Gödel's Disjunctive Argument

Wesley Wrigley*

Draft of March 2020

Abstract

According to Gödel's disjunctive argument, the incompleteness theorems entail that the mind is not a machine, or that certain arithmetical propositions are absolutely undecidable. Gödel's view was that the mind is not a machine, and that no arithmetical propositions are absolutely undecidable. I argue that the Gödelian position in both cases rests on the assumption that the idealised mathematician can execute a certain non-recursive procedure. I identify Gödel's hypothesised ability as one variety of the *recursive ordinal recognition ability*. I show that we have this ability if, and only if, there are no absolutely undecidable arithmetical propositions. These considerations are developed into an argument for the existence of absolutely undecidable arithmetical propositions. I argue that no recognizable example of such a proposition could be identified, in principle. This implies a certain form of quietism about the limits of our arithmetical knowledge.

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 propositions which are in some sense *absolutely* undecidable (1951, p.310).

This paper has two aims. The first is to give a thorough assessment of Gödel's response to his disjunction. Given the scarcity of source material on the subject, this will involve some degree of reconstruction from his published

*Wadham College and Faculty of Philosophy, University of Oxford

and reported views on anti-mechanism and absolute undecidability. I'll argue that the Gödelian view in both cases rests on the presupposition that we can execute (at least "in principle") a certain non-recursive procedure. The second aim of this paper is to develop the evaluation of Gödel's presupposition into a novel argument for the existence of absolutely undecidable arithmetical propositions.

§§1–2 concern the absolute undecidability disjunct. I present Gödel's *evidence argument*, his most compelling case against the absolute undecidability of problematic arithmetical propositions, notably the consistency sentences of certain extensions of Peano arithmetic. I'll argue that the success of the argument depends on our ability to enumerate a non-recursive set of numbers which codes information about the ordinals. Drawing on Feferman's completeness theorem, I'll further argue that if Gödel's argument is successful, then it establishes the decidability of not only consistency sentences and the like, but of every arithmetical proposition whatsoever.

§§3–4 concern the anti-mechanical disjunct. Gödel himself wrote frustratingly little on the subject, but drawing on the scant remarks available, and the development of the view by Lucas, I present a view called *Gödelian anti-mechanism.* Once again deploying considerations relating to Feferman's completeness theorem, I argue that any remotely plausible version of this view rests on exactly the same presupposition as the evidence argument, namely that some particular non-recursive set can be enumerated by the idealised mathematician.

This gives us a single issue in terms of which we can address the disjunctive argument: does the idealised mathematician have the abilities required to vindicate Gödel's views? In §5, I identify Gödel's hypothesised ability as one kind of *recursive ordinal recognition ability*, and identify a weaker ability that would also serve some of Gödel's purposes. I then argue for a new disjunction: either we have the recursive ordinal recognition ability (in one of the two senses), or there are absolutely undecidable arithmetical propositions. I also briefly review the literature directly relevant to this ability, and find the existing arguments against our possession of this ability inadequate.

In §6–7 I consider two arguments in favour of our possession of the recursive ordinal recognition ability. One represents Gödel's *rationalistic optimism*, and the other is a novel argument in favour of our having the weaker recursive ordinal recognition ability. In each case, I argue that it would be an epistemic miracle if we had the proposed ability. On the other hand, a compelling case can be made *against* our possession of the recursive ordinal recognition ability, from which the existence of absolutely undecidable arithmetical propositions follows.

In §8, I discuss the prospects for exhibiting one of these propositions. I argue that, in principle, there is no recognizable example of such a proposition, despite the compelling case that can be made for their existence. This serves two purposes. Firstly, it excuses my own lack of a witness to the claim that such propositions exist, and secondly it suggests that a form of quietism about the limits of our arithmetical knowledge is the proper philosophical response to Gödel's theorems.

1 Gödel's Evidence 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 propositions are undecidable in an absolute sense. According to Gödel, '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). Hence, I'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] In considering whether certain propositions are absolutely undecidable, we are not considering whether *actual* human beings can produce a proof of them. After all, actual humans have a limited attention span, very limited life spans, and pressing needs beyond the production of arithmetical proofs. Rather, we consider what could be accomplished by the *idealised* mathematician. It is a little unclear *how much* idealisation is allowed in the debate by the various parties; but 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.

In the Gibbs lecture, Gödel does not offer an argument for a particular disjunct. His first (and best known) argument for the absolute decidability of all arithmetical propositions appears in handwritten notes from the 1930s (Gödel 193?). The argument there is terrifically condensed, so I'll try to spell it out with a little more clarity. Gödel identifies a special class of polynomial expressions that give rise to unsolvable *Diophantine problems.* The Diophantine problem corresponding to such a polynomial expression is to determine whether the equation $P(a_1, ..., a_m, x_1, ..., x_n) = 0$ has any solutions in the

[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 $\{s \mid \phi \vdash s\}$ as sound, and hence ϕ would not be absolutely undecidable according to the definition given.

integers for arbitrary integer values of the parameters (the a_is). As a consequence of the incompleteness theorems, each recursively axiomatized theory which can express all Diophantine problems of this kind and is sound with respect to them is incomplete with respect to them (Gödel 193?, p.165).[2]

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 problems of this kind which are unsolvable in an absolute sense. Without loss of argumentative force, I'll recast Gödel's argument in terms of the undecidability 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.[3] While not all sentences which entail the consistency of **PA** will be recognizable as such, a canonical consistency sentence like $Con_{\mathbf{PA}}$ 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. To use Gödel's terminology, the undecidable sentences associated with 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 associated with the incomplete-

[2] Strictly speaking, the disjunction presented by Gödel in the Gibbs lecture is that the mind is not a machine or that certain Diophantine problems of this kind are absolutely unsolvable. However, it is harmless to recast Gödel's disjunction in its more familiar form as between anti-mechanism on the one hand and the existence of absolutely undecidable arithmetical propositions on the other, because the Matiyasevich-Davis-Robinson-Putnam theorem implies that for each sufficiently expressive recursively axiomatized arithmetical theory **T**, there is a true (and constructible) *Diophantine sentence* $D_{\mathbf{T}}$ undecided by the theory, which says that a certain Diophantine equation has no solutions. Hence the corresponding Diophantine problem is unsolvable just in case the Diophantine sentence is undecidable.

