Department of Philosophy
University of Bristol

PHILM0016 Philosophy of Mathematics

What is the significance of Gödel's theorems for Hilbert's Programme?

Daniel Wood
February 2010

Abstract

In this essay we shall examine the effects of Gödel's incompleteness theorems on Hilbert's Programme, focusing on the second theorem. We shall conclude that while Gödel's theorems are devastating for Hilbert's hope for a proof of consistency, they do not affect the underlying structure of Hilbert's Programme, the dichotomy between contentual (finitary) and ideal mathematics.[1] We shall also respond in detail to the claim by Detlefsen in [5] that the second theorem does not end hopes for a consistency proof.

1 Introduction

In §2 we shall outline Hilbert's Programme.[2] In §3.1 we shall state and sketch the proofs of the incompleteness theorems, and in §3.2 we shall demonstrate how they affect Hilbert's Programme. §4 will address the question of whether the arguments in §3.2 are in fact correct, and we shall conclude in the affirmative. In §4.1 we shall look at the precise nature of finitary mathematics and discuss whether the proofs of Gödel's theorems are in fact finitary, concluding that they indeed are; and in §4.2 we will refute the most sophisticated argument put forward against our thesis, namely that in [5]. §5 will see how we can deal with the aftermath of Gödel's theorems, in which we shall argue our main thesis. Finally, in §6 we shall draw our conclusions.

2 Hilbert's Programme

Hilbert's Programme was outlined in the early 20^{th} century by the great mathematician David Hilbert. It was an attempt to put mathematics onto solid foundations after the discovery of paradoxes in set theory, the most famous of which being Russell's Paradox.[3] Hilbert gave the most developed outline of his programme in [8], and we shall take this as our main guide for this section.

[1] This is not to say that I think that the underlying structure is a *good* philosophy of mathematics, but rather that Gödel's theorems *do not add* to its problems.

[2] We shall not discuss the various general philosophical aspects of Hilbert's Programme, except those that Gödel's theorems relate to directly. I will however, for the sake of the interested reader, try to highlight such general issues in passing, doing my best to suggest further reading.

[3] We shall not discuss the historical or philosophical motivations behind the development of Hilbert's Programme in detail. There are many good texts to be found that do go into these issues, some examples being [5], [6], chapter IV of [10], [11], [12], chapters 2 and 3 of [13], and chapter 6 of [15].

Let us outline the underlying structure of Hilbert's Programme. Hilbert divided mathematical statements into two classes: *contentual* and *ideal*.[4] We shall discuss them in turn.

A detailed discussion of the nature of (semantic) content is beyond the scope of this essay, but for our purposes the following rough characterisation is sufficient: A statement has content, i.e. is *contentual*, if it makes a statement about the world. As such, contentual mathematics is necessarily consistent (assuming that the world is consistent) and *bivalent*, i.e. every statement is either true or is false.[5] But what constitutes contentual mathematics? Well, in Hilbert's philosophy, contentual mathematics is precisely finitary mathematics. We shall discuss the nature of finitary mathematics in §4.1, but for the time being it can be thought of as arithmetic. Importantly, by our previous remark, finitary mathematics is thus necessarily consistent and bivalent.

We now come to ideal statements. Unlike contentual statements, ideal statements do not have content; they are used primarily for their instrumental value, such as mathematical simplicity or scientific applicability. For example, in analysis one introduces the number $i = \sqrt{-1}$ in order to deal with problems that cannot be solved with real numbers alone. According to a Hilbertian, however, i does not in fact exist, other than as a symbol, and is used simply for instrumental value. To quote Hilbert, we introduce i 'to preserve in simplest form the laws of algebra' ([8], p. 145); i is an 'ideal element', and statements in which it occurs are ideal.

So, in the dichotomy of finitary and ideal mathematics, ideal statements are the *invention* of mathematicians, while finitary ones are *discovered* by mathematicians. Finitary mathematics is the foundation upon which mathematicians build (ideal) mathematics. The philosophical merits and drawbacks of such an ontology – regardless of Gödel's theorems – cry out for a detailed discussion, but they are outside the scope of this essay; an interested reader may wish to read chapter 5 of [10] and chapter 2 of [13].

Prima facie this dichotomy between contentual and ideal mathematics looks interesting, but after a little thought we come to the following question: how exactly are we to conduct ideal mathematics if it has no content? Well, in order to conduct ideal mathematics, we can use formal languages. This works because a formal language requires no interpretation of the symbols; one simple needs to manipulate them correctly. Hilbert, who championed this approach to logic, put it best himself: when describing his rigorous axiomatisation of Euclidean geometry, he said, 'One must be able to say at all times – instead of points, straight lines, and planes – tables, chairs, and beer mugs' (p. 57 of [12]).

The precise formulation of our formal language does not matter for the purposes of this essay; our only technical requirement is that it be recursively axiomatisable, which we shall assume

[4] In the literature, the word 'real' is often used instead of 'contentual' in this context. I feel that the term 'contentual' is better, since it avoids confusion with statements referring to the mathematics of the real line $\mathbb{R}$, which are often in fact ideal in the sense of Hilbert's Programme. Moreover, 'contentual' is a much better translation of 'inhaltlich', the original German word that Hilbert used.

[5] Throughout this essay we shall assume that all statements and formulas are well-formed; that is, they obey the syntax of the language to which they belong.

throughout.[6] But for the sake of clarity, we shall take our formal language to be classical first-order logic (or perhaps second- or higher-order; it doesn't really matter). Indeed, it was Hilbert, with his student Bernays, who first formulated what we know today as first-order logic in a series of lectures at Göttingen during 1917–1921, so this seems appropriate.

