Utrecht University

Can the Human Mind Escape Gödel?

Linda van Vliet
5648319

Bachelor Thesis

Utrecht University, Faculty of Humanities
Department of Philosophy and Religious Studies
Theoretical Philosophy

Supervisor
Dr. H. L. W. Hendriks
Second reader
Dr. M. A. Peters

June 19, 2020

Summary

In 1931, Kurt Gödel proved his revolutionary incompleteness theorems, which demonstrate that formal mathematical systems are fundamentally limited in regards to what they can prove.[1] While they are strictly mathematical results, they are sometimes argued to imply that a computational model cannot exactly model the human mind. Such arguments, often called Gödelian arguments against mechanism, attempt to show that the human mind is not limited in the way formal systems are as demonstrated by the incompleteness theorems.

This thesis examines two influential Gödelian arguments which are both shown to be unsuccessful. The first is by J. R. Lucas,[2] an argument appealing to intuition, yet it makes a severe mistake of unjustifiably assuming consistency of the mind. Lucas argues that we as human beings can "grasp" that the Gödel sentence is true; however, there is serious reason to doubt this claim. The second Gödelian argument considered is by Storrs McCall.[3] This is a more sophisticated argument based on the claim that while the truth value of the Gödel sentence is unknown to us, we can see that it diverges from its provability. Understanding this divergence, McCall argues, is a demonstration of a uniquely human ability. McCall makes several mistakes in his argumentation as well, which causes him to fail in refuting mechanism.

Finally, I cautiously present an argument for why mechanism may never be disproven by an argument from the incompleteness theorems. Since a Gödelian argument must contain a claim of the mind being able to prove or understand something a machine cannot, it must make precise claims about how the mind deduces and reasons. Our knowledge of the human mind, I argue, is extremely unlikely to ever be precise enough and the anti-mechanist may be wise to look beyond Gödel's work for an argument in support of his claim.

[1] Kurt Gödel, "On formally undecidable propositions of *Principia Mathematica* and related systems I," in *Collected Works I*, ed. Solomun Feferman (Oxford: Oxford University Press, 1986), 144-196.

[2] J. R. Lucas, "Minds, Machines, and Gödel," *Philosophy* 36 (1961), 112-127.

[3] Storrs McCall, "Can a Turing Machine Know that the Gödel Sentence is True?" *The Journal of Philosophy* 96 (1999), 525-532.

Contents

Introduction

Kurt Gödel published in 1931 a revolutionary paper. His "On formally undecidable propositions of *Principia Mathematica* and related systems I" settled some of the most pressing questions of the debate on the foundations of mathematics (often called "Grundslagenstreit") that occurred in the beginning of the twentieth century. Furthermore, it is almost surely safe to say that no other mathematical result has aroused as much interest among non-mathematicians then Gödel's two incompleteness theorems.

The theorems are results about formal mathematical systems and show how these are fundamentally limited in regards to what they can prove. Some have argued that they have serious implications for the philosophy of mind. They argue by so called 'Gödelian arguments' that the theorems are evidence that human intellect cannot be reduced to a computational model. The idea is simple. If we can know that the human mind does is not limited in the way that formal systems are, it cannot be modeled by a formal system or a machine. These arguments, if correct, would have severe implications to cognitive sciences. Specifically, they would imply that the program of strong Artificial Intelligence is not viable.

The purpose of this thesis is to evaluate these claims. Do Gödel's incompleteness theorems show that a mechanist view of human mathematical intellect is unwarranted? An answer to this question requires an examination of Gödelian arguments against mechanism and identification of their strengths and weaknesses. I examine two of the most referenced Gödelian arguments in particular. The first is proposed by J. R. Lucas; the second by Storrs McCall. Their arguments will be carefully reconstructed and criticized in order to expose their philosophical assumptions. Only when we have a proper understanding of the premises and actual consequences of Gödel's incompleteness theorems are we able to do this. Thus, an answer to this question also requires an overview of the theorems themselves.

The thesis is divided into three chapters. The purpose of the first chapter is to provide the tools necessary to fairly assess the arguments. In this chapter, I will define the mechanist position, as well as stating the results of Gödel's incompleteness theorems together with some important steps in their proof. Then, in chapters two and three I will examine the anti-mechanist arguments by Lucas and McCall respectively.

Although Lucas' argument is easily brushed aside by philosophers, there are several reasons for why I think an analysis of his argument is relevant

to the philosophical debate surrounding the implications of Gödel's incompleteness theorems for the philosophy of mind. The debate is usually considered to have started with the Lucas argument, presented in his 1961 paper "Minds, Machines, and Gödel." Not only did it inspire many of the more recent Gödelian arguments against mechanism, a number of them depend on precisely the same philosophical assumptions. Thus an analysis of Lucas' argument reveals strengths and weaknesses of other anti-mechanist perspectives. Most importantly, however, I think his argument captures a prevailing intuition of the consequences of the incompleteness theorems. Even among mathematicians it is not uncommon to find interpretations of Gödel's first theorem stating that the Gödel sentence of Peano Arithmetic is true despite it being unprovable in its system. Nonetheless, such an interpretation is based on a fundamental misunderstanding of the results of Gödel's theorems and, as I will argue, is unwarranted.

Many Gödelian arguments against mechanism resemble Lucas' original argument. Storrs McCall is one of the few to have presented a radically different approach to disproving mechanism while still basing his arguments on the results of the incompleteness theorems. McCall, acknowledging the problems of the Lucas argument, provides a much more convincing argument. He claims that Gödel's first theorem reveals a sharp dividing line between human and machine thinking. Although his attempt is more sophisticated than earlier Gödelian anti-mechanist arguments, it cannot be saved from certain critiques.

