Polskie Towarzystwo Semiotyczne

STUDIA SEMIOTYCZNE

Tom XXXIV • nr 1

PÓŁROCZNIK

HUMANS, MACHINES, AND GÖDEL

Warszawa • 2020

Założyciel „Studiów Semiotycznych” (*Founding Editor*):

Jerzy Pelc
(1924–2017)

Redakcja językowa:
Martin Hinton, Carl Humphries, Christopher Moore

Skład elektroniczny:
Dominik Dziedzic

Adres redakcji: Krakowskie Przedmieście 3, 00-047 Warszawa
e-mail: studiasemiotyczne@pts.edu.pl
http://studiasemiotyczne.pts.edu.pl/

ISSN 0137-6608; e-ISSN 2544-073X

CONTENT

STUDIA SEMIOTYCZNE (SEMIOTIC STUDIES), t. XXXIV, nr 1 (2020), s. 5–8
ISSN 0137-6608, e-ISSN 2544-073X
DOI: 10.26333/sts.xxxiv1.01

From the Editor

ROMAN KOSSAK [*]

PREFACE

Cantor's set theory combined with the development of formal logic changed the foundations of mathematics forever. In the late 1920s, David Hilbert, who was one of the most influential mathematicians of his time, was advancing a program the aim of which was to establish once and for all the validity of infinitistic methods that proved to be so powerful in various areas of classical mathematics. Hilbert's program suffered a fatal blow when Kurt Gödel announced his incompleteness theorems. Some, including John von Neumann, instantly grasped the importance of Gödel's results, but it took the mathematical community much longer to realize what the results meant for the foundations of mathematics and for mathematical practice. Craig Smoryński tells an interesting story about it in *Hilbert's Programme* (1988).

It was only in 1950's, after seminal work in proof theory and computability theory, by Ackerman, Bernays, Church, Gentzen, Hilbert, Kleene, Post, Turing, and many others, that one could say with confidence that we now know how to formalize mathematics. Equipped with the new conceptual framework, certain foundational issues became approachable, and one could hope to establish results about them with mathematical precision. In this vein, John Lucas in the 1960s and later Roger Penrose in the 1990s came up with arguments, based on Gödel's theorems, to show that mathematics as human activity cannot be reduced to a single algorithmic procedure, or, more poetically, that human minds are not machines. Prior to that, Gödel himself, in his Gibbs lecture in 1951, formulated and argued for what is now known as Gödel's disjunction:

[*] City University of New York, The Graduate Center. E-mail: rkossak@gc.cuny.edu. ORCID: 0000-0002-8475-5747.

> Either the human mathematical mind cannot be captured by an algorithm, or there are absolutely undecidable problems.

Gödel believed that both disjuncts were true, and was convinced that a rigorous confirmation of the disjunction could be given, but he could not see a way to do it for either of the disjuncts. Lucas and Penrose argued that the first disjunct holds. Recently, in a series of articles, Peter Koelner provided a formal framework to validate Gödel's disjunction (2016; 2018a; 2018b); but, after a thorough analysis, the arguments of Lucas and Penrose have been rejected by the logic community. Stanisław Krajewski's essay in this collection provides a detailed analysis of Lucas' proof and two proofs given by Penrose. It was the initiative of the editors of *Semiotic Studies* to invite mathematicians and philosophers to respond to Krajewski's essay and to comment on related issues from todays perspective.

While not much can be added the logical analysis of the arguments of Lucas and Penrose, the question of mechanization of mathematics gives rise to a discussion that touches upon central problems in the philosophy of mathematics. As computer-assisted proofs become routine, it is also relevant to current mathematical practice. In the June/July 2018 issue of the Notices of the American Mathematical Society, Jeremy Avigad gives a survey of recent advances in automated theorem proving, and in the conclusion he writes:

> The history of mathematics is a history of doing whatever it takes to extend our cognitive reach, and designing concepts and methods that augment our capacities to understand. The computer is nothing more than a tool in that respect, but it is one that fundamentally expands the range of structures we can discover and the kinds of truths we can reliably come to know. This is as exciting a time as any in the history of mathematics, and even though we can only speculate as to what the future will bring, it should be clear that the technologies before us are well worth exploring. (2018)

The role of proof in mathematics is not just to discover mathematical truths, but rather to provide insights into why this or that particular statement is true. Surely, such insights cannot be provided by a machine that only spits out true mathematical statements one by one; most of them are simply uninteresting. However, it would be a mistake to underestimate what machines can actually do. Stephen Wolfram, the founder of *Mathematica*, discusses this in his blog:

> At some level I think it's a quirk of history that proofs are typically today presented for humans to understand, while programs are usually just thought of as things for computers to run. Why has this happened? Well, at least in the past, proofs could really only be represented in essentially textual form—so if they were going to be used, it would have to be by humans. But programs have essentially always been written in some form of computer language. And for the longest time, that language tended to be set up to map fairly directly onto the low-level operations of the computer—which meant that it was readily "understandable" by the computer, but not necessarily by humans. But as it happens, one of the main goals of

> my own efforts over the past several decades has been to change this—and to develop in the Wolfram Language a true "computational communication language" in which computational ideas can be communicated in a way that is readily understandable to both computers and humans. (2018)

The articles in this issue can be divided into three groups. Krajewski's article, Yong Cheng's contribution, and a short note by Rudy Rucker, provide detailed mathematical analysis of Lucas-Penrose type arguments. In the second group, with articles by Arnon Avron, Stepan Holub, Panu Raaikiainen, and Albert Visser, the authors discuss the status and various methodological and technical problems of the anti-mechanist arguments. In essence: what does the problem of "minds vs. machines" really mean, and how can it, and how should it, be formulated? Moreover: How to evaluate the merit of arguments that mix formal mathematics and philosophical considerations? The third group consists of the articles that, while including issues from the other two groups, concentrate of more specific themes: an analysis of Georg Kreisel's observation that it does not logically follow from the fact that a formal system is subject to the second Gödel incompleteness theorems that there are absolutely no means available to prove its consistency (Jeff Buechner); Per Martin-Löf's proof that there are no absolute unknowables in constructive mathematics (V. Alexis Peluce); diagonal arguments and Chomsky's approach to linguistic competence as contrasted with arithmetic competence (David Kashtan); and the role in the anti-mechanist arguments of difficulties in capturing the nature of natural numbers in formal systems (Paula Quinon).

All articles in this issue, directly or indirectly, address the limits of mathematical knowledge. While we have a precise definition of provability in formal systems, the question of what is knowable is vague. In a series of recent papers, Peter Koelner approached this problem, by formalizing aspects of mathematical truth and knowability in a way that allows him to give a rigorous argument validating Gödel's disjunction. This theme is taken up in Yong Cheng's article in this issue.

Finally, Wilfred Sieg's article gives a historical account of the seminal contributions of Gödel and Turing that made possible all later developments partially described in this issue.

REFERENCES

Avigad, J. (2018). The Mechanization of Mathematics. *Notices of the American Mathematical Society*, *65*(6), 681–690.

Koellner, P. (2016). Gödel's Disjunction. In: L. Horsten, P. Welch (Eds.), *Gödel's Disjunction: The Scope and Limits of Mathematical Knowledge* (pp. 148–188). Oxford University Press.

Koellner, P. (2018a). On the Question of Whether the Mind Can Be Mechanized. Part 1: From Gödel to Penrose. *Journal of Philosophy*, *115*(7), 337–360.