Let us take this opportunity to clarify some terminology and notation. By a *formal system* (or just *system*), we shall mean a theory built from a logical language and a set of axioms. For example, by the *system of Peano Arithmetic* (PA), we mean the language $\mathcal{L}_{\mathrm{PA}} = \{0, 1, +, \cdot, <\}$, together with the usual axioms. By the prefix *meta-*, we mean it in the usual sense to denote something outside the formal language in question. For example, the statement 'Peano Arithmetic is a powerful system' would be a *metastatement*, since it is a statement *about* PA, rather than a statement *in* PA; in this example English would be the *metalanguage* and PA would be the *object language*. I have spelled this point out because it is very important that we make this distinction clear.

So far, so good: we have outlined the underlying structure of Hilbert's Programme, but we now need to describe Hilbert's Programme itself. Hilbert wanted to place mathematics on firm foundations, and he thought the best way to do this was to prove, using only *finitary* methods, that ideal extensions of finitary mathematics are consistent ([8], pp. 150–151). After all, it would be no good proving theorems in an ideal system if the system in question turned out to be inconsistent.

So we now know the goal of Hilbert's Programme; we shall see that Gödel's theorems show that this goal cannot be achieved. But before we move on, three digressions are in order.

Our first digression regards Hilbert's hope for completeness. This wasn't so much an aim of Hilbert's Programme as it was an underlying belief: Hilbert assumed that every properly formulated problem in mathematics can be solved.[7] Hilbert's belief in such a hope is perhaps best illustrated by the epitath on his gravestome: 'Wir müssen wissen, wir werden wissen' ('We must know, we will know').[8] We shall see in §3.2 that Gödel's first incompleteness theorem shows this belief to be misplaced.

Our second digression regards the subtle link between consistency and conservativeness. Roughly speaking, an extension is *conservative* over its base if it cannot prove anything new about its base. Thus, it would seem that an ideal extension of finitary mathematics is consistent iff it is conservative over its finitary base.[9] Why? Well, consider a finitary statement φ. Since finitary mathematics is bivalent, φ is either true or false; without loss of generality, assume that it is true. The only way an ideal extension could prove something new about φ would be to prove it false, which would be a contradiction if our extension were consistent. Conversely, if an ideal extension is inconsistent, it

[6] For the uninitiated, this means that we could programme a computer, say in C++, to check whether any given formula obeys the syntax and whether any given use of a rule of inference or an axiom is valid. This can be and is indeed done: for example, the programme Fitch that accompanies [2] does precisely this.

[7] This is what Detlefsen calls the *Axiom of Solvability* in [6].

[8] Further evidence for Hilbert's faith that there is no *ignorabimus* in mathematics can be found on p. 150 of [8] and p. 81 of [12].

[9] In [5] and [16], the conservativeness of ideal mathematics over finitary mathematics is called *real-soundness*.

can prove anything, in particular that φ is false, which is a new finitary statement. Unfortunately, things aren't quite so simple as this, for while finitary mathematics is bivalent, there is a question as to whether finitary mathematics is closed under negation (as we shall see in §4.1), which leads to problems formalising the above argument. We do not have space to go into this issue, but discussions can be found in Chapter 29 of [16] and on pp. 124–129 of [5].

In our last digression, I would like to raise a point that I have not seen discussed in the literature: What is the Hilbertian to make of an ideal system that is not an extension of finitary mathematics? For example, neither group theory or graph theory appear to be extensions of arithmetic. Indeed, even geometry, Hilbert's forte, does not appear to be an extension of arithmetic. It would seem that by ideal mathematics, Hilbert had in mind analysis or infinitary set theory, but he must have been aware of these other ideal theories. My only conclusion is that he would have seen them as extensions of finitary mathematics in the sense that they should be studied using finitary proof theory (which we shall discuss in §4.1). Okay, digressions over.

3 Gödel's theorems

In §3.1 we shall state and sketch the proofs of Gödel's incompleteness theorems. Then in §3.2 we shall describe how they are significant for Hilbert's Programme.

3.1 The incompleteness theorems

In this section we shall start by quickly going over some necessary background material. We will then state and sketch the proofs of Gödel's first and second incompleteness theorems. We will finish by stating and sketching the proof of an improvement of the first theorem by Rosser. Technical details can be found in [3] and [16].

Before we can state the theorems, we need to go over extensions of arithmetic and ω-consistency. We shall start with the former. The base system of arithmetic that we will be using is EA (*Elementary Arithmetic*). The precise details can be found on p. 236 of [16], and we shall go over them in §4.1, but roughly speaking, EA is a system of arithmetic in which one can perform basic arithmetic, exponentiation and bounded induction.[10] We then define (in a technical sense) T to be an *extension* of EA iff $\mathcal{L}_{\mathrm{EA}} \subseteq \mathcal{L}_T$ and every axiom of EA is an axiom of T.

Let us now discuss ω-consistency. Before we can state the definition, we need to describe how we can express numerals in a formal system. When we are referring to numerals in the metalanguage, we denote them in normal font, e.g. $1, 2, 3$, or in italics, e.g. k, m, n. However, a formal language has a strictly defined set of permissible symbols, and numerals are built up in a precise way; we use bold type to highlight this distinction. For example, in EA we build up numerals using a logical constant symbol $\mathbf{0}$ (which should be thought of as the interpretation of 0 in EA) and a unary function symbol $\mathbf{S}$ (which should be thought of as the interpretation of the successor function in

[10] EA is often denoted as '$\mathrm{I}\Delta_0 + \exp$' in the literature; I have chosen to go with 'EA' for the sake of brevity.