The failure of the aforementioned Gödelian arguments does not settle the matter definitely. However, in the concluding chapter I will reflect on a common weakness to the arguments which, as I will argue, suggests that the mechanist position cannot be refuted by an argument based on the incompleteness theorems.

1 Setting the Stage

Before we delve into Gödelian arguments against mechanism, we must clarify some terminology. In the first section of this chapter, I will provide a definition of mechanism that is relevant to our discussion. In the second I will provide an overview of Gödel's theorem. Finally, I will briefly argue how the theorems are relevant to discussions on mechanical models.

Defining mechanism

Mechanism can generally be understood as the view that the human mind is, or can be accurately simulated by, a computational system like a Turing machine. This thesis is broad and, depending on the debate, can be interpreted in different ways. Relevant to our discussion is the question whether human *mathematical* intellect surpasses the abilities of Turing machines. Thus, I will take mechanism as the claim that for all human beings there is a Turing machine that proves the exact same set of arithmetical sentences as that person proves.

A Turing machine is an idealization of a computer; it is formal model of a computational system that is not limited by things like time or memory. The debate surrounding mechanism is therefore interesting only when we idealize on the human intellect as well. Consequently, the mechanist and anti-mechanist are not concerned with what sentences are provable by a particular mathematician; they are interested in statements that are *in principle* knowable to an ideal mathematician.

Gödel's incompleteness theorems

Gödel's incompleteness theorems are mathematical results about formal systems of arithmetic, though we will get into the reason why the theorems are related to Turing machines later. For now, let us consider formal systems. We call T a formal system if it is a set of sentences, called *axioms*, together with inference rules, that satisfies the following property. It must be decidable by a finite machine whether a sentence is among the axioms of T and whether it can be inferred from the axioms of T by the specified rules of proof. In mathematics, a proof is a finite sequence of sentences, all of which are either axioms or are inferred from previous sentences according to the specified rules of proof. Consequently, provability is a property *relative to* a

system.

Gödel's theorems concern formal systems of *arithmetic*, that have axioms for the multiplication and addition of natural numbers. His original proof was formulated for a system similar to what we now call Peano Arithmetic (PA), though his theorems hold for any formal system that includes PA.

Gödel ingeniously thought of a way to express certain metamathematical notions *about* formal systems like consistency and provability within the language of arithmetic *itself*. In other words, Gödel showed how metamathematical statements can be discussed inside mathematics. He constructed a method to assign each formula φ of the language of T a unique natural number. This number is detonated by $\ulcorner\varphi\urcorner$ and called the *Gödel number* of φ. Then, since proofs are finite sequences of sentences, proofs too can be assigned a unique code number. Gödel furthermore showed that the *Proof-in-T* relation, "x is the Gödel number of a proof in T of a sentence with Gödel number y", is expressible in the language of arithmetic: $\mathrm{Prf}_T(x, y)$. Then the same holds for the provability predicate $\mathrm{Bew}_T(x)$, named after the German *Beweis*, which expresses that "there is a proof in T of the sentence with Gödel number x". Using the diagonal method on the provability predicate, one can construct a sentence G_T, called the *Gödel sentence*, such that PA proves the following:

$$G_T \leftrightarrow \neg\mathrm{Bew}_T(\ulcorner G_T \urcorner).$$

From this, Gödel proved his *first incompleteness theorem*: Let T be a formal system of arithmetic, then T satisfies the following:

1. If T is consistent, then $T \nvdash G_T$.
2. If T is ω-consistent,[4] then $T \nvdash \neg G_T$.

Using the provability predicate, one can construct a predicate for the consistency of T, $\mathrm{Con}(T)$, which denotes the absence of a proof for a contradiction in T. Clearly $\mathrm{Con}(T)$ is expressible in the language of arithmetic as

[4]A system T is ω-consistent if and only if there is no formula φ such that T can prove $\varphi(n)$ for each natural number n, yet also prove $\exists x\neg\varphi(x)$. This condition implies consistency. The distinction need not worry us though, as Barkley Rosser proved in 1936 that mere consistency is sufficient for incompleteness. His proof makes use of a sentence different from the Gödel sentence. See Rosser, "Extensions of Some Theorems of Gödel and Church," 87-91; or Craig Smoryński, "The Incompleteness Theorems," 840-841.

well. From the fact that the formalization of the first incompleteness theorem, $\mathrm{Con}(T) \leftrightarrow G_T$, can be proven in PA follows the *second incompleteness theorem:* If T is consistent, then $T \nvdash \mathrm{Con}(T)$.

Formal systems or machines?

One final detail must be clarified. Why do philosophers and mathematicians connect the incompleteness theorems to machines? After all, they are results explicitely about *formal systems* - not machines. But equating a formal system to a Turing machine is legitimate as Gödel points out in a postscript added to the reprinting of one of his 1934 lectures:

> Turing's work gives an analysis of the concept of 'mechanical procedure' ... A formal system can simply be defined to be any mechanical procedure for producing formulas, called provably formulas. For any formal system in this sense there exists one ... that has the same provable formulas.[5]

[5]Kurt Gödel, "On undecidable propositions of formal mathematical systems," in *Collected Works I*, ed. Solomun Feferman (Oxford: Oxford University Press, 1986), 369-370

2 The Lucas Argument

An early - and perhaps the best known - argument for mechanism resting on the first incompleteness theorem was given by philosopher John Randolph Lucas. In his 1961 article "Minds, Machines and Gödel", Lucas argues that the theorem enables him to prove an arithmetical sentences that cannot be proved by any formal system. Hence, his mind cannot be exactly modeled by any formal system.