Koellner, P. (2018b). On the Question of Whether the Mind Can Be Mechanized. Part 2: Penrose's New Argument. *Journal of Philosophy*, 115(9), 453–484.

Smoryński, C. (1988). Hilbert's Programme. *CWI Quarterly*, *1*(4), 3–59.

Wolfram, S. (2018). Logic, Explainability and the Future of Understanding. Retrieved from: https://writings.stephenwolfram.com/2018/11/logic-explainability-and-the-future-of-understanding/

STUDIA SEMIOTYCZNE (SEMIOTIC STUDIES), t. XXXIV, nr 1 (2020), s. 9–56
ISSN 0137-6608, e-ISSN 2544-073X
DOI: 10.26333/sts.xxxiv1.02

Essay

STANISŁAW KRAJEWSKI [*]

ON THE ANTI-MECHANIST ARGUMENTS BASED ON GÖDEL'S THEOREM

SUMMARY: The alleged proof of the non-mechanical, or non-computational, character of the human mind based on Gödel's incompleteness theorem is revisited. Its history is reviewed. The proof, also known as the Lucas argument and the Penrose argument, is refuted. It is claimed, following Gödel himself and other leading logicians, that anti-mechanism is not implied by Gödel's theorems alone. The present paper sets out this refutation in its strongest form, demonstrating general theorems implying the inconsistency of Lucas's arithmetic and the semantic inadequacy of Penrose's arithmetic. On the other hand, the limitations to our capacity for mechanizing or programming the mind are also indicated, together with two other corollaries of Gödel's theorems: that we cannot prove that we are consistent (Gödel's Unknowability Thesis), and that we cannot fully describe our notion of a natural number.

KEYWORDS: Gödel's theorem, mechanism, Lucas's argument, Penrose's argument, computationalism, mind, consistency, algorithm, artificial intelligence, natural number.

1. Introduction

Several philosophical consequences of the celebrated Gödelian incompleteness results have been indicated by logicians and philosophers. Here, only one issue is examined: namely, the alleged Gödel-based proof of the non-mechanical character of the human mind. In more modern terms, this equates with the refutation of the (strong) computationalist thesis identifying the mind with a computer. According to that thesis, the mind can be imagined as a program, where this need

[*] University of Warsaw, Faculty of Philosophy. E-mail: stankrajewski@uw.edu.pl. ORCID: 0000-0002-1142-8112.

not necessarily correspond to a (computational) mechanism; therefore, “computationalism” seems to be a more appropriate term. Nevertheless, for historical reasons, I will continue using the term “mechanism”. Ever since Gödel himself, logicians have argued—against the claims of many non-logicians, including philosophers and mathematicians—that anti-mechanism is not implied by Gödel’s theorems alone. The present paper aims to set out the logicians’ argument in its strongest form.

Recently, another problem relating to the computationalist thesis has appeared: our thinking, or at least some manifestations of our intelligent behavior, no longer seem to be limited to human beings, in that they can be present in computers or networks of computers, too. The question, then, is whether Gödel’s limitative results imply limitations regarding our abilities to mechanize intelligence. Here, again following Gödel himself, the answer would seem to be positive.

Even if it should not be, the controversy surrounding the value of the anti-mechanist corollaries of incompleteness results remains very much a live one, with scholars as prominent as Roger Penrose claiming, against Gödel, that the latter’s theorem proves the non-mechanical nature of the mind. This stance is also reiterated in popular expositions, such as Goldstein (2005). Indeed, the continuing widespread support for this claim provides one of the principle justificatory motivations for the present paper.[1] Here, the Gödel-based arguments for anti-mechanism, commonly referred to as the Lucas argument and the Penrose argument, will be reviewed once again. The refutations of both versions will be set forth in this context in a more explicit way than were those proposed by Gödel and, subsequently, by other leading logicians. Even so, the essence of these refutations was, in fact, revealed by Gödel himself. The present paper is based on Krajewski (2003), a book-length study in Polish (summarized in Krajewski, 2004) where some topics are treated much more extensively and a wider range of authors are quoted, but also reflects this author’s presentation (also in Polish) of the anti-anti-mechanism arguments (Krajewski, 2012), as well as two other papers that further refine this critique (Krajewski, 2007; 2015). Compared to earlier publications, there will be more stress here on the generality of the anti-Lucas and anti-Penrose theorems and, following (Krajewski, 2015), on ways to explain Penrose’s approach by identifying an additional premise that he implicitly adopted. I also find it important to endorse the corollaries that do follow from Gödel’s theorems: that we cannot prove that we are consistent, and that we cannot fully describe our notion of a natural number.

Section 2 contains some background. However, a standard knowledge of Turing machines, recursive functions, Church’s Thesis, and Gödel’s theorems will be assumed. To be specific, G_T is Gödel’s sentence for any (first-order) theory T that includes elementary arithmetic. For any T that is consistent and (minimally) sound, G_T is independent of T (unprovable and not refutable). Soundness means

[1] There exist, to be sure, competent presentations that avoid such errors, e.g., (Franzen, 2005; Berto, 2009).

semantical adequacy: provable formulas are true. For those wishing to avoid the inherently unclear notion of truth, Gödel introduced a notion of restricted soundness, referred to as ω-consistency: for no formula φ all of the following are provable in T: $(\exists x)\neg\varphi(x)$ and $\varphi(\mathrm{S}^{(n)}0)$ for all $n = 0, 1, 2, \ldots$; here, "$\mathrm{S}^{(n)}0$" denotes the n-th successor of zero—that is, the number n. Minimal soundness (the above principle being applied only to formulas with restricted number quantifiers) is called 1-consistency. G_T can be seen as a natural formalization of the statement that T is consistent. It can be expressed as a Π_1 formula: all the unrestricted number quantifiers are universal, and they all appear in front of the rest of the formula. Due to the Matiyasevich-Robinson-Davis-Putnam theorem, this statement can be expressed as the absence of solutions to a specific (dependent on T) Diophantine equation. According to standard accounts, G_T is independent and true. For those for whom the notion of truth is unclear, it would probably be easier to admit this notion for the purposes of the statement that there is no integer solution to a particular, logically simple equation.

In Section 3, the history of the anti-mechanist argument is sketched. In Section 4 the argument is reconstructed as a procedure performed in four steps, and each step is analyzed. Then, two main issues are discussed: the "dialectical" character of the argument and its algorithmic nature. Section 5 contains a general theorem demonstrating the inconsistency of anyone who systematically applies the Lucas-style argument, and Section 6 contains a similar theorem for Penrose-style arguments. In Section 7, Gödel's position is briefly described, including the well-known Gödel's Disjunction. In Section 8, another well-known claim, the impossibility of a rigorous proof of our consistency, is mentioned, and I name this assertion Gödel's Unknowability Thesis. Afterwards, a claim is presented to the effect that we human beings cannot fully define our (human) understanding of natural numbers.