EA). So, for example, if we wanted to express 3 in EA, we would write $\mathbf{S}(\mathbf{S}(\mathbf{S}(\mathbf{0})))$. Now, expressing numerals in this way would soon become very long-winded, so we introduce some shorthand: To express the numeral n in EA, we write $\mathbf{n}$. So, to express 3 in EA, for example, we would write $\mathbf{3}$. We can now define ω-consistency:

Definition 3.1. An extension of arithmetic T is *ω-inconsistent* iff for some formula φ, $T \vdash \varphi(\mathbf{n})$ for every natural number n but $T \vdash \neg\forall x\varphi(x)$. T is *ω-consistent* iff it is not ω-inconsistent.

ω-consistency implies consistency, but the converse does not hold. Why? Well, the former is easy to prove via a contrapositive argument: if T is inconsistent, then it can prove anything, in particular $\varphi(\mathbf{n})$ for all n and $T\neg\forall x\varphi(x)$, and thus is ω-inconsistent. The latter is slightly harder to prove: Let us contruct a theory of arithmetic W whose language is $\mathcal{L}_{\mathrm{W}} = \{\mathbf{0}, \mathbf{S}, \varphi\}$, where φ is a one-place relation symbol, and whose axioms consist of $\neg\forall x\varphi(x)$ and the schema $\varphi(\mathbf{n})$ for all n. W is then consistent (since it does not have any form of induction) but ω-inconsistent. Now, we will initially require ω-consistency for Gödel's first incompleteness theorem, although we shall state an improvement by Rosser that weakens this requirement to consistency. However, even without this improvement, Gödel's theorem would still be significant, since if we want a formal system to capture what we mean by arithmetic, it had better be ω-consistent. We shall state the theorems:

Theorem 3.2 (Gödel's first incompleteness theorem)**.** *Let T be a ω-consistent extension of* EA*. Then there exists a sentence φ of T such that $T \nvdash \varphi$ and $T \nvdash \neg\varphi$.*[11]

Theorem 3.3 (Gödel's second incompleteness theorem)**.** *Let T be a consistent extension of* EA*. Then T cannot prove its own consistency.*

The key idea behind the proofs is *Gödel numbering*. Gödel's genius was to turn statements *about* arithmetic into statements *of* arithmetic. He did this by coming up with a method of turning formulas into natural numbers, and vice versa. Let us illustrate this with an example. Suppose our language consists of the symbols

$$\text{`}\exists\text{'},\ \text{`}\neg\text{'},\ \text{`}x\text{'},\ \text{`}P\text{'},\ \text{`(', and `)'}.$$

So an example of a formula in this language would be '$\exists x\neg P(x)$'. We can label the symbols of our language with numbers:

$\exists$	$\neg$	x	P	(	)
1	2	3	4	5	6

We can then write our formula '$\exists x\neg P(x)$' as the number

$$2^1 \cdot 3^3 \cdot 5^2 \cdot 7^4 \cdot 11^5 \cdot 13^3 \cdot 17^6;$$

the indices of the successive primes correspond to the labels of the symbols. By elementary calculation, one finds this number to be equal to 2768298641077907242305 0; this is the *Gödel number* of '$\exists x\neg P(x)$'. So, we know how to encode formulas as numbers, but how can we decode them? Well, by

[11] Such a sentence is said to be *undecidable*.

the Fundamental Theorem of Arithmetic (which says that every natural number has a unique prime factorisation, up to the order of the factors), if we were given the number 27682986410779072423050, we could decompose it into its prime factorisation and recover the formula '$\exists x \neg P(x)$' (since we order the primes). With this technique under our belt, we can introduce some notation. We use so-called 'Quine' or 'corner quotes' ($\ulcorner \varphi(x) \urcorner$) around a formula to denote the numeral in the system that corresponds to the Gödel number of that formula. So, in our example,

$$\ulcorner \exists x \neg P(x) \urcorner = \mathbf{27682986410779072423050}.$$

We now know how to use the language *of* arithmetic to talk *about* arithmetic. But how about *proving* statements about arithmetic in arithmetic? Well, the remarkable thing is that EA can in fact prove a great deal about itself. In particular, EA can prove statements about what it can prove: we shall not go into the details, but we can introduce a predicate 'Prv_T' into EA such that the following holds for every sentence A of T:

$$T \vdash \mathrm{Prv}_T(\ulcorner A \urcorner) \text{ iff } T \vdash A \tag{1}$$

All we need now is the following lemma, which again we shall not prove. This is Gödel numbering in action:

Lemma 3.4 (Diagonalisation Lemma)**.** *Let T be an extension of* EA*. Then for every formula φ of one free variable, there exists a sentence γ of T such that $T \vdash \gamma \leftrightarrow \varphi(\ulcorner \gamma \urcorner)$.*[12]

We can now prove the first theorem. Using the Diagonalisation Lemma and the proof predicate 'Prv_T', there exists a sentence G of T such that

$$T \vdash G \leftrightarrow \neg \mathrm{Prv}_T(\ulcorner G \urcorner).^{13} \tag{2}$$

We now ask the question 'Can T prove or disprove G?' Well, first suppose that T proves G; that is, $T \vdash G$. Then, by (1), $T \vdash \mathrm{Prv}_T(\ulcorner G \urcorner)$. But by (2), we have $T \vdash \neg \mathrm{Prv}_T(\ulcorner G \urcorner)$, contradicting T's consistency. Now suppose that T disproves G; that is, $T \vdash \neg G$. Then by (2), $T \vdash \mathrm{Prv}_T(\ulcorner G \urcorner)$. But then by (1), $T \vdash G$, again contradicting T's consistency. So $T \nvdash G$ and $T \nvdash \neg G$ and Theorem 3.2 is proved. Stunning.