In the first section of this chapter I present Lucas' argument against mechanism. By a reconstruction the crucial assumption of consistency of the human mind will be exposed. This assumption is controversial as I will then argue. Any possible way of establishing consistency compromises Lucas' argument. I argue that Lucas fails in refuting mechanism and, moreover, that *any* argument for the incompatibility of the incompleteness theorems and mechanism necessarily fails when it depends on the claim that human intellect is consistent or sound.

Reconstructing Lucas' argument

Lucas' argument may be put as follows. Suppose, for reductio, that mechanism holds. Then there is a consistent formal system, T, that proves exactly the same arithmetical sentences that Lucas is able to prove. It is known that T can formally prove $\mathrm{Con}(T) \rightarrow G_T$; Lucas, by assumption, can prove the same. By the first incompleteness theorem, T cannot formally prove G_T. Since Lucas' mind is consistent, T is consistent, and Lucas can apply modus ponens to $\mathrm{Con}(T) \rightarrow G_T$ which allows Lucas to prove a sentence, G_T, that T is unable to prove. It follows that T cannot be an exact model of Lucas' mind and as this argument can be applied to any arbitrary formal system, mechanism is refuted.[6]

A frequently raised objection to Lucas' argument and similar anti-mechanist Gödelian arguments concerns assuming that the mind is consistent to begin with. Hilary Putnam was among those raising the objection, doing so in as early as 1960.[7] It may well be possible that the mind is inconsistent. Moreover, as a contradiction in first-order logic implies *any* sentence, an in-

[6] This is a paraphrase of the argumentation in Lucas, "Minds, Machines, and Gödel," 112-116.

[7] Hilary Putnam, "Minds and Machines," in *Dimensions of Minds*, ed. Sidney Hook (New York: New York University Press, 1960), 138-164.

consistent mind would be able to derive *all* sentences. Its set of derivable sentences would be trivially equal to those of an inconsistent formal system. It is clear that the success of Lucas' argument rests on the assumption that the mind is consistent, which requires further justification.

This problem did not go unnoticed by Lucas. In his initial paper he provided an argument for the consistency of the mind by making an appeal to empirical evidence. According to Lucas, our reasoning behavior indicates that we are not like inconsistent formal systems. Though we assert contradictory statements on occasion, these are quite obviously the results of mistakes that we would prefer to correct. He argues that "If we really were inconsistent machines, we should remain content with our inconsistencies, and would happily affirm both halves of a contradiction. Moreover, we would be prepared to say absolutely anything - which we are not."[8]

Establishing Con(T)

The Lucas argument is appealing as it certainly agrees with our intuition. However, even if we ignore for the time being whether or not we find Lucas' evidence for the consistency of the mind convincing, there are fundamental problems with his argument. I will argue in this section that these problems cannot be solved and that his argument is unsuccessful in refuting mechanism.

Crucial to the argument is applying modus ponens to Con(T) and Con(T)$\rightarrow$ G_T to infer G_T. However, Lucas is mistaken when he equates the metamathematical claim of the consistency of the mind to the arithmetical sentence Con(T). The latter is mathematically constructed as the absence of T-proofs of contradicting sentences in T, a precisely defined sentence that is required to establish T's Gödel sentence. Lucas has merely provided evidence of the metamathematical claim. Perhaps it is possible to establish Con(T) so that Lucas' argument might still succeed. In what follows I will critically examine the two possible strategies according to which Con(T) might be established: formally and informally.

First, let us consider the possibility that Lucas formally establishes Con(T) where T is a formal system supposedly modeling his mind. His argument then rests on the following three assumptions.

(1) For reductio, mechanism holds;

[8]Lucas, "Minds, Machines, and Gödel," 121.

(2) The formal system T modeling his mind is consistent;

(3) Lucas can formally prove Con(T).

These assumptions allow him to infer G_T that contradicts the assumption that our mind is exactly modeled by system T. Lucas' position is that we are forced to reject assumption (1), though I will demonstrate that this position is problematic.

Lucas may be compelled to reject (1), but as he has not actually provided a formal proof of Con(T), which is needed to reach the desired contradiction, accepting premise (3) simply begs the question. Therefore, the mechanist is entirely justified in rejecting the third premise. Another legitimate possibility that remains is for the mechanist who trivially accepts premise (1), to accept premise (3) while rejecting (2). She might content that Lucas can formally prove Con(T) and, by mechanism, that $T \vdash$ Con(T). Indeed, Gödel's second incompleteness shows that T is an inconsistent formal system, contradicting assumption (2). However, Lucas' argument does not prevent the mechanist in rejecting (2).

All Lucas has shown is that premises (1) through (3) are jointly contradicting: human beings cannot be exactly modeled by a formal system T while being both consistent *and* having the ability to formally prove Con(T). Lucas' argument does not disprove mechanism in view of the fact that the mechanist may simply reject the assumption that she can be modeled by an inconsistent formal system.