2. Background

2.1. Mechanism

Historically, mechanism arose in the age of Enlightenment. Earlier, Descartes had come close, saying that animals are machines. Humans, according to him, were more than machines, as "there are no men so dull […] as to be incapable of joining together different words, and thereby constructing a declaration by which to make their thoughts understood; and that on the other hand, there is no other animal […] which can do the like" (Descartes, 1637, Part 5). At the same time, Descartes was sure that no mechanism could imitate specifically human behavior: "although such machines might execute many things with equal or perhaps greater perfection than any of us, they would, without doubt, fail in certain others from which it could be discovered that they did not act from knowledge […]" (*ibidem*). Yet a hundred years later, La Mettrie, a doctor who saw himself as a follower of Descartes, in his work *Man-Machine*, turned Descartes's argument upside down: he claimed that man *is* a machine, in both body and mind. The body was likened to a huge, ingeniously built clock. It is no surprise that he chose the clock for comparison, as this was the most complicated artificial mechanism known at the time. Thinking seemed to him "so inseparable from organized matter that it appears to be one of its qualities as much as is electricity, movability, non-penetrability, extension" (La Mettrie, 1747). At that time, almost 300 years ago, it was a matter of faith whether a machine could be constructed that would be like man—or that would actually *be* man. And, indeed, this remains an open question, despite the progress in robotics. It is not surprising that a hundred years ago the brain was compared to a telephone switchboard, the most complicated network in use at that time, while in our own time the comparison is made with a computer.

2.2. Artificial Intelligence

The ideology of Artificial Intelligence (AI) constitutes the modern version of mechanism as applied to the mind. We can discern two interpretations: either the computer is supposed to imitate the effects of our activities (the weaker thesis), or it should imitate the structure of our thinking—the way the mind operates (the stronger thesis). No involved analysis of the differences is needed here, as the argument based on Gödel's theorem has always been used to demolish even the weakest AI thesis. For a similar reason, we should not be troubled by the fact that no definition of the mind seems to be possible. We just need to take advantage of a few well-known effects of the mind's activity, and require no insight into its essence. Only some features of the mind are called for, and among these is the capacity to understand Gödel's theorem.

On the other hand, as we study the alleged refutation of the thesis that the mind is mechanical or can be simulated by a machine, we should be able to define what a machine is. For example, we would not accept as a machine a device with a little homunculus hidden inside it. We would accept computers, including their hitherto unknown versions. What, then, is a machine? A definition is difficult to formulate, though it may be easier than formulating a definition of the mind. However, we can happily refer to Church's Thesis. Information processing machines, whatever they are, present a product that can be described as a recursive function. So far, all attempts to define an abstract machine have produced concepts equivalent to recursive functions and Turing machines. Obviously, the equivalence here pertains to the results, not the way of operating. But this, fortunately, is just what the weaker AI thesis is concerned with.

The mechanist thesis in its fullest form amounts to the one advocated by La Mettrie: that the human being is a machine. A more restricted thesis concerns the mind only, while a still more restricted one applies only to mathematics. Ultimately, moreover, we arrive at the most restricted thesis of all, which is applied to the arithmetic of natural numbers (integers): that the operation of the mind in the field of arithmetic is mechanical.

Each of these theses can be expressed in a weaker version speaking not about the activities of man and the mind, but only the results of those activities. The weaker mechanist thesis admits the possibility that something essentially non-mechanical takes place there, but it claims that by using an appropriate machine we can simulate the mind so that exactly the same results are achievable. The weakest variant reads as follows: the operation of the mind in the field of arithmetic can be simulated by a machine.

To those theses we could add even more restricted versions, based on our knowledge of the shape of Gödelian formulas. Thus the weakest thesis could refer to the operations of the mind to the extent needed to establish the non-existence of solutions of Diophantine equations. It follows from all the other ones, so to refute it is to refute them all. According to Lucas and Penrose, their arguments refute all of the above theses of mechanism and AI.

3. The Anti-Mechanist Argument

Many people who have learned about Gödel's results have felt that they provide such a limitation on the capabilities of machines broadly conceived (i.e. computers and robots, as well as their networks) that the limitation cannot apply to humans. Consequently, it seems that a fundamental difference between the human mind and machines has been demonstrated. The basic idea is very simple indeed: if a machine produces mathematical truths, then it cannot produce the Gödelian sentence constructed for the totality of those truths without falling into a contradiction. On the other hand, we can prove that the Gödel sentence is true. Thus—hooray!—we are better than any machine.

3.1. The History of the Gödel-Based Argument

The first printed mention of some form of the argument can be found in Alan Turing's fundamental paper (1950). It was not a new idea even then, as is indeed clear from his presentation. Turing wanted to convince the reader that machines can think—or, rather, that they can perform certain functions that we normally associate with intelligence. He admits that "mathematical" arguments, in the sense of considerations based on Gödel's Theorem or directly on Turing's theorem, are relevant, as "it is argued" that they prove "a disability of machines to which the human intellect is not subject". We feel we are better, and the feeling is not "illusory", writes Turing, and adds, "I do not think too much importance should be attached to it" (Feigenbaum & Feldman, 1995, p. 22). What is this added remark supposed to mean? It seems that what Turing wanted to say was that the building of robots was such a worthwhile undertaking that it would remain so even if robots were subject to some limitations.

Even before Turing, and also around the same time, similar thoughts were expressed by Emil Post, one of the pioneers of modern mathematical logic. In 1941, the latter wrote that "[a] machine would never give a complete logic; for once the machine is made we could prove a theorem it does not prove" (Post, 1941, p. 417). He claimed that he had entertained a thought of this sort already in 1924. Only later did he take Gödel's results into account. Post's paper was published much later, in the anthology of Davis (1965). The quoted sentence is not a straightforward expression of the thesis that the mind is not mechanical, but we can see that this is suggested by the phrase "we could prove".

At the end of his exposition of mathematical logic, Rosenbloom wrote that Gödel's theorem shows that "some problems cannot be solved by machines, that is, brains are indispensable" (Rosenbloom, 1950, p. 208). Man, he says, "cannot eliminate the need to use intelligence" (p. 163). Similar in spirit, only much more comprehensive and penetrating, are the considerations put forward later by Douglas Hofstadter (1979) in his bestseller, which served to make the general public aware of Gödel's results.

Before Hofstadter, the most popular exposition of Gödel's achievements for a wider public was that available in the book by Nagel and Newman (1989). The authors write there that "the brain appears to embody a structure of rules of operation which is far more powerful than the structure of currently conceived artificial machines [...] the structure and the power of the human mind are far more complex and subtle than any non-living machine yet envisaged" (Nagel, Newman, 1989, pp. 101–102). The reservations expressed by the phrases "currently conceived" and "yet envisaged" testify to the authors' caution. It could seem that their approach was manifesting a certain hesitancy as regards the thesis concerning the non-mechanical character of the mind, in that it allows for the appearance of machines in a new, hitherto unknown, sense; Gödel's method would not apply to those machines, and they could, in fact, be equivalent to the mind. However, the authors refrain from drawing this conclusion. Their attitude is also apparent in their response to the criticism of Putnam, who wrote that theirs was a "misapplication of Gödel's theorem, pure and simple" (Putnam, 1960a, p. 207). According to them, Putnam "dogmatically" assumed that every conceivable proof of the consistency of a machine hypothetically equivalent to human mind could also be constructed by the machine (Nagel and Newman, 1961, p. 211). This remark seems to mean that for Nagel and Newman, some capabilities of the mind are assumed to be—or at least are allowed to be—fundamentally non-mechanical. This early controversy makes it clear that our attitude to Lucas's argument may depend strongly on a basic assumption about whether or not it is possible for a machine to imitate arguments created by the mind.

The debate was continued by, among others, Kemeny (1959) and Smart (1960). In the 1950s, more and more analytic philosophers saw the anti-mechanist consequences of the limitative theorems as quite apparent, though probably only a few would swear that the argument contained no mistakes. It was Lucas who, with no hesitation whatsoever, presented the allegedly indubitable mathematical proof of man's superiority over machines—and even over matter.

