Academic Publishers.

Mellor, H., ed. (1990). *F.P. Ramsey: Philosophical Papers*. Cambridge University Press.

Millican, P. and A. Clark, eds. (1996). *Machines and Thought: The Legacy of Alan Turing*. Oxford University Press.

Moore, A. (1998). "More on the 'The Philosophical Significance of Gödel's Theorem'". In: *Grazer Philosophische Studien* 55, pp. 103–126.

Moore, G., ed. (2014). *The Collected Papers of Bertrand Russell, Volume 5*. Oxford University Press.

Moschovakis, Y. (1980). *Descriptive Set Theory*. Amsterdam: North Holland.

Müller, G., ed. (1976). *Sets and Classes: On the Work of Paul Bernays*. Amsterdam: North-Holland.

Munévar, G., ed. (1996). *Spanish Studies in the Philosophy of Science*. Dordrecht: Springer.

Nash, L. (1963). *The Nature of the Natural Sciences*. Boston, MA: Brown, Little.

Newton, I. (1687). *The Principia: Mathematical Principles of Natural Philosophy*. Snowball Publishing.

Oliver, A. (1998). "Hazy Totalities and Indefinitely Extensible Concepts: An Exercise in the Interpretation of Dummett's Philosophy of Mathematics". In: *Grazer Philosophische Studien* 55, pp. 25–50.

Oliver, A. and T. Smiley (2016). *Plural Logic* (2nd. ed.) Oxford University Press.

Parsons, C. (1995). "Platonism and Mathematical Intuition in Kurt Gödel's Thought". In: *Bulletin of Symbolic Logic* 1, pp. 47–74.

— (2008). *Mathematical Thought and its Objects*. Cambridge University Press.

Paseau, A. (2007). "Boolos on the Justification of Set Theory". In: *Philosophia Mathematica* 15, pp. 30–53.

— ed. (2017). *Philosophy of Mathematics Volume I*. London: Routledge.

— ed. (2017a). *Philosophy of Mathematics Volume IV*. London: Routledge.

Penrose, R. (1994). *Shadows of the Mind*. Oxford University Press.

Potter, M. (2001). "Was Gödel a Gödelian Platonist?" In: *Philosophia Mathematica* 9, pp. 331–346.

— (2004). *Set Theory and its Philosophy*. Oxford University Press.

Putnam, H. (1960). "Minds and Machines". In: Putnam, H. (1975a), pp. 362–385.

— (1975). "What is Mathematical Truth?" In: Putnam, H. (1979), pp. 60–78.

— (1975a). *Philosophical Papers Volume 2*. Cambridge University Press.

— (1979). *Philosophical Papers Volume 1* (2nd ed.) Cambridge University Press.

Quine, W. (1951). "Ontology and Ideology". In: *Philosophical Studies* 2, pp. 11–15.

— (1951a). "Two Dogmas of Empiricism". In: Quine, W. (1980), pp. 20–46.

— (1980). *From a Logical Point of View* (2nd ed.) Cambridge, MA: Harvard University Press.

— (1990). *Pursuit of Truth*. Cambridge, MA: Harvard University Press.

Ramsey, F. (1925). "The Foundations of Mathematics". In: Mellor, H. (1990), pp. 164–224.

Rayo, A. and G. Uzquiano, eds. (2006). *Absolute Generality*. Oxford University Press.

Reinhardt, W. (1974). "Remarks on reflection principles, large cardinals, and elementary embeddings". In: Jech, T. (1974), pp. 189–205.

Roberts, S. (2017). "A Strong Reflection Principle". In: *The Review of Symbolic Logic* 10, pp. 651–662.

Russell, B. (1907). "The Regressive Method of Discovering the Premises of Mathematics". In: Moore, G. (2014), pp. 571–580.

— (1908). "Mathematical Logic as based on the Theory of Types". In: *American Journal of Mathematics* 30, pp. 222–262.

— (1912). *The Problems of Philosophy*. Oxford University Press.

— (1914). *Our Knowledge of the External World*. At: Project Gutenberg. URL: `http://www.gutenberg.org/files/37090/37090-h/37090-h.htm` (visited on 07/09/2018).

Schindler, R. (2000). "Proper Forcing and Remarkable Cardinals". In: *Bulletin of Symbolic Logic* 6, pp. 176–184.

Schirn, M., ed. (1998). *The Philosophy of Mathematics Today*. Oxford: Clarendon Press.

Shapiro, S. (1998). "Incompleteness, Mechanism, and Optimism". In: *Bulletin of Symbolic Logic* 4, pp. 273–302.

— (2016). "Idealization, Mechanism, and Knowability". In: Horsten, L. and P. Welch (2016), pp. 189–207.

Shapiro, S. and C. Wright (2006). "All Things Indefinitely Extensible". In: Rayo, A. and G. Uzquiano (2006), pp. 255–304.

Shelah, S. (1974). "Infinite Abelian Groups, Whitehead Problem, and Some Constructions". In: *Israel Journal of Mathematics* 18, pp. 243–256.

Sklar, L. (1992). *Philosophy of Physics*. Boulder, CO: Westview Press.

Smith, P. (2013). *An Introduction to Gödel's Theorems* (2nd ed.) Cambridge University Press.

Tait, W. (1981). "Finitism". In: *The Journal of Philosophy* 78, pp. 524–546.

— (1998). "Zermelo's Conception of Set Theory and Reflection Principles". In: Schirn, M. (1998), pp. 469–483.

— (2005). "Constructing Cardinals from Below". In: Tait, W. (2005a), pp. 469–483.

— (2005a). *The Provenance of Pure Reason*. Oxford University Press.

Tieszen, R. (2011). *After Godel: Platonism and Rationalism in Mathematics and Logic*. Oxford University Press.

Turing, A. (1939). "Systems of Logic Based on Ordinals". In: *Proceedings of the London Mathematical Society (2)* 45, pp. 161–228.

Wang, H. (1974). *From Mathematics to Philosophy*. New York: Humanities Press.

— (1987). *Reflections on Kurt Gödel*. Cambridge, MA: MIT Press.

— (1996). *A Logical Journey: From Gödel to Philosophy*. Cambridge, MA: MIT Press.

Welch, P. (2017). "Obtaining Woodin's Cardinals". In: Caicedo, A. et al. (2017), pp. 161–176.

Wilder, R. (1965). *Introduction to the Foundations of Mathematics* (2nd ed.) New York: Wiley.

Wright, C. (1985). "Skolem and the Skeptic". In: *Proceedings of the Aristotelian Society, Supplementary Volumes* 59, pp. 117–137.

— (1994). "About 'The Philosophical Significance of Gödel's Theorem': Some Issues". In: McGuinness, B. and G. Oliveri (1994), pp. 167–202.

Wrigley, W. (2018). "Sider's Ontologese Introduction Instructions". In: *Theoria* 84, pp. 295–308.

Zermelo, E. (1930). "On Boundary Numbers and Domains of Sets". In: Ewald, W. (2005), pp. 1219–1233.