Equating formal and informal proofs

We have seen that the validity of Lucas' argument rests on his ability to establish Con(T). This is an unfortunate position considering the consistency of his own mathematical reasoning is easily jeopardized by the results of Gödel's second incompleteness theorem. We consider the second manner in which Con(T) can be established: informally. At first glance this seems to solve the issues discussed so far as the second incompleteness theorem does not preclude a formal system from giving an informal proof of its consistency. Furthermore, Lucas' reasoning seems to favor this claim. After all, he has given empirical evidence in support of his claim that his mind is consistent. So let us assume for now that Lucas can informally establish Con(T) so that he may apply modus ponens to Con(T)$\rightarrow G_T$.

It appears that this change allows the Lucas argument to succeed, but as pointed out by Paul Benacerraf in his article "God, the Devil, and Gödel", it gives rise to new and severe problems. Benacerraf stresses that Lucas falsely equivocates two different senses of 'proving'. The equivocation is made explicit by our reformulation of the argument: it occurs when we infer that Lucas can formally prove G_T from the informal proof of $\mathrm{Con}(T)$. Clearly, this inference is invalid and we must acknowledge that from the assumption that Lucas proves $\mathrm{Con}(T)$ informally, it merely follows that he can prove G_T in an informal sense as well. Benacerraf reminds us that all that follows from the first incompleteness theorem is that T, if consistent, cannot prove its Gödel sentence from *its* axioms according to *its* inference rules. Lucas' position is, of course, that his mathematical abilities are beyond that of a formal system, but it is clear that Lucas cannot prove T's Gödel sentence from T's axioms and inference rules either. So the claim that Lucas can mathematically outperform a formal system does not follow from the theorems for it is not at all clear that T is limited in its ability to conjure up informal proofs. Perhaps T can carry the Lucas argument on itself and convince itself of G_T like Lucas does. Benacerraf goes on to write, "To be sure, one might reply that no machine ... can be said to convince itself that formulas are true. But of course, if that's why [it] can't, then we hardly need Gödel's theorems to establish it."[9] In any case, his point is that Lucas' argument fails in demonstrating that Gödel's incompleteness theorems are irreconcilable with mechanism.

The nature of a Turing machine's proof

In a response to Benacerraf's criticism, Lucas argues that there is a fundamental difference in the way he and formal systems are able to construct proofs. He claims that a formal system, by *virtue of* being a formal system, is necessarily incapable of giving informal proofs. Given a formal system T, there is a Turing machine M_T that produces exactly the same set of sentences that T proves. So a theorem of T can be seen as an output statement of some formally computational model. The operations M_T performs in order to establish a sentence is "governed by [its] programme, and would correspond to *a* formal system."[10] So T's proofs cannot be informal in nature. Therefore,

[9] Paul Benacerraf, "God, the Devil, and Gödel," *The Monist* 51 (1967), 20.

[10] J. R. Lucas, "Satan Sultified: A Rejoinder to Paul Benacerraf," *The Monist* 52 (1968), 147.

if T is consistent, it is subject to Gödel's theorems and are shown to be incapable of proving either $\mathrm{Con}(T)$ or G_T.

Lucas' response to Benacerraf is unsatisfactory as his contention that he has the ability to outperform a formal system remains problematic. Lucas argues that it is impossible that a formal system convinces itself of something it does not prove formally. This is indeed the case for systems modeled by a deterministic Turing machine; their procedures for proving are governed by strict rules. However, one could conceive of a system that can be modeled by a stochastic Turing machine. Such a system could assess the probability of sentences being true and accept or reject them accordingly. Perhaps our intellectual abilities can be modeled by such systems. As input these models can take the same information that Lucas takes to show we are consistent and conclude that we are indeed *probably* consistent. They can perform the Lucas argument and be convinced that their Gödel sentence is probably true as well. Though Lucas may be convinced that he proves the consistency of the mind with absolute certainty, it is clear that he does not do this.

Revisions of Lucas' argument

Lucas' conviction of the superiority of human minds to machines is based on his claim that he can see the truth of the Gödel sentence while a machine cannot. A reconstruction of his argument has revealed the fallacy of assuming consistency of the mind, something that cannot be proven without sabotaging the entire argument. We must conclude that Lucas' argument indeed fails in demonstrating that the incompleteness theorems are incompatible with mechanism.

Several others have attempted to revise this type of Gödelian argument in which the human ability to see the truth of the Gödel sentence is pivotal. One well-known revision is made by the physicist Roger Penrose in his *Shadows of the Mind*. He argues that human mathematical reasoning is 'sound'; that is, the mind cannot prove false mathematical sentences.[11] He argues that mathematical theorems are 'unassailably true', a claim that has frequently been criticized in the mechanism debate.[12] The fallacy remains: for the same reason that Lucas cannot assert his consistency without undoing the rest of

[11]Roger Penrose, *Shadows of the Mind: A search for the missing science of consciousness*, (New York: Oxford University Press, 1994).

[12]For examples, see David J. Chalmers, "Minds, Machines, and Mathematics," *Psyche* 2 (1995), 11-13; or Drew McDermott, "Penrose is Wrong," *Psyche* 2 (1995), 66-82.

his argument, Penrose cannot assert his soundness without doing the same. A change in terminology does not bring forth a change in logic.

Many Gödelian arguments fail in disproving mechanism for the same reason. Evidently, this does not mean that mechanism is true. Perhaps Gödel's incompleteness demonstrate something different that is exclusive to the human intellect. In order to refute mechanism from these theorems we require an argument that does not rests on the soundness or the consistency of the mind. Storrs McCall provides precisely such an argument. In the next chapter, I will consider his argument as well as some possible objections to it.