The anti-mechanist argument was by no means universally accepted. On reflection, Post had fundamental doubts: "The conclusion that man is not a machine is invalid. All we can say is that man cannot construct a machine which can do all the thinking he can" (Post, 1941, p. 423). Later, many authors would draw attention to the weak points of Lucas-style arguments. As a matter of fact, amongst mathematical logicians the currently dominant view is that Lucas's argument is wrong. In addition to Gödel himself saying so in his 1951 Gibbs lecture (though this analysis was published much later), the first published critical mentions of Lucas's argument (which in fact preceded Lucas's paper) were Putnam's (1960) and (1960a). Boolos called them "classic" (Boolos, 1995, p. 254). Criticism was voiced by, among others, Quine, Benacerraf (1967), and Wang (1974). Later, criticism was directed against Penrose's version of the argument; among the most important papers were those by Feferman (1995) and Putnam (1995). Further criticism was offered by several logicians, for example Shapiro (1998) and Lindström (2001). A recent account of the debate is available in the collection of papers edited by Horsten and Welch (2016).

The argument based on Gödel's theorem retains its "mystical" charm. Many a philosophically minded scientist labors under its spell—as, increasingly, do other authors who refer to Gödel in order to state general theses not just about the mind, but also the limits of rationality, the incomprehensibility of the world, etc.[2] For some, the motivation is *de facto* religious: a desire to confirm with mathematical rigor the existence of the soul and free will. This is explicit in Lucas's later book (1970).

Roger Penrose, an outstanding mathematician and theoretical physicist, developed his own version of Lucas's argument in his books *The Emperor's New Mind* (1989) and *Shadows of the Mind* (1994). His position remains scientific: he speculates that the quantum-mechanical level can provide an explanation of the non-mechanical character of the mind and consciousness. According to Putnam, Penrose "mistakenly believes that he has a philosophical disagreement with the logical community" (Putnam, 1995, p. 370).

3.2. Two Ways of Criticizing Lucas's and Similar Arguments

Although logicians mostly agree that Lucas's (and also Penrose's) argument must be rejected, one must admit that a certain disconcerting ambiguity keeps on arising. There is more than one way to demonstrate the error in the Lucas or Penrose arguments. Two main approaches are used, both well summarized by John Burgess. For some, "the mistake lies in overlooking the possibility that it might in actual fact be the case that the procedure generates only mathematical assertions we can see to be true, without our commanding a clear enough view of what the procedure generates to enable us to see that this is the case". (Burgess, 1998, p. 351) For others, the error results from the fact that "even if we do see that the procedure generates only mathematical assertions we think we see are

[2] Chapter IV of the present author's book-length study in Polish (Krajewski, 2003) treats this phenomenon at length.

true, it might be rational to acknowledge human fallibility by refraining from concluding that the procedure generates only mathematical assertions that are in actual fact true" (Burgess, 1998, p. 351). To put it in a simpler and more picturesque way, the first line of attack reveals that it is not excluded that we are consistent machines but don't know it, and the second line shows that it is not excluded that we are inconsistent machines. The first method was introduced by Gödel, while the second—though also mentioned by Gödel—was made known by Putnam.

This ambiguity engenders a perplexing consequence: no criticism of Lucas's argument seems definitive. The first method assumes our consistency, and the other allows for the opposite to be the case. The assumptions contradict each other, so a supporter of Lucas can use this to say that the matter is not settled, since the opponents cannot agree among themselves. Still, the two methods taken together constitute a strong refutation: either we are consistent or not, and in both cases Lucas is wrong.

In this paper, both approaches will be taken into account, and in addition Lucas's argument will be refuted in yet another way: without assuming anything about our, or Lucas's, consistency, we will show (in Section 5.2) how every Lucas-style argument leads to either a vicious circle or a contradiction.

It is important to stress that all methods of refuting Lucas- and Penrose-style arguments are based on the insights expressed by Gödel himself, especially in 1951. (For more details, see Section 7 below.) According to the one-sentence summary of the argument given in (1951) that Gödel presented to Wang in 1972,

> [O]n the basis of what has been proved so far, it remains possible that there may exist (and even be empirically discoverable) a theorem-proving machine which in fact is equivalent to mathematical intuition, but cannot be proved to be so, nor even be proved to yield only correct theorems of finitary number theory. (Wang, 1974, p. 324; 1996, pp. 184–185)[3]

The present paper may be seen as constituting a somewhat extended footnote to the above sentence.

4. Analysis of the Gödel-Based Arguments

4.1. Steps (L1)–(L4)

Lucas's argument reads as follows: no machine is equivalent to the mind, because the mind can recognize the truth of the Gödelian formula for the machine, while a machine cannot do so—due to Gödel's theorem—without being inconsistent, in which case it would certainly not be equivalent to the mind. To perform a critical analysis of Lucas's argument, we must present its main points, or

[3] The term "finitary" has its proper meaning in the framework of Hilbert's program. Here it means the Π_1 statements of elementary "finite" number theory.

reconstruct it. While some degree of arbitrariness is unavoidable, my version, to the best of my knowledge, is faithful and accurate. It can be presented as four simple steps, from (L1) to (L4). The division into steps makes it much easier to incorporate in an orderly fashion all the considerations and critical points made in the literature. The aim is to "out-Gödel" the machines.

> **(L1)** First of all, we can see that machines—referred to by Lucas as "cybernetical machines"—are necessarily equivalent to formal systems. Each machine *M* has a definite finite number of states and instructions, and therefore corresponds to a specific formal system *S* of the kind studied in logic: *S* is given by axioms formulated in a specific formal language and by formal rules of inference. A calculation, or a sequence of operations performed by *M*, corresponds to a formal proof in *S*.
>
> **(L2)** If the machine *M* models the mind, it "must include a mechanism which can enunciate truths of arithmetic". The formulas *M* can "produce as being true" correspond to the theorems of *S*.
>
> **(L3)** Now, we can use Gödel's technique to construct a formula G that is not provable in *S*—i.e. not a theorem of *S*. We assume, of course, that *S*, or at least its arithmetical part, S_{ar}, is consistent. (Otherwise, G is a theorem, since in an inconsistent theory every formula is derivable using classical logic.) If *S* were inconsistent, it would obviously be inadequate as a model of the mind. Thus, due to Gödel's theorem, *M* cannot produce G as being true.
>
> **(L4)** On the other hand, we can see that the formula G is true. We can follow Gödel's proof and see that G is not a theorem of *S* and that it is true. Its truth is a consequence, even an expression, of its unprovability in *S*. We, our mind, can do something that *M* cannot. It is impossible to simulate all of the mind's capabilities at once. The mind is not equivalent to *M*, so it is equivalent to no machine. "The Gödelian formula is the Achilles' heel of the cybernetical machine" (Lucas, 1961, p. 116).

These four steps constitute a careful rendering of the argument proposed in Lucas (1961). The case has not changed since then. No essentially new elements of logical reasoning appear in his subsequent publications containing replies to criticism—i.e. Lucas (1968) and (1970), followed by Lucas (1996; 1997; 1998). To be sure, various points are discussed and some aspects are emphasized: for example, the "dialectical" character of the argument (see Section 4.6 below). In a later book he briefly repeats the Gödelian argument, noting only that it is "highly controversial" (2000, p. 219).