So how do we prove the second theorem? Well, we shall not go into the details – although we shall discuss some important aspects of them in §4.2 – but we prove Theorem 3.3 by formalising the proof of Theorem 3.2 in T. Using the predicate Prv_T, we come up with a consistency sentence of T, which we denote Con_T:

$$\neg \exists y (y = \ulcorner \mathrm{Prv}_T(\ulcorner \mathbf{0} \neq \mathbf{0} \urcorner) \urcorner).^{14} \tag{3}$$

By 'formalising the proof of Theorem 3.2 in T', we mean that we prove that

$$T \vdash \mathrm{Con}_T \rightarrow G. \tag{4}$$

[12] We have $\varphi(\ulcorner \gamma \urcorner)$, rather than $\varphi(\gamma)$, since φ is a formula of arithmetic, and thus must take a numeral as its argument, not a formula.

[13] This sentence G is often called the *Gödel sentence* of T.

[14] Putting '$\mathbf{0} \neq \mathbf{0}$' is fairly arbitrary here; we could replace it with any other absurdity, say '$\mathbf{0} = \mathbf{1}$' or '$\mathbf{1} + \mathbf{1} = \mathbf{3}$'.

By (4), if Con_T were provable in T, then we could prove G in T, contradicting (our proof of) Theorem 3.2. Thus T cannot prove its own consistency, and we have Theorem 3.3.

Before we move on, we shall quickly mention and sketch the proof of an improvement on Theorem 3.2 due to Rosser ([14]). In the statement of Theorem 3.2, we had to assume ω-consistency. Rosser showed that we can weaken our requirement to just consistency.

Theorem 3.5 (Rosser)**.** *Let T be a consistent extension of* EA*. Then there exists a sentence φ of T such that $T \nvdash \varphi$ and $T \nvdash \neg\varphi$.*

The proof is similar to that of Theorem 3.2, but instead of using the Gödel sentence G that informally says 'I am unprovable in T', we use a *Rosser sentence*, which informally says 'If I am provable in T, then there is a shorter proof of my negation.'

Before we move on to showing how Gödel's theorems affect Hilbert's Programme, let us introduce two abbreviations. We shall refer to Theorem 3.3 as '(G2)'. Since Theorem 3.5 is an improvement on Theorem 3.2, we shall refer to Theorem 3.5 as '(G1)' (it's best to use one's sharpest tools).

3.2 The Standard Argument

In this section we shall outline the so-called *Standard Argument* (SA)[15] that Gödel's theorems undermine Hilbert's Programme. We shall briefly mention how (G1) shows Hilbert's belief in completeness to be naïve,[16] but we shall focus on the effect of (G2) on the hope for a finitary proof of consistency.[17] We shall finish by pointing out that while (G2) shatters hopes for a finitary proof of the consistency of an ideal system, Gödel's theorems do not in fact affect the underlying dichotomy of finitary and ideal mathematics that underlies Hilbert's Programme.

The key premise behind (SA) is that finitary mathematics is an extension of EA. As we shall see in §4.1, there is some debate over the precise nature of finitary mathematics, and we shall argue that finitary mathematics must at least contain EA, but in this section we shall simply take it as read that finitary mathematics is an extension of EA.

So how does (G1) show that Hilbert's belief that there is no *ignorabimus* in mathematics to be misplaced? Well, (G1) tells us that for any extension of arithmetic, there are always statements that we can neither prove nor disprove. Moreover, the sentence G in the proof of Theorem 3.2 is true, since it says of itself that it cannot be proved, which indeed it cannot. So we have found a true but unprovable statement. So much for completeness!

We now come to the second incompleteness theorem. We shall demonstrate its effect on Hilbert's Programme by reductio ad absurdum. Let F denote finitary mathematics and let I be an extension of F; we shall abuse notation and write this as $F \subseteq I$. Thus, since we are taking it as read that $\mathrm{EA} \subseteq F$, we have $\mathrm{EA} \subseteq I$. The aim of Hilbert's Programme is to come up with a proof in F of

[15] We have taken this term from [5].

[16] For a fun and highly original exposition of this argument, I recommend [1].

[17] There are some people who believe that in fact (G1) alone shows that a finitary proof of consistency is unobtainable. We do not have room to discuss it here, but an interesting critique can be found in the Appendix of [5].

I's consistency. So, to obtain a contradiction, suppose we have such a proof; that is, $F \vdash \mathrm{Con}_I$. Now, (G2) says that $I \nvdash \mathrm{Con}(I)$ (since $\mathrm{EA} \subseteq I$). But since $F \vdash \mathrm{Con}_I$ and $F \subseteq I$, a fortiori we have $I \vdash \mathrm{Con}(I)$; this is a contradiction. Therefore a finitary proof of the consistency of I cannot be carried out, and consequently neither can Hilbert's Programme.