[3] 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.

ness theorems more generally, are not absolutely undecidable propositions.[4]

In a nutshell then, the evidence argument is 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. Something like this argument must be correct. Indeed, in other places, e.g. (Gödel 1946, p.151), Gödel takes some version of the argument to be essentially truistic. 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 associated with the incompleteness theorems. The question, then, is do we have a proof in some-or-another system of *all* such propositions?[5]

It's crucial to note that the evidence argument relies on an *intensional* relation between two theories, namely that the axioms of one are exactly as evident as those of another. In claiming that the relation *the axioms of* ϕ *are exactly as evident as the axioms of* ψ is intensional, all I mean is that if (a) **T** is a sound theory, (b) the consistency of **T** is sufficient for the truth of P, and (c) there is no proof or other rationally compelling reason to believe (b), then it *does not* follow that the axioms of $\mathbf{T} + P$ are exactly as evident as the axioms of **T**.

Clearly, if P is *recognizable* as a consistency sentence for **T**, then the axioms of $\mathbf{T} + P$ are exactly as evident as those of **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 **T**. In some cases there might be *independent* compelling reasons to believe P; but Gödel's argument is supposed to apply to any extension of a knowably sound theory by its canonical consistency sentence, 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.* 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.

It might be objected that thinking of the relation *the axioms of* ϕ *are exactly as evident as the axioms of* ψ in this way forces us to adopt an *inter-*

[4] 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.

[5] The argument could similarly be run using Gödel sentences. In this formulation, the procedure for constructing the Gödel sentence of a theory known to be sound makes it obvious that it is true. The same goes for the Diophantine sentences discussed above, 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.

nalist 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 we're considering the relation under an already high degree of idealisation. So if the consistency of $\mathbf{T}$ is sufficient for the truth of P, but there is no proof or other rationally compelling reason to believe this, it means 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.

So, given the idealisation here, any assertion that the axioms of $\mathbf{T} + P$ are exactly as evident as those of $\mathbf{T}$ is essentially a demand to change the subject, and stop thinking about what propositions creatures like us could possibly prove. Crucially, thinking that the relation *the axioms of* ϕ *are exactly as evident as the axioms of* ψ is intensional does not beg the question against the Gödelian, since thinking of the relation in this way does not involve asserting that any situation in which (a), (b), and (c) jointly hold does, or even could, obtain.

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 $\mathbf{T}$' all I mean is a sentence in the language of $\mathbf{T}$ which expresses that $\mathbf{T}$ 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. It is, of course, standard practice to give a canonical form for consistency sentences to take. For example we may stipulate that $Con_{\mathbf{T}} =_{df} \neg\exists m \; Prf_{\mathbf{T}}(m, \ulcorner 0 = 1 \urcorner)$, for any $\mathbf{T}$. But that definitional schema is not a functor into which the axioms of $\mathbf{T}$ 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 $\mathbf{T}$ are changed, even if this makes no difference whatsoever to the theory considered as a set of sentences. Consistency sentences, and related constructions which code information about axiomatic theories, are in general are sensitive to the particular axiomatic presentation of a theory. In many 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 consistency sentences to extensionally equivalent theories does *not*, however, preserve extensional equivalence in this way. This is because such sentences 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, **A** and **B**, both of which extend **PA**. Suppose further that **A** and **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. In such a case, the relation $Prf_{\mathbf{A}}$ and $Prf_{\mathbf{B}}$ might appear so different that the consistency sentences $Con_{\mathbf{A}}$ and $Con_{\mathbf{B}}$ 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 **T** is extensionally equivalent to **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 **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 **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 **T**, that $\mathbf{T} + Con_{\mathbf{T}}$ is indeed an extension of **T** by the addition of a canonical consistency sentence for **T**. As we shall see, this is an assumption to which Gödel is certainly not entitled.[6]

[6] 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 paper 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 large cardinal axioms *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.

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, **PA** is sound. Since we know this, we have just as much cause to believe in the soundness of $\mathbf{PA} + Con_{\mathbf{PA}}$. The chain of such theories (starting with **PA**) 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 **PA** is should also convince us that $\mathbf{T}_{n+1}$ is sound. Since $\mathbf{T}_{n+1} \vdash Con_{\mathbf{T}_n}$, we have, for any n an argument that $Con_{\mathbf{T}_n}$ isn't absolutely undecidable.

2 Reflection Principles and Ordinal Notations

Gödel's evidence argument is intimately related to the theory of *reflection principles*. 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. A *reflection sequence* based upon a theory **T** is the result of the iterated addition of some reflection principle to **T**. In the case of Gödel's argument, the relevant reflection principle (for successor ordinals) is

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

Essentially, the evidence argument is that, where $\mathbf{T}_0 = \mathbf{PA}$, the axioms of $\mathbf{T}_\alpha$ are exactly as evident as the axioms of **PA**, no matter the ordinal value of α.

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 sequence 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 the theories in our consistency reflection sequence 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 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 to code information about these *recursive ordinals* in the language of arithmetic. In keeping with the majority of work in the field, we'll use Kleene's $\mathcal{O}$ notation system, though the arguments of this paper will carry over just as well to an equivalent system. In Kleene's 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.

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]

[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

Crucially, 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 and the property of being a member of $\mathcal{O}$ isn't definable by any formula in the language of arithmetic. This is a big problem for Gödel's argument because, unlike in the finite case, we should not suppose that, where α is transfinite, we can effectively recognize that the α^{th} theory in a consistency reflection sequence is actually an extension of **PA** by transfinitely iterated consistency reflection. In consequence, we currently have no reason suppose that the axioms of such a theory really are exactly as evident as those of **PA**.[8]

To spell this out in a little more detail: let

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

here, 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 α^{th} theory in our reflection sequence as the α^{th} extension of **PA** by iterated consistency reflection if α 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 the ordinal index α is denoted in the arithmetical presentation of the theory $\mathbf{T}_\alpha$ by a notation in $\mathcal{O}$ of a recursive ordinal.

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

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] 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 Gödel's arguments break down.

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 relies crucially on the assumption that we can, in principle, always recognize a canonical consistency sentence of a suitable extension of **PA** for what it is. This is a substantive assumption: once we've iterated the addition of consistency sentences into the transfinite, it amounts to the claim that the idealised mathematician can always determine whether a natural number denotes an ordinal in $\mathcal{O}$ (or an equivalent coding system). 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.