Essentially the same argument is presented by other authors—most notably Penrose (1989). Later, in his (1994) and (1996), the latter presented a modified version as well: one which includes a defense against critical voices and takes into account Gödel's own position. (See below, Section 6.)

However, each step in Lucas's reasoning can be questioned. In the discussion below, I analyze each of points (L1) to (L4) in turn. Then I consider Lucas's main line of defense, the “dialectical” nature of the argument. It turns out that the initially disregarded problem of consistency is fundamental. Finally, I present a theorem demonstrating that the threat of inconsistency is fatal to both Lucas’s original argument and every argument of a similar character, even when the concept of truth is not utilized.

4.2. Re (L1): Must Machines be Equivalent to Turing Machines?

Step (L1) seems to be the least controversial of the four. A machine that has a finite number of states and instructions, and operates sequentially—one operation after another—is essentially equivalent to a Turing machine. To be more precise, Turing machines constitute mathematical idealizations of those physically possible machines because they disregard all practical limitations: in using Turing machines, we admit a fixed but arbitrary (that is, limitless) number of states and an arbitrary number of instructions, as well as a boundless amount of input (so that the number of the states or instructions or the size of the input can even transcend the number of elementary particles in the universe, according to current physics). We also make another important idealization: we assume that the tape, or memory, of the Turing machine is (potentially) infinite. The output of every such machine can be described as the totality of theorems of a certain formal system. To prove this, it is enough to note that the output is a recursively enumerable (r.e.) set—and that, due to Craig’s lemma, each such set of elementary arithmetical sentences is recursively axiomatizable in the standard logical calculus. Thus, if Lucas’s argument—that is, its remaining points—were correct, we would agree that the mind is equivalent to no idealized machine, as the mind beats each such machine at least in some respect: so, *a fortiori*, the mind beats each real machine. That conclusion depends upon the assumption that there are no machines of a different nature, ones not reducible to Turing machines. This is essentially Church’s Thesis. Is it incontestable?

It seems that the gradual progress made possible by parallel processing, genetic algorithms, neural nets, and machine learning brings no breakthrough: the class of computable functions remains the same. Of course, we are considering idealized computability, without limitations of time, space or memory. If we were to consider practical computability, new kinds of machines would make more functions practically computable. Yet with Lucas’s argument, we are dealing with computability in principle, not in practice.

How does a mind emerge? So far, we have known only naturally created minds; but are we sure that above a certain level of complexity, a machine cannot acquire a mind? Even Lucas admits this possibility. However, in such a case, he claims, “it would cease to be a machine” (Lucas, 1961, p. 126). On this approach, the controversy over mechanism would turn, at least in part, into a disagreement over the use of words. To preserve the real problem, let us consciously and explicitly assume that to be a machine means to operate according to rules that can be reduced to steps equivalent to those described by Turing. In applying this to the problem of mechanism, we should beware of a circularity: if we simply assume that the mind, which is self-conscious, does not operate according to those rules, then we assume what we are supposed to prove by means of Lucas's argument, and the whole business connected with Gödel's theorem becomes superfluous. To avoid this, we should assume as little as possible about the nature of the mind. We shall therefore accept only those features clearly discernible on the basis of introspection. (For an example, we may refer to the diagonal construction, in which we treat as obvious the fact that from a recursive sequence of recursive functions we can effectively form a diagonal function that is also recursive.)

To sum up, step (L1) can be confirmed in the sense that it, and thereby the whole of Lucas's argument, can apply to a machine *M* belonging, at least, to the very extensive class of machines that—considered as idealized structures—are equivalent to Turing machines. We can assume that the input is absent or fixed, or is even itself recursively enumerable. Inputs that are not recursively enumerable must not be allowed, because in that case the non-recursive complexity of the input could be expressed in the output. An input of sorts is mathematically unnecessary, because it could be positioned as a part of the (program of the) machine. However, we will allow for it, as it may prove necessary when considering the "dialectical" character of Lucas's argument.

4.3. Re (L2): What Does "True" Mean for a Machine?

The machine must qualify some output expressions as "true". Following Lucas, one can say that they are "produced as being true". While this manner of speaking is not particularly neat, at first glance it seems to be innocuous. It is, however, perceived as an equivocation by Benacerraf (1967), Wang (1974) and, in a more detailed treatment, Slezak (1982). The point is that we use at the same time an expression suitable for a machine ("produce") and an expression proper to humans ("true"). We must describe an act that the mind—and no machine—can carry out, so it must fit both the machine mode (hence the cold terms "produce", "generate", "print", or the matter-of-fact "output") and human perception, which includes understanding and acceptance (hence "true", "ascertain", etc.). The equivocation is not due to carelessness; it is, instead, inherent to the foundations of an argument that is supposed to consider machines and humans at the same time, but never allow their identification. "Hence the (Lucas) argument requires the conflation of truth and provability to reach its conclusion" (Slezak, 1982, p. 45).

If we speak about machines as counterparts to formal systems, then it is enough to talk about (formal) derivability. The notion of truth is not needed as a prerequisite to state Gödel's theorem; it is enough to say that a consistent system is (syntactically) incomplete: i.e. for some formula, neither it nor its negation is derivable in the system. Gödel's theorem makes sense on the syntactic level: to apply it to a theory *T* we do not even need to know what "true" means when applied to *T*'s formulas.

There seem to be two ways of overcoming the equivocation—understood as the use of truth and derivability in the same statement. First, perhaps the notion of truth can be applied to machines? Second, in the context of Lucas's argument, maybe we can dispense with truth altogether?

It would be incorrect, if tempting, just to assume that a machine cannot use the notion of truth and other semantic concepts. Possibly, further scientific progress will lead to an increasing level of sophistication on the part of computers in the area that, for us, constitutes the realm of meaning and sense. If we assume that "genuine" truth does not apply to machines, but does apply to humans, then Lucas's argument is completely dispensable, because we are simply assuming our superiority over machines, which is the thesis that was to be demonstrated.

As much as it is incorrect to assume our superiority over machines, it would be wrong to refute Lucas's argument by, again, merely assuming that machines can understand, and that when they are developed far enough the whole semantic realm will emerge automatically—in other words, by supposing that "the Chinese palace", due to its size, will overcome the limitations of "the Chinese room". Fortunately, we need no such assumption to continue our analysis.

While analyzing the argument of Lucas we should be neutral towards the problem of the applicability of the concept of truth to the relations between linguistic objects and machines, both present and future. In the present context, to make the Lucas argument as easy-going as possible (and then to demolish it), we can assume that the machine either has access to truth or just pretends that it does.

We can assume that the machine has a green light that lights up only when the output expression is "produced as being true". Rather than truth itself, we simply have a green light pretending to correspond to truth. Clearly, rather than the suggestive light, we can assume that the output expression is accompanied by some other special symbol indicating "truth". This is done by Penrose (1994), in his version of the argument; yet he also begins by saying that the purported machine "ascertains truths". Then a little star is used as the "imprimatur" symbol. It is enough to use the device for arithmetical formulas. Whatever their truth means to us, whatever it may "mean" for a machine, we are left with the problem of whether Gödel's theorem excludes the existence of a machine that lists precisely those arithmetical formulas that can be perceived as true by humans.

We have just shown that in (L2) the reference to truth is not necessary. Later, it will be shown that we can allow the anti-mechanist to reformulate the argument so that the notion of truth is not used at all, but the argument remains bound to collapse.