The devastating effect of (G2) on Hilbert's Programme is plain to see. But what do Gödel's theorems say about the dichotomy of finitary and ideal mathematics that underpins Hilbert's Programme? Well, they say very little: while (G1) and (G2) prevent the aim of Hilbert's Programme, namely a finitary proof of ideal consistency, they do not affect its foundation. We shall discuss this in more depth in §5.

4 Do Gödel's theorems actually affect Hilbert's Programme?

4.1 Finitism

In this section, we shall address our assumption in §3.2 that finitary mathematics is an extension of EA. To do this, we will discuss the nature of finitary mathematics.

Unfortunately, Hilbert never stated precisely what he considered to be finitary mathematics, which has led to much philosophical debate. We will not enter into this debate in detail, but we shall deal with those issues that are relevant to the scope of this essay.

Let us start by outlining of the key idea behind finitary mathematics. The main objects of finitary mathematics are 'numerical symbols' (p. 192 of [8]), for example:

$$|, ||, |||, ||||, |||||||||$$

These are often referred to as *Hilbert strokes*.[18]

While Hilbert never stated precisely what he considered finitary mathematics to be, in [8] he did specify some conditions. Let us specify them:

(i) The main objects are Hilbert strokes.

(ii) The *rudimentary* statements are those that do not involve unbounded quantification,[19] such as $4 < 5$ or $\forall x < 10(x + 3 < 20)$. They can be built up using the usual logical connectives of $\rightarrow$, $\vee$, $\wedge$ and $\neg$.

(iii) For unbounded quantification, we have *schemas*. So, the statement $\forall x\ x + 1 = 1 + x$ isn't a finitary statement itself, but becomes one when we replace x by an actual numeral. Importantly, the negation of a schema is not a finitary statement, since we cannot check all the numbers for a counter example. This leads to problems for formalising finitary mathematics, since it makes it unclear whether finitary mathematics is closed under negation (as we men-

[18] These strokes are considered to be types, not tokens; this does of course lead us to the philosophical debate regarding the type/token distinction, but we shall not discuss it here. For an account of these issues regarding Hilbert strokes, see pp. 101–103 of [10].

[19] Hilbert considered bounded quantification to be finitarily kosher since, in principle at least, one can go through all the numbers in question and check whether the given statement is true of them.

tioned in our second digression at the end of §2), but unfortunately we do not have space to discuss this.

(iv) We only allow *contentual* induction; that is, induction on rudimentary formulas. Moreover, the conclusion is only admitted as a schema, as in (iii) above. (See pp. 45–46, 59–62 of [5] and pp. 94–98 of [13].)

Well, we now know what we have to work with. So what shall we take to be finitary mathematics? There is a lot of debate in this area which do not have space to discuss, so our response will be brief. Our system EA is weaker a system of arithmetic than the main competitor in the debate, Primitive Recursive Arithmetic (PRA), which roughly speaking consists of first-order logic, all primitive recursive functions, and bounded induction (details can be found on pp. 104–105 of [16]).[20] Also, EA adheres to (i)–(iii) above. Now, it does not adhere to (iv), since it admits unbounded quantification, but this is true of all systems based on classical logic, in particular PRA. However, given a use of unbounded quantification we can interpret it as a schema and thus adhere to (iv). Accordingly, our assumption in §3.2 that finitary mathematics must be an extension of EA is justified.

EA consists of *Robinson arithmetic* (Q), bounded induction (= contentual induction in point (iv)), and exponentiation. The precise details of Q can be found on pp. 55-56 of [16], but its language is $\mathcal{L}_{\mathrm{Q}} = \{\mathbf{0}, \mathbf{S}, +, \cdot\}$ and its axioms state basic arithmetical truths, such as the commutativity of $+$ and $\cdot$, the distributivity of $\cdot$ over $+$, etc. One axiom that we should point out is $\forall x(x \neq \mathbf{0} \rightarrow \exists y(x = \mathbf{S}(y)))$. We add exponentiation because Q is too weak to express it (and we use it for Gödel numbering).

Now, importantly, Hilbert didn't restrict finitary mathematics to *just* Hilbert strokes. Hilbert considered the symbols in a formal system to be finitary objects too, and thus we can talk about them in a finitary way. This then allows us to use finitary mathematics as *metamathematics* or *proof theory*, i.e allows us to talk about finitary proofs of consistency and so forth, which is of course crucial for Hilbert's Programme. We will not go into how Hilbert expounded this finitary proof theory[21], since it is rather technical and, perhaps ironically, Gödel numbering allows us to use finitary mathematics as our proof theory.

Before we move on to the next section, let us address a question that has received very little discussion in the literature: Are the proofs of (G1) and (G2) finitary? For if they are not, perhaps a Hilbertian could argue that they are not valid as metatheorems. The only discussion I can find is on p. 80 of [5], where Detlefsen claims that (SA) is not finitary, since it involves an implicit universal quantifier ('for *all* extensions of EA...'), but that the argument is still 'cogent' and thus must be addressed. We do not have room to go into a detailed discussion, but let's make a couple of points. Firstly, (SA) *is* finitary, since we can take the universal quantification as a schema. Secondly, the proofs of (G1) and (G2) only require basic arithmetic and induction; they do not appeal to infinite

[20] PRA is supported by Tait ([17]) and, to a lesser degree, Detlefsen (p. 66 of [5]).
[21] See the Appendix in [12] for a summary.

ordinals, unlike Gentzen's proof of the consistency of PA, for example. Thus, a Hilbertian is not in a strong position to attack (SA) or the proofs of (G1) and (G2) on a technicality, since they are certainly finitary "in spirit". Indeed, it was Hilbert and his student and fellow Hilbertian Bernays who first *rigorously* proved (G2) in [9].[22] Moreover, Gödel did not initially think that his theorems were fatal for Hilbert's Programme (p. 40 of [7]), and Bernays had to win him round to the opposite position (p. 91 of [5]), which Bernays surely would not have done if he thought something was (finitarily) wrong with the proofs. All this suggests that Hilbert and his followers considered the proofs to be finitarily kosher.