Thorough evaluation of Gödel's assumption will have to wait until §§5-7. For now, it's important to note that Gödel's argument is far from truistic. Not only does the argument rest on a contentious assumption, significantly more follows from it (if it is successful) that the decidability of propositions of particular classes, such as consistency sentences, Gödel sentences, and Diophantine sentences. Somewhat surprisingly, if the axioms of any member of a reflection sequence based on **PA** really are exactly as evident as the axioms of **PA** itself, then no arithmetical proposition whatsoever is absolutely undecidable. This follows from a result due to Feferman, and will be of great importance to the evaluation of Gödel's disjunction. Stating the result, however, requires some technical preliminaries.

Consistency reflection is not the only principle that can be iteratively added to **PA** on the basis of its soundness. We can also formulate principles which express not just that the preceding theory in a reflection sequence is consistent, but also that it is sound. One such principle is *Feferman reflection* (a standard version of the *Uniform Reflection Principle*), which says that, for any formula ϕ, if, of every number, a theory proves that it is ϕ, then every number is ϕ.[9] 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

[9] 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.

formulae of $\mathcal{L}_{\mathbf{PA}}$ with a single variable (Feferman 1962, p.274). A Feferman reflection sequence can then be defined over the recursive ordinals as follows:

> $\mathbf{T}_0$ is the base theory (for example $\mathbf{PA}$)
>
> For successors, $\mathbf{T}_{\alpha+1} = \mathbf{T}_\alpha \cup \{\forall x\ Pr_{\mathbf{T}_\alpha}(\overline{\phi}(\dot{\overline{x}})) \to \forall x\ \phi(x)|\ \phi \in \mathcal{L}_{\mathbf{PA}}\}$
>
> If β, is a limit ordinal, the axioms are chosen such that $\mathbf{T}_\beta$ extends $\mathbf{T}_\alpha$ by Feferman reflection for each $\alpha < \beta$ (including 0).

As with consistency reflection, the members of any such 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, we need a means of representing transfinite ordinals arithmetically. As in Feferman's original work on the subject, we'll use Kleene's $\mathcal{O}$ to do so. A transfinite recursive *progression* on $\mathbf{T}$, 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).[10]

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 $\forall x\ \phi(x)$. So if α is a successor ordinal then we should accept that $\mathbf{T}_\alpha$ is sound if we believe that its predecessor in the sequence is sound. This is because $\mathbf{T}_\alpha$ is an extension of $\mathbf{T}_{\alpha-1}$ by Feferman reflection for each ϕ, which is soundness-preserving with respect to the domain. If β is a limit ordinal, 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 using ordinal notations. 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 is acceptable, providing 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.[11]

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

[11] 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.

So any Gödelian who believes that any extension of **PA** by iterated consistency reflection has axioms which are just as evident as those of **PA** should feel similarly about members of any reflection sequence based on **PA** using Feferman reflection. The soundness of **PA** justifies the iterated addition of Feferman reflection just as much as consistency reflection. Hence, our Gödelian thinks that the axioms of the α^{th} member of any Feferman reflection sequence based on **PA** are exactly as evident as the axioms of **PA**. Surprisingly, if every member of each such sequence really does have axioms which are exactly as evident as those of **PA**, then it follows that no propositions are absolutely undecidable. The central mathematical result behind this conclusion 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 uniform reflection progression, some particular path b in $\mathcal{O}$ is such that for any true arithmetical ϕ, there is some n in b such that $\mathbf{T}_n \vdash \phi$ and $|n| < \omega^{\omega^{\omega+1}}$ (Franzén 2004, §14.3), meaning that a reflection sequence up to a small ordinal proves every arithmetical truth. Indeed, the bound can be reduced to ω^{ω^2+1} (Franzén 2004a, p.386).

This result is certainly a striking one. Moreover, it provides for a significant strengthening of Gödel's evidence argument. The original version of that argument was aimed at showing that none of the metamathematical propositions generated by the incompleteness theorems were absolutely undecidable. The success of that argument hinged on our ability to recognize, or to enumerate, notations for recursive ordinals. Feferman's theorem shows us that if we actually have this ability, the idealised mathematician could recognize that an arithmetically complete theory is sound by recognizing that it is constructed from **PA** by the iterated addition of a soundness reflection principle. Since such a mathematician would be equipped with a recognizably sound and arithmetically complete theory, *no* arithmetical proposition whatsoever would be absolutely undecidable.

3 Anti-Mechanism

We've seen that the hypothesised ability to recognize or enumerate notations for transfinite ordinals is the crucial moving part of Gödel's argument against the existence of absolutely undecidable arithmetical propositions. In the next two sections, I will argue that this hypothesised ability is *also* at the heart of Gödelian anti-mechanism, and hence an assessment of whether or not the idealised mathematician has this ability is critical for our evaluation of the disjunctive argument as a whole.

For our purposes, mechanism is the claim that the idealised arithmetical output of the human mind is coextensive with the output of a Turing machine. So we can take anti-mechanism to be the claim that the idealised mathematician can prove more than can be proved by any Turing machine. With a suitable system of Gödel numbering fixed, there is a one-to-one correspondence between (the outputs of) Turing machines and (the deductive closures of) recursively enumerable theories. Hence, anti-mechanism (in the present context) amounts to the view that the idealised mathematician can deploy a non-recursive procedure in the course of producing their arithmetical output.

Although he was an anti-mechanist, Gödel's support for the view is only cautiously hinted at in the Gibbs lecture; it was later confirmed as his own view in (Wang 1974, p.324-326). But the most well-known argument for the anti-mechanist disjunct is the Lucas–Penrose argument, which I'll breifly review.[12]

Lucas' argument against mechanism 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 mechanist (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 $\mathbf{M}$', 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 $\mathbf{M}$ enumerates the Gödel numbers of the theorems of $\mathbf{PA}$, and perhaps other things too. Let $\mathbf{T_M}$ be the recursive theory the theorems of which are coded by $\mathbf{M}$'s output. Thanks to Gödel's theorem, there is some sentence $G_{\mathbf{M}}$ which, provided that $\mathbf{T_M}$ is consistent, is true and which $\mathbf{M}$ cannot prove (i.e. $\mathbf{T_M} \nvdash G_{\mathbf{M}}$). Lucas then

[12] For the most part, I'll set aside Penrose's version. This is 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' (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.

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 $\mathbf{T_M}$. Assuming that (the theory corresponding to) the machine is consistent, 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).