4.4. Re (L3): The Consistency of a Machine and of a Human Being

The construction of the Gödelian formula for the relevant theory is the key point in Lucas's argument itself, and in its other variants. If out-Gödeling is not carried out as indicated in (L3), reference may be made to a formula expressing consistency, or another incompleteness result can be utilized—in particular, Turing's theorem, as, for example, Penrose does. All these approaches are basically equivalent.

It is not hard to see that two facts undermine the philosophical significance of Lucas's argument—though Lucas (1961) hardly showed any awareness of those facts, and he also clearly underestimated their impact in later works. The first fact is that the method of constructing Gödel's formula is algorithmic, and thus in a broad sense mechanical; the second is that its application depends on the consistency of the theory for which the formula is constructed. Leaving the first point, the algorithmic nature of out-Gödeling, for later, let us take up the second issue. The reasoning performed in step (L3) can be divided into two cases:

> **Case I**: The theory *S* is consistent. In that case the Gödelian formula is used to out-Gödel the machine *M*.
>
> **Case II**: The theory *S* is inconsistent. In that case machine *M* is disqualified (as a model of the mind).

The main difficulty is how to distinguish Case I from Case II. Before considering this problem, let us note that Case II is not itself as unproblematic as is claimed above.

If a system were to be equivalent to the mind, it would necessarily be consistent, says Lucas. Why? Because we are rational. While we commit mistakes, rationality means logic, and this means avoiding contradictions. If we believed in two contradictory sentences, we would infer arbitrary statements. This is a way to affirm our rationality, but serious doubts remain. After all, we hardly infer an arbitrary sentence as a consequence of our beliefs, even though we often happen to fall into contradictions: we change opinions, tend to say "yes" and "no" at the same time, and find ourselves being reminded by others that we have just said something quite the opposite of what we said sometime earlier. What is more, although our minds seem very similar to each other, our opinions are often not: people with the same degree of rationality, and with similar knowledge, are sometimes convinced of the truth of opposing propositions. Clearly, for us—that is, for our minds—contradiction does not lead to the acceptance of every sentence. (And there exist logical systems that formalize such situations.)

Lucas disposes of the problem in two ways. First, jokingly: Humans are inconsistent? Well, "certainly women are, and politicians" (Lucas, 1961, p. 120). Let us keep this opinion in mind. Second, our inconsistencies are temporary, because once we learn about them, we correct them. "They correspond to occasional malfunctioning of a machine" (*ibidem*, p. 121) rather than to a genuine inconsistency. We are fallible, but self-correcting. This sounds convincing, but the issue does not stop here.

While we do indeed try to correct mistakes, we may still be fundamentally inconsistent. Could not some principles of thought lead to contradictions, just as soon as they are used in particularly unfavorable circumstances? How could we exclude this prospect? There are examples of contradiction in the thought processes of outstanding thinkers—and not just philosophers: even the greatest mathematicians have committed mistakes and created contradictions. What is more, according to William Byers (2007), inconsistencies are unavoidable, and also fruitful, in mathematics. Even logicians, who are particularly sensitive to the danger of contradictions, are not immune. The example of Frege is well known: his system of logic turned out to be inconsistent. And the danger has not disappeared. One can imagine that a contradiction arose, but mathematics continued to function as smoothly as ever, without difficulty, in normal domains and applications. Actually, precisely this did happen when the set-theoretical paradoxes appeared over a hundred years ago.

Although we cannot exclude a worst-case scenario—in which a contradiction arises and nobody knows how to eliminate it—it is beyond doubt that mathematics must not abandon the struggle for consistency. Consistency, even when we cannot be absolutely sure of it, is for mathematics something like a regulative idea in Kant's sense. Consistency in this sense guides all of our intellectual endeavors that are subject to the rigors of logic. In some fields, it is possible to overcome contradictions by pointing to the metaphorical character of the expressions involved (e.g., "I am myself and I am not myself"). Nevertheless, in the realm of natural numbers contradiction proves fatal.

Lucas, Penrose, and all those who employ Gödel's theorem to refute mechanism or computationalism, as well as Gödel himself and many others, assume that our mind is (i.e. we are) fundamentally consistent—and often, also, that we are fundamentally sound. However, it is one thing is to believe this and another to know it for sure. The fact is we cannot know such a thing with absolute certainty. In other words, we cannot demonstrate it in, to use Penrose's terms, an unassailable manner. This makes sense independently of Lucas's argument. (See Section 8.1 below.)

And what happens, let us ask, if we are not consistent? In that case, one could say, everything would be provable. This is, however, unconvincing, writes Wang (1974, p. 319). We do not function as a Turing machine, even if, deep down, something equivalent to a Turing machine underlies our functioning. Also, we are back with the problem of hidden inconsistency here. As with those large computer programs that contain bugs but function well in regular applications, contradiction, too, can be hidden or indirect and provoke no destructive consequences in normal life. Perhaps, then, we are inconsistent? Maybe we are inconsistent machines?

While the conclusion that we are really, hopelessly inconsistent cannot be excluded, it is very implausible to many people, including myself. Lucas is right that any proper modeling of thinking must contain, in some way, propositional calculus and elementary arithmetic, including the belief in the consistency of arithmetic. I also agree with Lucas that a serious acceptance of the idea of the unavoidable inconsistency of our mind reflects irrational views that make rational polemics with mechanism impossible (Lucas, 1996, p. 121–122).

It should not be surprising that we humans are not able to answer all questions concerning our mind. The statement of consistency has a special status: we really do seem to arrive at a positive answer just through contemplating our own minds. It is beyond doubt, though, that we can be mistaken. As explained before, even the sharpest minds can commit errors. In that case, out-Gödeling leads to another inconsistency. In fact, it will be shown below (in Section 5) that every procedure similar to out-Gödeling inevitably leads to a contradiction.

If we assume our fundamental consistency, then either (a) this is not formally expressible, or (b) it is, but in that case it is not provable (unless the proof is by methods not susceptible to formalization), as will be shown in due course in Section 8.1. In the case of (a), we basically assume that the mind is not a machine, while in that of (b), we do not exclude it being one. If we choose (a), then the aim of Lucas and like-minded thinkers—that of demonstrating that humans are better than machines—is achieved; however, the argument is circular, and we add little to the initial conviction that evidently we are not machines. Much the same has been observed by many commentators; for example, in connection with the version proposed by Penrose, Minsky says: "In effect, it seems to me, Penrose simply assumes from the start precisely what he purports to prove" (Brockman, 1995, p. 256). If, on the other hand, we opt for (b), then the analysis of Lucas's argument must be carried further.

4.5. Re (L4): How Do We Know the Truth of Gödel's Sentence?

Step (L4) consists in the realization that we see the truth of the formula G. Lucas invoked the phrase often exploited by believers in the metaphysical consequences of Gödel's theorem, asserting that while G is not provable (derivable) in the system in question, "we, standing outside the system, can see (it) to be true" (Lucas, 1961, p. 113). Some people think we are talking here about truth in a special sense. Standing outside a formal system would then correspond to some sort of extraordinary fact: one that mysteriously enables us to grasp unusual truths. These truths must be atypical, they would seem to think, if they cannot be proven even within a very strong system *S*. Our power to "see truth" thus acquires a quasi-mystical character. This, I believe, is a major source—possibly the main source—of the attractiveness of Lucas-style arguments. Yet the position is surely misguided. The sheer fact of being "outside the system" affords us no mysterious advantage, even though global properties of formal systems do exist. The truth of G is not specific; G is true in a normal mathematical sense, much as the statement that a given equation has no solutions is true.