3 The McCall Argument

Storrs McCall continues the search for a convincing anti-mechanist argument based on Gödel's incompleteness theorems. He contends that the unproven assumption of consistency of a system is necessary to derive its Gödel sentence.[13] Since the truth of the Gödel sentence is beyond what we can know, McCall attempts to find a different sentence that can be seen to be true by the human mind while fundamentally unprovable by formal systems. Such a sentence must exist, according to McCall, for Gödel's incompleteness theorems demonstrate that truth and provability diverge. His argumentation tactically refrains from making unwarranted assumptions about the consistency of systems or the human mind. McCall's argument thus has the potential to succeed where the Lucas argument fails.

In what follows I will examine his argument carefully. First I consider the particular sentence that is claimed to be true by Gödel's theorems, though unprovable to formal systems. It will become apparent that McCall misunderstand Gödel's proofs, which causes him to fail to provide a sentence only human minds can see to be true. Then I will focus my attention to McCall's central claims: that truth and provability diverge, and that this divergence cannot be understood by formal systems. After a careful analysis of these claims, it will be demonstrated that McCall's argument cannot be justified. He, too, fails in showing that mechanism is irreconcilable with Gödel's incompleteness theorems.

Reconstructing McCall's argument

McCall argues that "the domain of expertise of a Turing machine lies in the area of proof and provability, not in the area of truth. Human beings, on the other hand, are acquainted with both proof and truth, and also know of cases where the two diverge."[14] Let us consider McCall's motivation for this point carefully. Let T be a formal system, then by the first incompleteness exactly one of the following two cases must hold:

(1) T is consistent, then $T \nvdash G_T$;

(2) T is inconsistent, then $T \vdash G_T$.

[13] McCall, "Can a Turing Machine Know that the Gödel Sentence is True?", 526.
[14] Idem, 527

Considering the Gödel sentence, by construction, satisfies $G_T \leftrightarrow \neg\mathrm{Bew}_T(\ulcorner G_T \urcorner)$, it is indeed *true* when unprovable and *false* when provable. No illegitimate assumptions are made; these conditional statements hold categorically. It is simply a mathematical result that the truth value of the Gödel sentence is in opposition to its provability. Thus, McCall is indeed correct to conclude that truth and provability are not equivalent properties.

There must be some sentence that is *true but unprovable*. McCall then sets out to find one. Contrary to Lucas and Penrose, he contents that the Gödel sentence is not an appropriate candidate and he looks for his golden ticket elsewhere. Perhaps a conditional claim following from the incompleteness theorems, similar to statements (1) and (2), would suffice. Statements (1) and (2) in particular are not what McCall is after, since their formalization is a theorem of T. That is,

$$T \vdash \mathrm{Con}(T) \leftrightarrow \neg\mathrm{Bew}_T(\ulcorner G_T \urcorner).$$

It is the second part of the first incompleteness theorem that, according to McCall, belongs to the desired category:

(3) If T is consistent, then $\neg G_T$ is unprovable in T.

He argues that, firstly, (3) is a consequence of Gödel's first theorem and, secondly, that the formal counterpart of (3),

(3*) $\mathrm{Con}(T) \rightarrow \neg\mathrm{Bew}_T(\ulcorner \neg G_T \urcorner)$,

is unprovable in T. If his argument succeeds, we know of a particular sentence that can be seen to be true by the human intellect while being out of bounds of what a formal system can derive.

A true yet unprovable sentence

There are several problems with this argument. The first is that McCall does not provide a conclusive proof of (3*) being unprovable in T, simply stating that it is "unlikely" for it to be provable.[15] This, I think, does not make for a convincing argument.

Another problem is more obvious still: statement (3) is not exactly what follows from the first incompleteness theorem. Gödel explicitly required a

[15] Idem, 529.

formal system to be ω-consistent, a strictly stronger requirement than mere consistency, to conclude that $\neg G_T$ does not belong to its theorems. Though McCall attempts to show that the stronger requirement is unnecessary to establish the underivability of $\neg G_T$, Alexander George and D. J. Velleman point out that his proof makes tacit use of the assumption.[16] Paraphrasing McCall's proof,

> Suppose, for reductio, that T is consistent and that $\neg G$ is a theorem. Then there is a T-proof of $\neg G_T$, and, from the consistency of T we infer that there is no T-proof of G_T. At the same time, since $\neg G_T$ is equivalent to $\mathrm{Bew}_T(\ulcorner G_T \urcorner)$, from $T \vdash \neg G_T$ we derive $T \vdash \mathrm{Bew}_T(\ulcorner G_T \urcorner)$. Because the open formula $\mathrm{Bew}_T(x)$ weakly represents the property of being a T-proof, it follows that there is a T-proof of G_T, which completes the reductio.[17]

What McCall does not realize is that the truth of his claim, in particular the weak representability of a property, depends on $\omega-$consistency. To see why this is the case, consider what precisely follows from the assumption of $T \vdash \neg G_T$. From the construction of G_T, this is equivalent to $T \vdash \exists x \mathrm{Prf}_T(x, \ulcorner G_T \urcorner)$. However, without assuming ω-consistency, one cannot rule out the possibility that for all natural numbers n, $T \vdash \neg \mathrm{Prf}_T(n, \ulcorner G_T \urcorner)$ (see footnote 4 for a definition of ω-consistency). Thus, we may not conclude that there is an actual T-proof of G_T. Hence, the reductio is not completed.[18]

This fallacy makes it illegitimate to conclude that the human intellect knows statement (3) to be true. This results in the irrelevancy of claiming that its formal counterpart, (3*), is underivable in a formal system. Though this mistake by itself leaves the McCall argument severely troubled, there is another problem worth mentioning. One may wonder whether the argument applies to the following *true* statement:

---

[16] Alexander George and Daniel J. Velleman, "Leveling the Playing Field between Mind and Machine: A Reply to McCall," *The Journal of Philosophy* 97 (2002), 458.