4.2 Detlefsen, The Last Son of Hilbert('s Programme)

In this section we shall consider arguments put forward against the Standard Argument (SA) by Michael Detlefsen in [5].[23] Detlefsen appears to be unique in his belief that Hilbert's Programme is unaffected by Gödel's theorems, and we shall conclude that his arguments are flawed.

Detlefsen puts forward two arguments against (SA):[24]

The Stability Problem (SP). The statement of consistency in (G2) is just *one* (class of) statement(s) of consistency. There may be others that can be proved. That is, a formula's property of being unprovable may not be a *stable* property of it being a statement of consistency.

The Convergence Problem (CP). It may not be the case that all finitary proofs in a given ideal system are 'feasible' (we shall discuss what 'feasible' means shortly.) That is, finitary mathematics and feasible mathematics may not *converge*.

Let us explain how these apparently undermine (SA). The problem that (SP) causes for (SA) is clear: If there is a consistency statement of an ideal system I that can be proved in I, then it may be the case that a finitary proof of the consistency of I can be found. The idea behind (CP) is that even if (G2) does apply to I, the feasible part of I may not contain EA, and thus we may be able to use finitary methods that are outside of the feasible part of I to prove the consistency of the feasible part of I.

We shall argue against (CP) at length. We shall not attack (SP) directly due to lack of space, but instead we shall show that Detlefsen's proposed example of a system that can prove its own consistency is flawed.

Our attack on (CP) will be two-pronged: first we shall attack its foundation, and then, taking no prisoners, we shall further attack it directly. Detlefsen bases (CP) on what he calls the *Thesis of Strict Instrumentalism* (TSI):

> 'Of the infinitely many ideal proofs constructible in a given system T of ideal mathematics, only finitely many of them are of any value of instruments of

[22] Gödel only sketched the proof of (G2) in [7]; he planned to prove it rigorously in a later paper but never did (see footnote 68a in [7]).

[23] The reader should note that we have changed some of Detlefsen's terminology to fit with that of this essay. In particular, he uses 'real' where we use 'contentual'; c.f. footnote 4.

[24] Detlefsen in fact puts forward another argument against (SA), what he calls the *Problem of Strict Intrumentalism*; this is very closely related to the Convergence Problem, and thus for the sake of brevity we shall not discuss it.

human epistemic acquisition.' ([5], p. 84.)

We then define a proof to be *feasible* iff it has value as an 'instrument of human epistemic acquisition.' Before we argue against (TSI), let us explain why it implies (CP): if only finitely proofs of an ideal system are feasible, a fortiori, how do we know that all finitary proofs of the system are feasible?

Detlfsen's argument for (TSI) is as follows: Checking each line of a proof takes a finite expenditure of time and effort (even if only a small amount), and thus, since we only have a finite amount of time in which to check proofs, there are only finitely many characters that can appear in a proof, and hence we have (TSI). Initially this seems like a plausible argument, since indeed there are only so many hours in a mathematician's lifetime in which to prove theorems – even David Hilbert had to sleep. But then we realise that Detlefsen has made the classic mistake of confusing *actual* infinities with *potential* infinities. Humans, even after 10,000 generations and with the help of the finest computers IBM can produce, will only ever produce finitely many theorems. This much cannot be disputed. However, this is not to say that there is a limit on the number of theorems that can be proved: every proof is finite, so given enough time and effort, it can be proved. In short, only finitely theorems *will* be proved, but infinitely many *can* be proved.

There is an obvious objection to the above argument: *I* have confused potential and actual infinities. Detlefsen can say, 'you said it yourself: even after 10,000 years, humans will only have proven *finitely* many theorems.' The problem seems to arise over the conception of Hilbert's Programme. I suggest that Hilbert saw ideal mathematics as the realm of mathematical creativity, a place of boundless exploration, while Detlefsen sees ideal mathematics purely in terms of its 'value as an instrument of human epistemic acquisition' (p. 84 of [5]; slightly adapted) and as such, (TSI) is valid. On this point we shall have to agree to disagree, for alas our prophet David has moved on to the great Hilbert space in the sky. But where does this leave us? Well, let us try a different ploy: for the sake of argument, we shall concede (TSI), but now we shall attack (CP) directly.

To employ (CP) against (SA), Detlefsen changes the conception of a formal system from one of infinitely many theorems to one of finitely many feasible theorems; we call such a system a *feasible system*. So how does Detlefsen propose to set up such a feasible system? After all, Hilbert's Programme can only be carried out in this way if we can show how to actually construct such a feasible system. Detlefsen proposes a method using what he calls *Hilbertian residues*. He describes the process on p. 89 of [5]: Let T be a formal system. Successively eliminate proofs from T as follows:[25]

(i) Remove all unfeasible ideal proofs of finitary formulas.