Rather than explicating these points in more detail,[4] let us observe that even if the theory is consistent, we may be unable to know this. The problem, thus, is to determine the truth of $Cons_S$. Even when the output *S* of the machine that Lucas's argument is aimed at dealing with is consistent, we can lack sufficient

[4] This is done in (Krajewski, 2003) and, e.g., (Franzen, 2005).

grounds to know this. To ascertain the consistency of a theory can be very difficult. For instance, take Quine's set theory NF. We do not know whether it is consistent; therefore, we cannot tell if the arithmetical sentence $Cons_{\mathrm{NF}}$ is true. No amount of "standing outside", of following the course of the proof of Gödel's theorem, of thinking at different levels at the same time, can help here. Even though the formula $Cons_{\mathrm{NF}}$ is arithmetical, its truth is difficult to settle, because it codes a property involving the whole of the theory.

In regard to (L4), we have noted that the truth of G for S_{ar} is a consequence of our assumption concerning consistency—rather than of some unusual insight. The problem of the truth of Gödel's formula (as distinct from the unquestionable truth of Gödel's theorem) boils down to the question of whether we know that the theory for which the Gödelian construction is made is consistent. We need to know that the machine *M*, or theory *S*, is consistent. Still, even if it is, says Putnam, we can be unaware of this reality.

We turn now to the two most fundamental and decisive ways of criticizing Lucas's argument: first, that it is impossible to determine in general terms precisely when Case I or Case II applies, and second, that the trick utilized by Lucas can also be carried out by some machines themselves.

4.6. The "Dialectical" Character of Out-Gödeling

In a relatively recent paper, Lucas deploys an argument against the claim that in order to know that the Gödelian formula is true one must know the consistency of the corresponding theory. He states that "Putnam's objection fails on account of the dialectical nature of the Gödelian argument" (Lucas, 1996, p. 117). This is his favorite argument, traceable right back to his original 1961 paper and stressed as the central point in Lucas (1968), which is an answer to his critics—in particular Benacerraf (1967). The point is that his argument is not a normal argument demonstrating a thesis, but is instead a "dialectical", or conditional, argument: if somebody claims that a machine is equivalent to the human mind, then it is shown to him that he falls into a contradiction.

Let us accept the dialectical character, in this sense, of the argument. In fact, the points (L1) to (L4) are consistent with this interpretation. Why, however, should it be the case that it overcomes Putnam's criticism that we may be unable to know that the relevant theory is consistent, even if it is?

In the argument conceived as a game, the opponent—let us call him or her "the mechanist"—indicates some machine (cf. L2) as being equivalent to the human mind (in the realm of arithmetic), and Lucas responds by pointing to the appropriate Gödelian formula (cf. L3 and L4). In the game, the consistency of the proposed machine should be granted: "The consistency of the machine is established not by the mathematical ability of the mind but on the word of the mechanist" (Lucas, 1996, p. 117). Thus the mechanist is only required to present consistent machines M (i.e. those machines for which the corresponding theory *S* is consistent). Yet can we really impose such a requirement?

One major problem with doing so stems from the fact that there is no decision procedure for determining consistency. Therefore, it is not only difficult on a practical level, but also theoretically impossible to have an algorithm that always correctly decides whether (the set of arithmetical sentences produced by) a given machine is consistent. To be more precise, if M_1, M_2, …, M_n, … is an effective listing of all Turing machines, then the set C, C = $\{n\colon M_n \text{ is consistent}\}$ of all indices of consistent machines is not recursive. Moreover,

Fact: C is not recursively enumerable.

A proof of the Fact can be based on Gödel's Theorem. If C were to be r.e., then so would be the set D = $\{G_n\colon n \in C\}$ of all Gödelian formulas for consistent theories $T(M_n)$ corresponding to machines M_n. But then, for some k, we would have D = $T(M_k)$. D consists of true sentences, so it is consistent, which means that $k \in C$. Given the definition of D, G_k is in D, and so in $T(M_k)$, which contradicts Gödel's Theorem. The argument based on the above Fact was first used in the context of Lucas-style reasoning in Wang (1974), before being further strengthened in Bowie (1982) and Krajewski (2003). To require the mechanist to present only consistent machines means that we assume he or she has "superhuman" capabilities—or, at least, non-mechanical capabilities. This would mean that in order to prove the non-mechanical character of the mind, we would have to assume that the human mind is non-mechanical: an obviously circular way of thinking!

Lucas tries to defend himself by saying that what is needed is not the full power to determine consistency, but only the ability to do so in some circumstances: namely, when one is seriously presenting a machine as a model of the mind. Such a machine would need an appropriate recommendation, and that would include a certificate of consistency. However, the problem remains: the opponent must have access to a recommending authority that can—correctly!—determine consistency. The circularity remains: if out-Gödeling assumes that human beings are somehow in the position of being able to decide about a non-recursive property, the conclusion that they are in some sense better than machines is immediate, but it remains an assumption. In reply, Lucas (1996, p. 118; cf. also 1968) proposed an additional trick, which was to ask the mechanist an insidious question: Would the machine proposed by him ascertain as true its own Gödelian sentence? If he or she answers "Yes", the machine is inconsistent, so it cannot be equivalent to the mind. If the answer is "No", the machine is consistent, and then it can be out-Gödeled by the mind.

Yet the above trick does not do the job—for several reasons. First, because again we need to assume that the mechanist knows whether or not the machine really proves the appropriate Gödelian sentence, or whether or not it is consistent, which brings us back to the previously mentioned problem of circularity, the assumption of the non-mechanical character of the opponent. Second, the trick is dubious because Lucas himself can be asked precisely the same question. Would he be able to prove his own Gödelian formula, or, in other words, determine his own consistency? We are back with the problem discussed above. Maybe he cannot prove his own consistency, but does this say anything significant about him? Third, and this is the most fundamental issue, the trick can also be executed by a machine. To ask the right question (this being that of whether G_S is provable in the theory *S* corresponding to the machine *M*), and respond as explained above (if "Yes", then *S* is inconsistent, if "No", then G_S is unprovable and true), is algorithmic, completely mechanical! It requires no special capabilities, and can be executed by a suitably defined machine. This observation lands one of the most serious blows against every version of the Lucas-style argument.

4.7. The Algorithmic Character of Lucas's Argument

To produce the Gödelian formula, no insight into the nature of the theory is needed; it is enough to execute a certain algorithm, and Lucas's argument can therefore be performed by a machine. The dialectical character of the argument does not help. The effective nature of Gödel's construction was clear to its inventor. Judson Webb even claimed that the mechanization of the diagonalization can be considered the essence of Gödel's work (Webb, 1980, p. 151). I am not sure who first exploited that fact in connection with Lucas. Among early mentions are Irving Good (1967, p. 144), and Paul Benacerraf, who wrote that even if a Gödelian weak spot can be found in every machine, "it is conceivable that a machine could do that as well" (Benacerraf, 1967, p. 22).

Based on this observation, Webb (1980) built an elaborate defense of mechanism. In fact, the matter is more general than just the problem of analyzing Gödel's work. This "is the basic dilemma confronting anti-mechanism: just when the constructions used in its arguments become effective enough to be sure of", then, thanks to Church's Thesis saying that the humanly effective is recursive, "a machine can simulate them" (Webb, 1980, p. 232). Post made that observation in 1924, before Gödel began his research. If we can be "completely conscious" of something, he wrote, it can be mechanized. He called this principle the "Axiom of Reducibility for Finite Operations" (Davis, 1965, p. 424), and it can be seen as an early version of Church's Thesis.