[17] McCall, "Can a Turing Machine Know that the Gödel Sentence is true?" 529. Notations and definitions have been altered to better suit the style of this thesis; however, the deductions and argumentation are true to McCall's text.

[18] One might wonder whether Rosser's improvement, which showed that mere consistency suffices to demonstrate negation-incompleteness, might save McCall's argument. However, the fundamentally different construction of Rosser and Gödel sentences make it so that the argument does not work. The modified formal counterpart of (3*) where the Gödel sentence is replaced by the Rosser sentence is derivable in T. See George and Velleman, "Leveling the Playing Field," 461.

(4) If T is ω-consistent, $\neg G_T$ is unprovable in T.

As (4) is unequivocally true, perhaps *its* formal counterpart,

(4*) ω-Con(T)$\rightarrow \neg$Bew$_T(\ulcorner \neg G_T \urcorner)$,

can be shown to be unprovable in T. Then we would nonetheless have concrete evidence of some sentence that is knowable exclusively to the human intellect. All things considered, McCall was too optimistic in claiming to demonstrate a *true yet unprovable* sentence, as George and Velleman prove that (4*) is in fact formally derivable in T.[19]

Can a Turing machine differentiate between truth and provability?

There is a more essential point to McCall's argument that, regardless of failure to provide a *particular* sentence that can be seen to be true yet unprovable, may still be the successful to demonstrating that human mathematical reasoning is irreducible to a machine. His arguments rests on the claim that *true but unprovable* sentences exists at all, and that they form a category that is fundamentally out of reach of any machine. Indeed, McCall is correct to assert that the category exists as the incompleteness theorems demonstrate that truth and provability diverge for the Gödel sentence. However, to claim that this divergence can only be understood by the mind is problematic. Recall that the construction of the Gödel sentence is done in PA (or a formal system extending it):

$$\text{PA} \vdash G_T \leftrightarrow \neg\text{Bew}_T(\ulcorner G_T \urcorner).$$

By simple logic, the following equivalency holds as well:

$$\text{PA} \vdash (G_T \wedge \neg\text{Bew}_T(\ulcorner G_T \urcorner)) \vee (\neg G_T \wedge \text{Bew}_T(\ulcorner G_T \urcorner)).$$

The above holds for any arbitrary T, and since PA is an axiomatizable theory, a machine can be constructed that outputs all sentences of this form. Such a machine can recognize, similar to us, that truth and provability diverge for some sentences.

McCall may object to the significance of the formalization above. His original claim, *if T is consistent, then G_T is unprovable yet true*, is an informal claim that requires a semantic notion of truth. It can be argued that for a

[19] George and Velleman, "Leveling the Playing Field," 459-461.

formal system the mere assertion of a sentence is not equivalent to expressing the *truth* of that sentence via a truth predicate. If a formal system is capable of expressing such a predicate, it has sufficient tools to differentiate between truth and provability. Given a formal system T, a truth predicate may be denoted $\text{True}_T(x)$ and should satisfy $\text{True}_T(\ulcorner\varphi\urcorner) \leftrightarrow \varphi$ for all sentences φ of T's language. This approach is considered by McCall in his original paper. In it, he argues that Tarski's undefinability theorem demonstrates that this cannot be done.[20]

The undefinability of 'truth'

McCall claims, rightfully so, that all a Turing machine can know is that which it can prove: that is, what can derive from its axioms using its well-defined rules of proof. Truth, for a formal system, is something different entirely according to McCall. He claims that from Tarksi's theorem,

> The notion of "truth" in PA is ... on quite a different footing from that of "provability." The latter concept is represented by an open arithmetical formula; no analogous formula expresses (much less represents) the former, thus reinforcing the hypothesis that "truth," though meaningful to humans is a closed book to a [formal system].[21]

Tarski's undefinability theorem is an important result in mathematical logic which states that a formal consistent system is incapable of having a truth predicate for its own language.[22] To see why this is the case, suppose otherwise. Let T be a formal and consistent system extending PA with language $\mathcal{L}$, satisfying

$$T \vdash \text{True}_T(\ulcorner\varphi\urcorner) \leftrightarrow \varphi$$

for every $\varphi \in \mathcal{L}$. As T extends PA, the diagonalization trick applies to the negation of the truth predicate, so there is a sentence ψ such that T proves

$$\psi \leftrightarrow \neg\text{True}_T(\ulcorner\psi\urcorner).$$

[20] McCall, "Can a Turing Machine Know that the Gödel Sentence is True?" 530.

[21] Idem, 530.

[22] Alfred Tarski, "The Concept of Truth in Formalized Languages," in *Logic, Semantics, and Metamathematics*, ed. J. Corcoran (Oxford: the Claredon Press, 1983), 260.

However, by assumption T proves

$$\psi \leftrightarrow \mathrm{True}_T(\ulcorner \psi \urcorner).$$

Contradicting the consistency of the system, we must concede that such a truth predicate cannot exist. Since a formal system cannot define truth for its own language, McCall concludes that it cannot know of the category *true but unprovable* altogether. This is the essence of the McCall argument: human intellect fundamentally differs from what a formal system can derive due to its ability to understand that truth and provability do not always coincide.