Good (1967, p.146) first emphasised that a finite machine can be constructed corresponding to $\mathbf{PA}_\omega$, via the use of notations for recursive ordinals.[13] This means 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 crucial issue then, is whether we have any reason to believe that Lucas can keep beating the mechanist by applying reflection principles at transfinite stages of this process. Lucas certainly believes that he can. Before we assess this claim, it's important to note that if Lucas can always one-up the mechanist's proposed machine, then he can beat the mechanist overall. The soundness of **PA** implies not only the truth of its Gödel sentence, but also the relevant instances of Feferman reflection. Moreover, the construction of Good's machine for $\mathbf{PA}_\omega$ is still possible when the reflection principle of concern is Feferman reflection, rather than the iterated addition of Gödel sentences. So whatever Lucas could do in the original game against the mechanist, he can do in the higher-stakes version played using Feferman reflection. Thanks to Feferman's theorem, we know that there exists a branch b in $\mathcal{O}$ such that $\bigcup_{n \in b : |n| < \omega^{\omega^2+1}} \mathbf{T}_n$ is complete. Call any arithmetically complete theory constructed by transfinite reflection on **PA** a *Feferman arithmetic.* If Lucas can one-up the mechanist at every stage of the game played on some such b, he will *certainly* win; after sufficiently many moves, Lucas will have produced a Feferman arithmetic, which none of the mechanist's machines will be able to do.

If Lucas can beat the machine at every stage on some suitable b, then he has the ability to recognize that all of the theories corresponding the the mechanist's machines are indexed by notations for recursive ordinals. After

[13] Here $\mathbf{PA}_\omega$ is the first member with a transfinite index of a reflection sequence based on **PA** where the construction proceeds by the iterated addition of Gödel sentences.

all, it is in recognizing that the index of the theory codes such an ordinal that Lucas would recognize that it extends **PA** by the iterated addition of a soundness-preserving principle, and hence this recognition is where Lucas would find his justification for applying Feferman reflection and thereby proving something that the machine cannot.

If Lucas can enumerate $\mathcal{O}$, then he can certainly do this. And supposing that the idealised mathematician has this ability would not, I think, make the account any less Gödelian. Gödel himself certainly thought that we possess certain non-recurisive abilities. Although he acknowledges that the notion of a non-recursive procedure is far from clear, 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 Gödel thought that we can, in principle, enumerate ω_1^{CK} (the set of the order types of the recursive well-orderings of the integers), it is consonant with his view to suppose that we have the intimately related ability to enumerate $\mathcal{O}$, especially since his evidence argument depends on this supposition. If we can non-recursively enumerate $\mathcal{O}$, then we could construct a Feferman arithmetic; correspondingly, I'll call the view that the idealised mathematician can axiomatize a Feferman arithmetic and deploy it in their production of proofs *Gödelian anti-mechanism.*[14]

Before we come to the issue of our hypothesised ability to enumerate $\mathcal{O}$, it is instructive to consider whether the Gödelian anti-mechanist can beat the mechanist using fewer resources. Lucas does offer an argument to the effect that the anti-mechanist needn't presuppose that the idealised mathematician can enumerate $\mathcal{O}$. 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 Turing machine the corresponding theory of which is indexed by such a notation. According to Lucas, this means that whatever

[14] One might object that I am unfairly ascribing to Gödel the view that the idealised mathematician can carry out a transfinite number of reasoning steps, since this appears to be required for the construction of a Feferman arithmetic. I'm not sure if a transfinite number of steps of reasoning really is required to execute such a construction; the issue strikes me as decidedly murky. But in any case, 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 my proposal still represents a reasonable reconstruction of Gödel's position from the meagre textual evidence that is available, even if it does involve a contentious view about the abilities of the idealised mathematician.

machine was presented by the mechanist cannot have an arithmetical output coextensive with his own (Lucas 1996, pp.111-112).

This argument, however, is unconvincing. Even if it is a requirement of the dialectical scenario that the mechanist put forward as a proposed model of the mind a machine the corresponding theory of which is *actually* an extension of **PA** by the transfinite iterated addition of a reflection principle, 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-Gödelize' it (to borrow his term) if he can determine the theory corresponding to the machine, verify that it is indexed by an ordinal notation, and then apply a reflection principle 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 recognize ordinal notations when presented with them, which comes to the same as the ability to enumerate $\mathcal{O}$.

So much for Lucas' argument. The failure of that argument does not, however, show that the anti-mechanist really must presuppose that we have the ability to enumerate $\mathcal{O}$, as I've claimed. To see that this is so, we need to examine one further aspect of the kinds of reflection progression that we've been considering; this reveals that, without the presupposition that they can enumerate $\mathcal{O}$, the anti-mechanist is left with a *disastrous* epistemology of arithmetic.

4 The Failure of Autonomy

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 $a \leq_{\mathcal{O}} d$ 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 $\omega^{\omega^{\omega}}$ 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 without the presupposition of the idealised mathematician's ability to enumerate $\mathcal{O}$ (or perhaps just the ability to enumerate some branch of it suitable for the construction of a Feferman arithmetic, though this would be rather *ad hoc*).

Consider the theory **TA**, or 'true arithmetic'. This theory is axiomatized by every true sentence of the language of arithmetic, i.e. $\mathbf{TA} = \{\phi \mid \phi \in \mathcal{L}_{\mathbf{PA}} \wedge \phi \text{ is true}\}$. **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 the theory as presented presupposes the notion of arithmetical truth, 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. But since there is no recursive procedure for determining of an arbitrary formula in the language of arithmetic whether it is true, we cannot straightforwardly determine what the axioms of **TA** *are*, given that those are just the truths of arithmetic.

The ineliminable use of non-autonomous methods means that any Feferman arithmetic we might construct is defective in the same way as **TA**. 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 ability to construct a Feferman arithmetic goes hand-in-hand with the idealised mathematician having the ability to enumerate $\mathcal{O}$, rather than the weaker ability to simply follow a path within $\mathcal{O}$ up to ω^{ω^2+1}. After all, can't we just start with **PA**, add the Feferman reflection principle at successor stages of the sequence, and take the union of our previous theories at limits? If we need only the ability to *follow a path* within $\mathcal{O}$, 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 ω^{th} 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 true Π_1 sentences (Feferman and Spector 1962, p.384). Moreover, there are $\aleph_0$ such paths (1962, p.389).