(ii) Remove all ideal proofs of finitary formulas that have an equally short and simple finitary proof.

(iii) Remove all finitary proofs of finitary formulas.

[25] I have adapted the precise quotation; in particular, I have used the word 'finitary' instead of Detlefsen's 'real'.

Once (i)–(iii) have been carried out, take all the axioms of T that are used in the remaining proofs and close under logical operations. The new system is the *Hilbertian residue* of T, denoted T_H. Detlefsen's idea is that (G2) may well not apply to such a system. Prima facie this looks pretty good: the system T_H may not contain EA, and thus (G2) may not apply to it, while in the meantime T_H contains all the theorems we can feasibly prove. But on a little further thought we notice a very serious flaw: step (i) is simply not well-defined, let alone effective. Where are we to draw the line on what counts as feasible and what does not? This is not only a problem of vagueness though: presumably over time the boundary of feasibility will be pushed forward, since as the centuries go by our idea of what is feasible will expand. For example, the proof of the Four Colour Problem by Appel and Hanken in 1976 would no doubt have been considered 'unfeasible' in the 19^{th} century, considering the computing power needed to carry it out. Thus we find that our feasible systems will be time- (and technology-)dependent, which surely is not anyone's idea of a formal mathematical system, let alone Hilbert's. In fact, this problem of the vagueness and temporal-dependence of feasibility isn't just a problem for the particular method of Hilbertian residues: it strikes at the very core of (CP), for if we don't know what feasible actually means, how can we talk of finitary mathematics and feasible mathematics 'converging'?

Now, one might point out that we may be able come up with a definition of *feasible* using computational complexity theory. For example, we might say (perhaps arbitrarily) that a function is *feasible* iff it can be carried out in polynomial time. This would eliminate the problem of vagueness, but we would still have the problem of time- and technology-dependence: In 10,000 years time we might decide that polynomial time is far too limited to draw the line of feasibility, and that exponential time – or perhaps an even higher complexity class – is more appropriate. Our objection to Hilbertian residues and (CP) still holds.

We shall finish by demonstrating that Detlefsen's attempt to show that (SP) cannot be solved is flawed. He tries to do this by giving an example of a formal system that can prove its own consistency. The system in question was first developed by Rosser and is based on first-order logic, but it is equipped with a way of "cheating" inconsistency: We order the sentences of the language, say lexicographically, and then add the rule that before we can admit a sentence as a theorem (assuming that we have a conventional proof of it), we check whether it contradicts any previous theorems: if it does not, then we admit it as a theorem; if it does, then we do not. The system is consistent by construction.

I called this "cheating" for a reason: This goes against what Hilbert meant by a formal system, and indeed against what we mean by mathematics. For whether a statement is or is not a theorem should not be order-dependent; different orderings can lead to different theorems. Moreover, we can construct such systems that are consistent but nevertheless are very silly: For example, take EA and add $\mathbf{0} \neq \mathbf{0}$ as an additional axiom. Then place $\mathbf{0} \neq \mathbf{0}$ first in your ordering; now $\mathbf{0} \neq \mathbf{0}$ will be a theorem of this "consistent" system but $\mathbf{0} = \mathbf{0}$ will not. We don't want this.

5 Dealing with the Aftermath

In this section we shall argue the main thesis of this essay, that while Gödel's theorems are devastating for the aim of Hilbert's Programme, that of a finitary proof of ideal consistency, the underlying structure of finitary and ideal mathematics is left unaffected.

How do (G1) and (G2) affect Hilbert's Programme? Well, we have seen that (G1) shows that Hilbert's underlying faith in completeness was naïve and, more importantly, that (G2) shows that a finitary consistency proof cannot be achieved. But what of the underlying structure of finitary and ideal mathematics? Do (G1) or (G2) affect that? As we suggested at the end of §3.2, we surely must conclude that they do not, since neither (G1) nor (G2) says anything *about* the dichotomy; they only say what we cannot achieve *with* the dichotomy.

Now, one might say that the use of finitary mathematics as metamathematics is brought down by (G2), since we know that finitary mathematics cannot prove ideal consistency. But surely any other possible metamathematics, if it is to be viable, must contain finitary mathematics, and thus must also be affected by (G2). In short, (G2) shows us that no metamathematics can prove consistency, not just finitary metamathematics.

So what is our proposal exactly? Well, we suggest that a Hilbertian may happily carry on using finitary mathematics as metamathematics and proving ideal theorems, but – in light of (G2) – be aware that one day they may come across an inconsistency. And, should our Hilbertian discover an inconsistency in a given ideal system, they must not despair and abandon the system entirely, but rather go back to the axioms of the system and try to fix the source of the inconsistency. Indeed, the discovery of an inconsistency does not make all previous work in that system worthless, for many of the proofs may well not use the inconsistent axiom. Our analogy is that of a scientist who proposes a theory. If the theory is disproved, our scientist should not simply throw in the towel, but rather should go back to the drawing board and, if possible, fix the error. This is how science often develops, and so too can mathematics. Indeed, we offer a major historical example of such a process occurring in mathematics. When Russell's Paradox was discovered in 1903, mathematicians saw the source of the problem, namely unrestricted set-comprehension,[26] and set about resolving the issue. And indeed, just five years after Russell discovered this antimony in Frege's work, Zermelo published some axioms of set theory that removed the troublesome construction; these axioms in fact form the basis of modern set theory.