McCall is correct when he states that Tarski's theorem prevents a formal (consistent) system of being able to express "truth" for its own language and therefore that knowledge of sentences being *true but unprovable* is limited. However, it is important that we are careful with interpreting the consequences of Tarski's theorem. A truth predicate cannot exist for a systems *own* language, yet it is well known that the theorem does not prohibit truth predicates for languages that are less rich. Let us illustrate this idea for a formal system T in language $\mathcal{L}$. Then there might be a language $\mathcal{L}^*$ included in $\mathcal{L}$ for which we can find a truth predicate, $\mathrm{True}(x)$, such that for all $\mathcal{L}^*$ formulas φ, $\mathrm{True}(\ulcorner \varphi \urcorner) \leftrightarrow \varphi$ is derivable in T. As long as the possibility exists that a formal system can define a partial truth predicate McCall's argument is in serious trouble. It cannot be said definitively that formal systems are incapable of recognizing that truth and provability in some cases diverge. As McCall fails to consider this possibility, his argument is unconvincing and mechanism may still be a legitimate theory of the mind.

A truth predicate for the Gödel sentence

A similar and more specific objection is raised by Panu Raatikainen in his critique of McCall's argument. Not only is the hypothetical existence of a partial truth predicate in the language of arithmetic a problem for McCall's argument, Raatikainen points out that Gödel sentences belong to a specific class of sentences for which a partial truth predicate has been defined.[23] The class of Π_1-sentences[24] consists of *universal* arithmetical sentences to which Gödel sentences belong. As proven by Smoryński, formal systems extending

[23]Panu Raatikainen, "McCall's Gödelian argument is invalid," *Facta Philosophica* 4 (2002), 168.

[24]To be more precise: Π_1-sentences are of the form $\forall x \varphi(x)$ or $\neg \exists x \varphi(x)$ where $\varphi(x)$ is a bounded arithmetical formula.

PA can express a truth predicate for Π_1-sentences without violating Tarski's undefinability theorem.[25] Let us denote such a predicate by $\mathrm{Tr}_1(x)$. Then for any formal system T extending PA, PA can prove that $G_T \leftrightarrow \mathrm{Tr}_1(\ulcorner G_T \urcorner)$. Hence, PA can prove

$$(\mathrm{Tr}_1(\ulcorner G_T \urcorner) \wedge \neg \mathrm{Bew}_T(\ulcorner G_T \urcorner)) \vee (\neg \mathrm{Tr}_1(\ulcorner G_T \urcorner) \wedge \mathrm{Bew}_T(\ulcorner G_T \urcorner))$$

for any arbitrary formalized T. Again, a Turing machine can be constructed to enumerate all such facts. As a result, it is able to differentiate truth from provability, even when the former is taken as a substantial property.

Concluding remarks

Though McCall's argument is indeed innovative and seems to be much stronger than the Lucas argument at face value, it is subject to severe problems that have not been rectified. McCall claims to have found a true sentence that is unprovable to formal systems, but George and Velleman have demonstrated that he is mistaken in even claiming it to be true. Moreover, the corrected statement *is* provable in formal systems. McCall therefore fails to provide an example of a sentence that is *true but unprovable*.

The stronger claim, that the category of *true but unprobavable* sentences exists at all yet is unrecognizable to machines, was examined next. McCall appeals to Tarski's undefinability theorem to argue that a machine is not equipped to understand "truth" in the same way that it can understand "provability". This argument is unconvincing since machine can have partial truth definitions and can therefore understand, in some cases, that truth and provability diverge. Moreover, the Gödel sentences were proof that the category exists, but machines can easily be equipped with a truth-predicate for Gödel sentences as well. Thus, a machine can recognize that truth and provability diverge for Gödel sentences. McCall's argument, therefore, fails in demonstrating any way in which the human intellect has superior mathematical abilities than machines.

[25] Craig Smoryński, "The Incompleteness Theorems," in *Handbook of Mathematical Logic*, ed. J. Barwise (Amsterdam: North-Holland Publishing Company, 1977), 843.

Conclusion

The two Gödelian arguments against mechanism that have been examined in this thesis are shown to be unsuccessful. Lucas claims the human mind is superior to a machine since it can see that the Gödel sentence is true. This claim, however, has been shown to be based on an illusion. In order to establish the Gödel sentence, we must be able to establish the antecedent of $\mathrm{Con}(T) \rightarrow G_T$, something that, as I have argued, is impossible in a way that does not compromise the argument altogether. Many Gödelian anti-mechanist arguments are based on the same illusion and fail for the same reasons. McCall's approach differs considerably: it is not the Gödel sentence that is true yet unprovable, but rather the more complicated *if T is consistent, then $\neg G_T$ is unprovable in T.* However, it has been shown that this argument is false as well. McCall furthermore appeals to the fact that truth and provability diverge by the incompleteness theorems, though his conclusion that this is knowable only to human beings has also been refuted.

While an analysis of Lucas' and McCall's arguments shed light only on a part of the mechanism debate, I think there is something more general to be said about Gödelian anti-mechanist arguments. The arguments we have considered, including that of Penrose, all share a noteworthy feature. They claim that human minds are superior to machines since they are able to prove, contrary to formal systems, some particular arithmetical sentence. It seems that it is safe to say that any Gödelian argument must take this form as Gödel's theorems demonstrate nothing more than that (consistent) formal systems are limited in this way. However, a claim to our abilities to prove some particular sentence will probably be far too inexact to be of value.