The Gödelian anti-mechanist claims that, in principle, we can axiomatize and make use of a Feferman arithmetic. The Feferman–Spector theorem shows that to do so, 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 new 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 without knowing in advance whether a 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 our inability to construct a Feferman arithmetic using only resources internal to our reflection sequences, when we attempt to construct such a sequence we find ourselves in the following situation: for certain problematic sentences ϕ in the language of arithmetic, some theory in the reflection sequence is sound just in case ϕ is true. If ϕ really is true then we 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. This situation prompted Turing to claim that his Π_1-completeness theorem was 'of no value'.[15] As he 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

[15] This result is similar to Feferman's theorem, but is restricted to Π_1 sentences and is obtained with a lower ordinal bound. The progressions considered by Turing are also based only on consistency reflection, rather than Feferman's stronger principle. Proving Turing's result similarly makes ineliminable use of non-autonomous methods. This shows that the arguments of the present section apply also to significantly weaker forms of anti-mechanism which only credit the idealised mathematician with the ability to determine the truth of all Π_1 sentences. Since Penrose restricts his Gödelian argument to such sentences (Penrose 1994, p.96), this is of some significance.

a certain [number] as an ordinal [notation]' (1939, §9).

Of course, if we have the power to enumerate $\mathcal{O}$, then we needn't *assume* that some number is a notation; we could simply check this directly. So the Gödelian is better off supposing that the idealised mathematician has the ability to enumerate $\mathcal{O}$ in order to effect the construction of a Feferman arithmetic. Without this presupposition, the anti-mechanist is forced to credit the idealised mathematician with the ability to prove any arithmetical truth by deploying a theory the axiomatization of which relies on inexplicable prior knowledge of (certain of) its theorems. I take it that this is a position which no serious epistemology of arithmetic can tolerate.[16] So although there is some motivation for thinking that we can apply reflection principles justifiably in something like the manner Lucas has in mind (because, for example, we know that **PA** is sound), the general success of anti-mechanism hinges on exactly the same assumption as Gödel's evidence argument, namely the assumption that the idealised mathematician can enumerate $\mathcal{O}$.

5 A New Disjunction

Surprisingly then, the key to the entire disjunctive argument seems to be our ability to enumerate notations for recursive ordinals in the natural numbers. Gödel's evidence argument succeeds only if we have the ability to enumerate $\mathcal{O}$; if we cannot enumerate $\mathcal{O}$, then we don't have the ability to determine whether an arbitrary number denotes an ordinal, and hence we have no guarantee that the axioms of any extension of **PA** by iterated reflection are exactly as evident as those of **PA**. On the other hand, the Gödelian anti-mechanist must also think that we have the ability to enumerate $\mathcal{O}$; otherwise the axiomatization of a Feferman arithmetic by the idealised mathematician would be an inexplicable miracle. So the evidence argument and anti-mechanism stand or fall together, and both rest on the presupposition that the idealised mathematician can enumerate $\mathcal{O}$.

This gives us a much clearer setting in which to make an evaluation of Gödel's view. Rather than consider the general nature of the human mind, or consider specific potential cases of arithmetical absolute undecidability, we can simply examine whether the idealised mathematician can enumerate $\mathcal{O}$. Call the ability to enumerate $\mathcal{O}$, under some acceptable idealisation of our

[16] 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 arithmetical axioms prove that $2 + 2 = 4$. 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**.

actual mathematical abilities, the *strong recursive ordinal recognition ability*. It will be helpful to identify a related, though weaker ability: say that we possess the *weak recursive ordinal recognition ability* just in case there is some branch b in $\mathcal{O}$ such that for any true arithmetical ϕ, there is some n in b such that $\mathbf{T}_n \vdash \phi$ and for each $\mathbf{T}_n$ such that $n \in b$, we can recognise that $\mathbf{T}_n$ is indexed by a notation for a recursive ordinal under some suitable idealisation of our mathematical abilities.

Every arithmetical truth is absolutely provable just in case we possess the recursive ordinal recognition ability in either the strong or the weak sense. If we have the ability in the strong sense, then the idealised mathematician can prove every arithmetical truth by constructing a Feferman arithmetic. If we have the ability merely in the weak sense, then every arithmetical truth can be proved in principle, regardless of whether the idealised mathematician can construct an arithmetically complete theory or not. For suppose that there is some branch b, with the desired completeness property, such that for each n in b we can recognize under some-or-another idealisation that $\mathbf{T}_n$ is an extension of **PA** by iterated Feferman reflection. Hence each such $\mathbf{T}_n$ can, in principle, be recognized as sound, and since each each arithmetical truth ϕ is provable from some such $\mathbf{T}_n$, each true ϕ is provable by us (under some suitable idealisation) using some recognizably sound theory.

If we have the ability in neither sense, then there are absolutely undecidable arithmetical propositions: suppose that each true ϕ is provable by us under some acceptable idealisation. Then for some b with the required completeness property, all members of each $\mathbf{T}_n$ such that $n \in b$ would be provable too. Hence, all such $\mathbf{T}_n$s would be recognizable as sound. Since these theories are sound only if their index denotes an ordinal, each theory would be recognizably indexed by a recursive ordinal under some suitable idealisation, and hence we would have the weak recursive ordinal recognition ability. Contraposing, if we lack the weak ability (and *a fortiori* the strong ability), there are absolutely undecidable arithmetical propositions.

Hence we can consider a new disjunction: either we possess the recursive ordinal recognition ability, or there are absolutely undecidable arithmetical propositions. The remainder of this paper will be devoted to arguing for the latter disjunct. A helpful place to start is with the (small) literature discussing whether or not the idealised mathematician can enumerate $\mathcal{O}$.

In early work on the subject, Turing highlighted 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 in support of his view that the human ability to recognise ordinal notations outstrips that of any machine (Lucas 1996, p.111). He also cites Gödel and Wang as rejecting mechanism *because* we can enumerate ordinal notations (1996, p.111). While the ascription of this view to Gödel seems fair, the ascription of it to Wang is somewhat suspect: 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). Regardless, the problem is that whether the idealised mathematician can perform some ingenious operation which no machine could ever do is precisely the point at issue in the anti-mechanism debate. 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.

More substantive discussion is given Shapiro (1998, p.289), who distinguishes between a weak and a strong version 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 $\mathcal{O}$.[17]

Shapiro claims that the weak version is hopelessly vague, since it involves 'machine enumerating ordinal notations' as a parameter. Although this is admittedly not precise, I think that we can make enough sense of it to see that the anti-mechanist must make a stronger claim. $\mathcal{O}$ is not recursively enumerable, but for any $n \in \mathcal{O}$, $\{m \mid m <_{\mathcal{O}} n\}$ *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. Hence the weak anti-mechanist thesis is too weak to distinguish humans from machines in the required fashion.