Our argument is similar to that of Curry, who sees mathematics as the 'science of formal systems' (p. 154 of [4]); that is, the study of what different formal systems can and cannot prove. But unlike Curry, who is a formalist "all the way down to his metawear," [27] we point out that Gödel's theorems do not force us to change our *meta*mathematics from finitary mathematics. That is, while Curry takes the view that we need to formalise even our metamathematics in light of

[26] The precise source was Frege's now infamous Basic Law V, which, in modern language, stated that for any property P, the set $\{x : P(x)\}$ exists. The problem occurs when we take P to be $x \notin x$.

[27] I owe this light-hearted description to Øystein Linnebo and Richard M. Nixon.

Gödel's theorems, our argument in the beginning of this section shows that this view is mistaken. Furthermore, our position is in fact better than Curry's, since building formal systems on contentual finitary foundations avoids various objections that can be levied against his form of formalism. For example, Curry has to deal with the contention that all of mathematics cannot consist entirely of formal systems: mathematics was practised for many centuries before it was ever formalised and history shows us that, typically, a branch of mathematics is formalised *after* it has been developed (see p. 170 of [15]). Our position also avoids the regress of Curry's formalism: If metamathematics is itself a formal sytem, then can we not study it just like any other formal system? But this would require metametamathematics, and so on ad infinitum. Now, there are of course other problems associated with finitary mathematics as metamathematics, perhaps most importantly its precise nature, *but these problems do not stem from Gödel's theorems.*

6 Conclusion

In this essay we started by covering the nature of Hilbert's Programme and explaining how Gödel's theorems affect it. We addressed the nature of finitary mathematics, concluding that the proofs of Gödel's theorems are finitary, and then moved on to rebuke various arguments put forward by Detlefsen in [5] against the claim that Gödel's theorems do not affect Hilbert's Programme. Finally we considered the aftermath of Gödel's theorems, concluding that while the hope for a finitary proof of ideal consistency is completely destroyed by Gödel's theorems, the underlying structure of finitary and ideal mathematics is left unaffected.

References

[1] Apostolos, D. and Papadimitriou, C.H., *Logicomix: An Epic Search for Truth*, London: Bloomsbury Publishing, 2009.

[2] Barwise, J. and Etchemendy, J., *Language, Proof and Logic*, Stanford: CSLI Publications, 1999.

[3] Boolos, G.S., Burgess, J.P., and Jeffrey, R.C., *Computability and Logic*, Cambridge: Cambridge University Press, 4th ed., 2002.

[4] Curry, H.B., 'On the Definition and Nature of Mathematics', *Philosophy of Mathematics*, edited by P. Benacerraf and H. Putnam, pp. 152–156, New Jersey: Prentice-Hall, Inc., 1964.

[5] Detlefsen, M., *Hilbert's Program: An Essay on Mathematical Instrumentalism*, Synthese Library, Dordrecht: D. Reidel Publishing Company, 1986.

[6] Detlefsen, M., 'Formalism', *The Oxford Handbook of Philosophy of Mathematics and Logic*, edited by S. Shapiro, pp. 236–317, New York: Oxford University Press Inc., 2005.

[7] Gödel, K., 'On Formally Undecidable Propositions of *Principia Mathematica* and Related Systems I (1931)', *Gödel's Theorem in focus*, edited by S.G. Shanker, pp. 17–47, London: Routledge, 1988.

[8] Hilbert, D., 'On the Infinite', *Philosophy of Mathematics*, edited by P. Benacerraf and H. Putnam, pp. 134–151, New Jersey: Prentice-Hall, Inc., 1964. Originally delivered on 4th June 1925 before a congress of the Westphalian Society in Munster, in honour of Karl Weierstrass. Translated by Erna Putnam and Gerald J. Massey from *Mathematische Annalen* (Berlin) vol. 95 (1926), pp. 161–190.

[9] Hilbert, D. and Bernays, P., *Grundlagen der Arithmetik*, Berlin: Springer, 1934–1939. Two volumes.

[10] Körner, S., *The Philosophy of Mathematics*, London: Hutchinson & Co Ltd, 1971. First published 1960.

[11] Kreisel, G., 'Hilbert's Programme', *Philosophy of Mathematics*, edited by P. Benacerraf and H. Putnam, pp. 157–180, New Jersey: Prentice-Hall, Inc., 1964.

[12] Reid, C., *Hilbert, with an appreciation of Hilbert's mathematical work by Hermann Weyl*, Berlin: Springer-Verlag, 1970.

[13] Resnik, M.D., *Frege and the Philosophy of Mathematics*, London: Cornell University Press Ltd., 1980.

[14] Rosser, B., 'Extensions of Some Theorems of Gödel and Church', *The Journal of Symbolic Logic*, vol. 1, no. 3: pp. 87–91, 1936.

[15] Shapiro, S., *Thinking about mathematics*, New York: Oxford University Press Inc., 2000.

[16] Smith, P., *An Introduction to Gödel's Theorems*, Cambridge: Cambridge University Press, 2007.

[17] Tait, W.W., 'Finitism', *The Journal of Philosophy*, vol. 78, no. 9: pp. 524–546, 1981.

Acknowledgements

I wish to thank Øystein Linnebo for his invaluable comments, suggestions and help with the German language. I would like to thank Josephine Salverda for her brilliant proofreading. I would also like to thank Kate Hodesdon, Dave Mendes Da Costa, Michelle Montague and Mark Pinder for invaluable discussion. All mistakes are, however, entirely my own.