What type of sentence would we be able to prove that by Gödel's theorems is unprovable in a formal system? For Lucas and Penrose this statement was the Gödel sentence, G_T; for McCall is was a sentence constructed with G_T and $\mathrm{Con}(T)$. Moreover, I don't see how an anti-mechanist can use Gödel's theorems to argue for the human provability of some sentence that does not depend on G_T or $\mathrm{Con}(T)$. It is crucial, therefore, to be conscious of how these sentences are to be understood. Both $\mathrm{Con}(T)$ and G_T are mathematically defined using the provability predicate of system T. In non-mathematical literature they are often written as meaningful sentences. For example, $\mathrm{Con}(T)$ may be defined as "there is no sentence such that there is an T-proof of this sentence and an T-proof of its negation". While not incorrect, this definition is incomplete without a clarification of what constitutes an T-proof. Sure,

we know that a provability predicate for formal system T can be effectively constructed as Gödel provided a method to do so, but to actually construct a provability predicate for a system T requires knowing the axioms and rules of proof of the system. In other words: T must be given.

This is a considerable problem for the Gödelian anti-mechanist. We may know how to construct the provability predicate for, say, Peano Arithmetic, but the Gödelian does not claim that we are superior to formal systems because we can prove G_{PA}: certainly formal system PA + G_{PA} can prove the same. The Gödelian claims that we are superior because we can prove some particular sentence depending on G_T or $\mathrm{Con}(T)$ for any arbitrary formal system T. Ignoring all other problems such an argument may face, this claim is justified only if we can explicitly or effectively define a provability predicate for T. When T may be any arbitrary formal system, this is an unattainable task. Benacerraf makes a similar point in his criticism of Lucas' argument:

> If given a black box and told not to peek inside, then what reason is there to suppose that Lucas or I can determine its program by watching its output? But I must be able to determine its program (if that makes sense) if I am to carry out Gödel's argument in connection with it... If the machine is not designated in such a way that there is an effective procedure for recovering the machine's program from the designation, one may well know that one is presented with a machine but yet be unable to do anything about finding the Gödel's sentence for it.[26]

Benacerraf's point is that precise definitions are required for the notions of provability within a system. Any argument appealing to the capabilities of a mechanical model is unconvincing without knowledge of its program. It would simply be too hypothetical to yield any concrete results. Ignoring all other problems a Gödelian argument against mechanism may have, it must at the very least be precise in its claims about how any particular relevant model of a formal system can derive sentences. Granted, this is not an easy task. Benacerraf goes on to write, "In a relevant sense, if I am a Turing machine, then perhaps I cannot ascertain which one."[27]

Proving that the incompleteness theorems are incompatible with mechanism requires that the notion of provability is defined. Therefore, until

[26]Benacerraf, "God, the Devil, and Gödel," 28.
[27]Idem, 29.

such time that we understand the intricate workings of the human mind (or the formal system that allegedly models us exactly) well enough to give an explicit provability predicate for it, it is unlikely that mechanism can be refuted by the results of Gödel's theorems. The anti-mechanist may need to look beyond Gödel's work in order to support his claim.

References

Benacerraf, Paul. "God, The Devil, and Gödel." *The Monist* 51 (1967), 9-32.

Chalmers, David J. "Minds, Machines, and Mathematics: a review of *Shadows of the Mind* by Roger Penrose." *Psyche* 2 (1995), 11-20.

George, Alexander and Daniel J. Velleman. "Leveling the Playing Field between Mind and Machine: A Reply To McCall." *The Journal of Philosophy* 97 (2000), 456-461.

Gödel, Kurt. "On formally undecidable propositions of *Principia Mathematica* and related systems I (1931)." In *Collected Works I*, edited by Solomon Feferman, 144-196. Oxford: Oxford University Press, 1986.

Gödel, Kurt. "On undecidable propositions of formal mathematical systems (1934)." In *Collected Works I*, edited by Solomun Feferman, 346-371. Oxford: Oxford University Press, 1986.

Hájek, Petr and Pavel Pudlák. *Metamathematics of First-Order Arithmetic.* Berlin: Springer-Verlag, 1993.

Lucas, J. R. "Satan Sultified: A Rejoinder to Paul Benacerraf." *The Monist* 52 (1968), 145-158.

Lucas, J. R. "Minds, Machines, and Gödel." *Philosophy* 36 (1961), 112-127.

McCall, Storrs. "Can a Turing Machine Know that the Gödel Sentence is True?" *The Journal of Philosophy* 96 (1999), 525-532.

McDermott, Drew. "Penrose is Wrong." *Psyche* 2 (1995), 66-82.

Penrose, Roger. *Shadows of the Mind: a search for the missing science of consciousness.* New York: Oxford University Press, 1994.

Putnam, Hilary. "Minds and Machines." In *Dimensions of Minds*, edited by Sidney Hook, 138-164. New York: New York University Press, 1960.

Raatikainen, Panu. "McCall's Gödelian Argument is Invalid." *Facta Philosophica* 4 (2002), 167-169.

Rosser, Barkley. "Extensions of Some Theorems of Gödel and Church." *Journal of Symbolic Logic* 1 (1936), 87-91.

Smoryński, Craig. "The Incompleteness Theorems." In *Handbook of Mathematical Logic*, edited by J. Barwise, 820-865. Amsterdam: North-Holland Publishing Company, 1977.