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. 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 somewhat misses the point.[18] Crucially, claiming that a human reasoner could, in principle, enumerate $\mathcal{O}$ is *not* to claim of anyone that they are arithmetically omniscient. It is rather to claim that every arithmetical truth is provable by the *idealised* reasoner,

[17] A similiar distinction was drawn with less detail by Good (1969, p.357).

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

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, for Gödel at least, presented as an alternative to the view that there are absolutely undecidable arithmetical propositions.

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 (2004, p.191) has proved that there is a unary primitive recursive function f from sentences in $\mathcal{L}_{\mathbf{PA}}$ to natural numbers such that ϕ is true iff $f(\phi) \in \mathcal{O}$. Hence if we had a good reason to think that we could enumerate $\mathcal{O}$, that ability *could* be used to explain why all arithmetical propositions are provable: for any ϕ, we could determine it's truth by applying f and checking the output against our enumeration.

Shapiro's criticisms miss their mark, but this doesn't change the fact that we still have no reason yet to believe that we have the recursive ordinal recognition ability. In the next two sections, I will argue that there is no good reason to believe that we possess the ability in either the strong or the weak sense. I don't have an argument that the idea that we have the recursive ordinal recognition ability 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 do have the ability, even in principle. For those of us who can't take our possession of such an ability on faith, the existence of an absolutely undecidable arithmetical proposition is made enormously plausible by the evidence to be presented.

6 Rationalistic Optimism

I remarked above that Gödel never offered an argument for anti-mechanism like Lucas and Penrose. One reason is that he took anti-mechanism to be a consequence (via disjunctive syllogism) of *rationalistic optimism*, the principle that no well-defined mathematical problem is unsolvable in an 'absolute' sense (Shapiro 2016, p.191). A consequence of this view is that we have the recursive ordinal recognition ability in the strong sense, but in this section I'll argue against the idea that we have any good reason to follow Gödel on this point.

Rationalistic optimism represents Gödel's mature view on the decidability of arithmetical propositions. Narrowly speaking, it is the view that any well-posed arithmetical proposition can be proved or can be refuted (Wang 1974, p.324-326). More broadly, it is the view that 'for clear questions posed by reason, reason can find clear answers' (Gödel 1961/?, p.318). 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.[19] 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 a paper from the 1940s 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.

Since rationalistic optimism credits us with the ability to prove or refute any well-defined mathematical proposition, if it is correct then we have the strong recursive ordinal recognition ability almost trivially. After all, for any n, the question '$n \in \mathcal{O}$?' is meaningful. Assuming rationalistic optimism then, we should be able to enumerate $\mathcal{O}$ by enumerating $\mathbb{N}$ and removing those numbers to which the answer to '$n \in \mathcal{O}$?' is 'no'.

However, rationalistic optimism on its own cannot provide an *argument* for our possession of the strong recursive ordinal recognition ability. This is because the 'explanation' of our ability to recognise ordinal notations proceeds in terms of a *prima facie* more contentious principle, namely that *any* mathematical problem is solvable. 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

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

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 premise is that all arithmetical propositions are expressible in a set-theoretic context. Any argument from rationalistic optimism to the recursive ordinal recognition ability will be more similar to this unconvincing argument than we ought to be comfortable with. If *any* problem is solvable, then *of course* any particular problem is solvable, including that of enumerating $\mathcal{O}$.

The situation would be different if we had some *independent* reason to be rationalistic optimists. Gödel never published an argument for the position, though he did communicate to Hao Wang a somewhat crypic argument which is of some relevance. With respect to the hypothesis that there exist absolutely undecidable 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).[20] Let's call this the 'irrationality argument'.

I hope we can all agree that Gödel's meaning here is difficult to discern. There are (at least) two readings of this argument that bear on the absolute undecidability debate. The more modest reading pertains specifically to potentially undecidable propositions of the kinds we've already considered: consistency sentences, Gödel sentences, etc. On this reading, Gödel's argument is that there is some kind of irrationality involved in thinking that we could 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 or Gödel sentence for those axioms by means of mathematical reasoning. This is a compelling 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 quite similar to the evidence argument. I'll return to this version of the argument in §8. For now, it of no concern to us because, understood as an argument pertaining specifically to arithmetical propositions, it cannot lend any significant support to the optimist's strategy of inferring the decidability of all arithmetical propositions from a more general principle about the epistemology of mathematics as a whole.

According to a stronger, less modest, reading of the argument, Gödel might be claiming that the mere existence of an absolutely undecidable math-

[20] 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.

ematical proposition of any kind is sufficient for some kind of inconsistency in human reasoning. Such an argument must 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 clearly posed question as a symptom of either affliction. What the argument requires is an account of the nature of mathematics that would make it irrational to be unable to answer a clearly-posed question. Unfortunately, Gödel does not offer such an account, and it is difficult to see how he could in light of his platonism. Presumably, a realist about mathematics could no more complain that human reason would be inconsistent if unable to answer a mathematical question than they could so complain if human reason were unable to answer a biological or chemical question.

So, in the end, rationalistic optimism is of no argumentative support for the Gödelian in establishing the idealised mathematician's possession of the strong recursive ordinal recognition ability. Despite Gödel's insistence that its failure would constitute some kind of scandal to human reason, rationalistic optimism really is *just* optimism, and it is hard to see how anything other than faith could compel us to believe its truth. So Gödel's position, though coherent, really has little to recommend it. We've been given no serious philosophical or mathematical reason to think that we can, even in principle, enumerate $\mathcal{O}$.

7 Between Mechanism and Optimism

Recall that we have the *weak* recursive ordinal recognition ability just in case there is some branch b in $\mathcal{O}$ such that for any true arithmetical ϕ, there is some n in b such that $\mathbf{T}_n \vdash \phi$ and for each $\mathbf{T}_n$ such that $n \in b$, we can recognise that $\mathbf{T}_n$ is indexed by a notation in $\mathcal{O}$ under some suitable idealisation of our mathematical abilities. In this section, I want to discuss a position which I'll call the *intermediate view*, which attempts to motivate the idea that we have the weak recursive ordinal recognition ability without presupposing rationalistic optimism.

The advocate of the intermediate position does *not*, unlike the Gödelian, assert that given time and paper, the idealised mind can execute a non-recursive procedure. Furthermore, the intermediate position denies that there is some absolutely undecidable arithmetical proposition which expresses the consistency of our idealised arithmetical output. Since 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, the view appears to lie somewhere between mechanism and rationalistic optimism.[21]

The intermediate position holds that our actual 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. Since we cannot, even in principle, execute a non-recursive procedure, we cannot enumerate $\mathcal{O}$. More than this, we cannot enumerate a path through $\mathcal{O}$, though presumably we can enumerate some paths *within* $\mathcal{O}$ that are non-maximal, in the sense of not assigning a notation to every recursive ordinal. Moreover, there are some arithmetical propositions which we cannot prove. But the intermediate position denies that any such proposition is *absolutely* unprovable.

According to the intermediate position, it is no accident that any path we can follow *within* $\mathcal{O}$ is not a path *through* $\mathcal{O}$, since since constructing a maximal path through $\mathcal{O}$ would require the execution of a non-recursive procedure. However, for any path within $\mathcal{O}$ that we can actually enumerate, we could have enumerated a longer path that includes it. This is because, according to the intermediate position, any limitation on our actual ability to recognize ordinal notations (other than limitations on executing non-recursive procedures) is somehow "accidental"; if we can actually follow a path within $\mathcal{O}$ up to, but not including n, then that isn't because n has some special property. If only we'd had a bit more time and paper, we could have extended this path further to some $m >_{\mathcal{O}} n$. Hence we can recognize m as a notation under a natural idealization of our current abilities. The claim has some intuitive appeal; after all it is difficult to see how there could be a path within $\mathcal{O}$ which we can enumerate, but which we could not extend even in principle. What relevant explanation could we possibly give of this state of affairs? Indeed, it seems that for any path which we can follow under some acceptable idealisation, we could have followed a slightly longer one under another admissible idealisation. According to the intermediate position, this gives us good reason to think that no undecidable proposition is absolutely so.

Here is the reasoning: suppose we accept the intermediate position's claim that our current arithmetical abilities have an output coextensive with some recursively enumerable set. Then there is some undecidable true proposition ϕ which is not a member of this set. Feferman's theorem shows us that there is some path b, such that for some n on b, $\mathbf{T}_n \vdash \phi$. We can enumerate at

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

least part of b, so under some acceptable idealisation of our abilities, we can enumerate enough of it to recognize that $\mathbf{T}_n$ is indexed by an ordinal notation, and hence recognize that the theory is sound and obtain a proof of ϕ. The same goes for any undecidable proposition, so it follows that there are no absolutely undecidable arithmetical propositions. If this is all correct, then we possess the weak recursive ordinal recognition ability, despite not possessing its strong counterpart.

The intermediate 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 recognition ability not because we can, under idealisation, recognize each of the notations required to construct an arithmetically complete theory, but because each of the required notations can, under some idealisation, be recognized as such by us.

Despite the intuitive appeal of such a position, it actually rests on optimism just as much as the original Gödelian position did. Suppose we grant that for any path which we can enumerate in principle, there is some longer path including it which we can also enumerate in principle. This 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 *for some path* through $\mathcal{O}$, any notation on that path is recognizable by us under some idealisation.

This would suggest that for certain paths through $\mathcal{O}$, every theory in a Feferman reflection progression on **PA** indexed by a member of these paths is recognizable as sound. But even if that is true, it does *not* entail that we have the recursive ordinal recognition ability, thanks to the Feferman–Spector theorem, which we encountered in §4:

> **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 true 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 intermediate position? Suppose we accept that our current abilities correspond to $\mathbf{T}_n$ and that for some path that includes n, any notation which lies on it is recognizable by us under some idealisation. All this means that for any ordinal less that ω_1^{CK}, we can recognize some notation for it. But the Feferman–Spector theorem shows that there are paths through $\mathcal{O}$ such that the theories in a Feferman reflection progression which are indexed to that path are collectively incomplete. So even if every notation on some path through $\mathcal{O}$ is recognizable by us on some suitable idealisation, this does not entail that we have the weak recursive ordinal recognition ability.

This is significant because it might seem plausible that we *can* provide a notation for each recursive ordinal if we idealise enough. For instance, while he doesn't endorse the intermediate position, Penrose claims that where our reflection sequence proceeds via the iterated addition of Gödel sentences, we can construct a theory $\mathbf{T}_\alpha$ for any $\alpha < \omega_1^{CK}$ (1994, p.114). I have no argument against that suggestion. And we can certainly give a system of notations for all the ordinals $< \epsilon_0$ (which is far greater than the ordinal bound required to deploy Feferman's theorem) by exploiting Cantor's normal form theorem. But what the Feferman–Spector theorem shows is that this is perfectly consistent with the existence of absolutely undecidable propositions. Even if we have the ability to recognize a notation for each recursive ordinal in principle, this does not mean that have the capacity to recognize the *right* ordinal notations in principle.[22]

Of the $2^{\aleph_0}$ paths through $\mathcal{O}$, only $\aleph_0$ have the desired completeness property (Feferman and Spector 1962, p.389). Hence, the intermediate position is only correct if, by some cosmic chance, the notations which we can recognize under idealisation constitute one of these special paths. And there is simply no reason to suspect that this is the case. So the intermediate position, much like rationalistic optimism, is a bare conceptual possibility; there is no reason to think that either is true.

8 Absolutely Undecidable Propositions

We've seen that two parties to the absolute undecidability debate, namely Gödelianism optimism and the intermediate 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 unwarranted by the evidence. It would simply be a miracle if we have the recursive ordinal recognition ability, even if we only have it in the weak sense. If we are unwilling to countenance such a miracle, it follows that some arithmetical propositions are absolutely undecidable, namely those instances of Feferman reflection corresponding to theories indexed by numbers which we cannot

[22] Thanks to Daniel Isaacson for highlighting the importance of this point to me. See (Franzén 2004, §11.3 and §13.2) for the technical details of how to obtain a part of $\mathcal{O}$ which functions as a canonical system of *autonomous* notations for ordinals below ϵ_0.

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 truths 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, which we first encountered in §6.

I argued above that 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 known 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. If my case against the recursive ordinal recognition ability is sound, it follows that the axioms of some theories in our Feferman reflection progression on **PA** are *not* exactly as evident as the axioms of **PA**, and that instances of Feferman reflection for these theories are absolutely undecidable. But this fact means that neither I, nor anyone else, can exhibit an absolutely unprovable arithmetical truth of the kind under discussion.

Recall that, according the modest understanding of Gödel's irrationality argument, there is some kind of irrationality or inconsistency in the idea that some axioms of a member of a reflection sequence based on **PA** are true, but unprovable. After all, how could it be that we are presented with a theory known to be sound, and yet be unable to offer a proof that the theory is consistent, or that its Gödel sentence is true? Unlike the more straightforward understanding of the irrationality argument, this involves no presupposition of rationalistic optimism, and hence is of considerably broader interest.

But even on this more interesting reading, the argument cannot be counted as a success. If we lack the recursive ordinal recognition ability, then there is some true sentence in the language of arithmetic which is *absolutely* undecidable. But there cannot be a recognizable *example* of such a proposition that would give rise to the irrational scenario sketched above.[23] Suppose that we were presented with an instance of Feferman reflection, 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 Feferman reflection,

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

and assert something of that theory which follows from its soundness. 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 $\mathbf{T}_{2^n}$. And *this* theory trivially decides all instances of Feferman reflection for $\mathbf{T}_n$ 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 the undecidable sentence 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 of this kind that are true *and* absolutely undecidable can't be recognised for what they are.

The undecidability of such propositions therefore does not mean that we can be in the 'irrational' position of simultaneously having a theory known to be sound and not having a means of proving the relevant instances of Feferman reflection: if we can't prove them it's *because* we can't recognize that the theory 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.

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.

Conclusion

According to Gödel's favoured resolution of his disjunctive conclusion, there are, in principle, no absolutely unprovable number-theoretic truths and the mind cannot be modelled, in principle, by a Turing machine. I've argued in both cases that the Gödelian position rests on a critical assumption: that the idealised mathematician has the ability to (non-recursively) enumerate $\mathcal{O}$. Thanks to Feferman's theorem, a weaker ability would also suffice for the absolute decidability of all arithmetical propositions. I've argued that our possession of either ability would constitute an epistemic miracle in which we have no serious philosophical or mathematical reason to believe. From our lack of either version of this *recursive ordinal recognition ability*, the existence of some absolutely undecidable arithmetical proposition follows. Although we can identify the broad kind to which such propositions belong, we cannot give a recognizable example of an absolutely unprovable arithmetical truth of this kind, even in principle. Given this, we are left with a peculiar species of quietism about the limits of our arithmetical knowledge.

Benacerraf, in his discussion 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, our arithmetical capabilities can be perfectly mimicked by a Turing machine, but we don't have the ability to recognize the machine when presented with it (indeed, Smith claims that the ability to spot which machine enumerates my idealised output would be 'godlike' (2013, pp.281–282)). A similar view is sketched by Gödel himself in the Gibbs lecture (1951, pp.309–310), according to which the mind would be unable to fully understand itself. I'm happy to remain silent on whether this kind of mechanism is true. But regardless of the ultimate status of mechanism, a similar form of quietism is forced upon us by absolute undecidability: we can't precisely delimit our ability to recognise notations for recursive ordinals, because we can't give an example of an absolutely undecidable proposition. Moreover, this isn't 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 the Benacerraf–Smith view, we should at least acquiesce in its quietism.

There is a decent positive story to tell about why our idealised arithmetical 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 in this paper. 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.[24]

[24] I am very grateful to Tim Button for invaluable feedback on many different incarnations of this paper, and to Owen Griffiths, Daniel Isaacson, Alex Oliver, Michael Potter, and Robert Trueman for extensive comments which have greatly improved it. I'm also grateful to the participants of the Oxford Graduate Class in Philosophy of Mathematics of Trinity term 2020, including my co-host Joel David Hamkins, for helpful discussion of this paper. Thanks also to an anonymous referee from *The Review of Symolic Logic* for comments on the parts of this paper concerned with anti-mecahnism. This research was generously funded by an Arts and Humanities Research Council Studentship.

Bibliography

Benacerraf, P. (1967). “God, the Devil, and Gödel”. In: *The Monist* 51, 9–32.

Feferman, S. (1960). “Arithmetization of Metamathematics in a General Setting”. In: *Fundamenta Mathematicae* 49, 35–92.

— (1962). “Transfinite Recursive Progressions of Axiomatic Theories”. In: *Journal of Symbolic Logic* 3, 259–316.

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. (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, 383–390.

Franzén, T. (2004). *Inexhaustibility.* Wellesley, MA: Association for Symbolic Logic.

— (2004a). “Transfinite Progressions: A Second Look at Incompleteness”. In: *Bulletin of Symbolic Logic* 10, 367–389.

Gödel, K. (193?). “Undecidable diophantine propositions”. In: Feferman, S. et al. (1995), pp. 164–175.

— (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.

— (1951). “Some basic theorems on the foundations of mathematics and their implications”. In: Feferman, S. et al. (1995), pp. 304–323.

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.

Good, I. (1967). "Human and Machine Logic". In: *The British Journal for the Philosophy of Science* 18, 144–47.

— (1969). "Gödel's Theorem is a Red Herring". In: *The British Journal for the Philosophy of Science* 19, 357–358.

Horsten, L. and P. Welch, eds. (2016). *Gödel's Disjunction.* Oxford University Press.

Kant, I. (1787). *Critique of Pure Reason* (2nd ed.) Trans. Guyer, P. and A. Wood (1998). Cambridge University Press.

Koellner, P. (2010). "Absolute Undecidability". In: Feferman, S. et al. (2010), pp. 189–225.

Lucas, J. (1961). "Minds, Machines and Gödel". In: *Philosophy* 36, 112–127.

— (1968). "Human and Machine Logic: A Rejoinder". In: *The British Journal for the Philosophy of Science* 19, 155–156.

— (1996). "Minds, Machines and Gödel: A Retrospect". In: Millican, P. and A. Clark (1996), pp. 103–124.

Millican, P. and A. Clark, eds. (1996). *Machines and Thought: The Legacy of Alan Turing.* Oxford University Press.

Penrose, R. (1994). *Shadows of the Mind.* Oxford University Press.

Shapiro, S. (1998). "Incompleteness, Mechanism, and Optimism". In: *Bulletin of Symbolic Logic* 4, 273–302.

— (2016). "Idealization, Mechanism, and Knowability". In: Horsten, L. and P. Welch (2016), pp. 189–207.

Smith, P. (2013). *An Introduction to Gödel's Theorems* (2nd ed.) Cambridge 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, 161–228.

Wang, H. (1974). *From Mathematics to Philosophy*. New York: Humanities Press.

— (1996). *A Logical Journey: From Gödel to Philosophy*. Cambridge, MA: MIT Press.