The algorithmic nature of the procedure consisting in the reference to the Gödelian formula is not preserved in the unlimited iteration of the procedure. The mechanist can always add the appropriate Gödelian sentence to the (theory corresponding to the) machine, and Lucas can always apply his procedure to the extended machine. Therefore it would seem natural to add at once all subsequent Gödelian sentences; but then Lucas would apply the procedure again to the machine extended by all those sentences. And so on. Transfinite processes arise naturally. The situation was analyzed, independently of the issue of mechanism, by Turing (1939), and then by Feferman (1962).[5] It turns out that while all Π_1

[5] A review is offered in (Feferman, 1988), and another in (Franzen, 2004a).

sentences are eventually decided, the result depends on the way transfinite ordinal numbers are presented. For Good (1969), this means that the point is not Gödel's theorem, but transfinite counting. This argument was employed also in Hofstadter (1979). According to the latter, the problem for Lucas results from the Church-Kleene theorem stating that there exists no recursive method to describe constructive ordinal numbers (corresponding to recursive well-orderings). Therefore, "no algorithmic method can tell how to apply the method of Gödel to all possible kinds of formal systems […] any human being simply will reach the limits of his own ability to Gödelize at some point" (Hofstadter, 1979, p. 476).[6] The transfinite iteration of the addition of Gödel's sentence, or stronger reflection principles, provides an intricate extension of the picture of incompleteness. Yet, says Shapiro, who considered the issue in (1998) and (2016), it is of no help in the debate about mechanism: "What we do not get, so far as I can see, is any support for a mechanist thesis, nor do we get any support for a Lucas-Penrose-Gödel anti-mechanist perspective" (Shapiro, 2016, p. 200).

Whatever is done in regard to the out-Gödeling is done according to a simple algorithm, and therefore is mechanical. And our attitude towards Church's Thesis is irrelevant as long as the machine, or rather its code, or, equivalently, its number in the accepted listing of all Turing machines, is known. (Usually, effective listings make the number directly dependent on the machine's specification and program.) This algorithm can be presented in technical detail, as is done by Webb (1980, p. 230). Moreover, the recursive function that generates "Achilles heels" of recursive functions can, with no problem, be applied to itself—that is, to its own number, resulting in its own "Achilles heel".

The Lucas argument against mechanism appears weak as soon as it becomes clear that it is itself mechanical. To counter that, Lucas attempts to distinguish two senses of the Gödelian argument: first, when we know an exact specification of the argument so that it can be carried out by a machine, and second, "a certain style of arguing, similar to Gödel's original argument in inspiration, but not completely or precisely specified, and therefore not capable of being programmed into a machine, though capable of being understood and applied by an intelligent mind" (Lucas, 1996, p. 113). Even so, I do not think that out-Gödeling involves any informal move; to use Gödel's theorem is to make a definite mathematical step. And again, if the informal, unspecified arguing is not algorithmic, then Lucas has assumed the non-recursive capabilities of the human mind—which is just what he was supposed to demonstrate. If, on the other hand, the argument is algorithmic, he stands refuted, as we will see in a moment. As a matter of fact, differentiation between the strict and the loose senses of out-Gödeling is rejected, due to the Theorem in Section 5.2, which applies to both the strict and the other senses, as long as the looser one does not beg the question by assuming the non-recursive capabilities of the mind.

[6] Hofstadter seems to have been unaware of the problem we have with establishing consistency. Therefore his analysis is not cogent.

Lucas admits that “an air of paradox remains” (Lucas, 1996, p. 114). A cogent, unformalizable argument, then? No, says Lucas: we are not talking about “absolutely unformalizable” arguments. Yet something must remain unformalized—for example, the use of the rules of inference. This is undoubtedly true, but the same can be said about machines: in computers, some rules are simply contained in the processors. Second, continues Lucas (1996, p. 117), the range of possible applications of his argument remains informal. He does not elaborate, but the remark misses the point in our context. We have considered all possible Turing machines, and they all are listed in a recursive sequence. The appropriate Gödelian formula depends only on the place in the sequence occupied by the machine in question. To out-Gödel, one must know that place, or the code, the program of the machine. However, it is fair to ask whether to know the machine means to know its code. This is highly improbable, even if many idealizations are made. Lucas rejects the issue, saying that we can know the code in principle. Well, then, this will be assumed in Section 5 below, where every Lucas-style argument is shown to involve a contradiction.

Putnam believed that in order “to simulate mathematicians who sometimes change their minds about what they have proved, we would need a program which is also allowed to change its mind”. While there are such programs, he writes, “they are not of the kind to which Gödel’s Theorem applies” (Putnam, 1995, p. 373).

Meanwhile, Benacerraf (1967) presents a precise version of the Lucas argument in order to show that we cannot exclude our mind being a machine, where we nevertheless do not know which one. I shall skip over that analysis, as the general anti-Lucasian argument of Section 5 cuts deeper.

In fact, what has been said so far does not exclude the possibility that our mind is a machine, but we do not know which one. This is the first of the two basic lines of attack against Lucas that were mentioned by Burgess (see Section 3.2). Gödel alluded to such possibilities in (1951)—which, of course, is not to say that he actually believed in their truth. Benacerraf’s analysis seems to be a commentary on that remark by Gödel.

The second line of attack mentioned by Burgess is that it is not excluded that we are inconsistent machines. This was expressed by Putnam and by Benacerraf; the first mention is also in Gödel (1951). It turns out that it is Lucas himself who is inconsistent—see the next section. And it also transpires that Penrose is “unsound”—see Section 6.

5. Lucas’s Inconsistency

To make the analysis as general as possible, we will first consider the assumptions made by Lucas, or, more generally, by the anti-mechanist (Mr. A), in order to out-Gödel his opponent, the mechanist. Four possibly weak conditions will be formulated that seem necessary for the application of some variant of the Lucas-style procedure, and it will then be proved that those general conditions are sufficient to defeat Mr. A by showing his inconsistency. (Of course, the claim is not that the mechanist is right, but only that he cannot be out-Gödeled.) The Inconsistency Theorem also applies to all reasonable modifications of the out-Gödeling procedure.

5.1. The Necessary Conditions for Out-Gödeling

Let us imagine a "dialectical" procedure, this being the most convenient one for Mr. A: he responds to every machine proposed by the opponent. What machines are admissible? All are, but in order to make Mr. A's life easier we assume that nobody will come up with machines that are not equivalent to Turing machines. In addition, we assume that the opponent must be able know the code of the machine and at least the number (in some fixed listing of Turing machines) of the Turing machine equivalent to the proposed one—either equivalent to it in general terms or, as a minimum, equivalent to it in the realm of the arithmetic of natural numbers. This is a limitation on the mechanist, because it excludes the possibility of the machine being a huge box, a network of unknown computers, or a fat volume containing the program. Otherwise we would paralyze Mr. A. The excluded cases amount to a reproach along the lines of "You are a machine, but you don't know which one". So, to avoid the paralysis we assume the following condition:

> **Condition 1.** Each machine proposed by the mechanist is equivalent to a Turing machine, and it is possible to exhibit one such machine.

We assume that each proposed machine can "prove" some statements in the language of arithmetic. The nature of this "proof" is not essential, nor is its connection to real proofs; it may be either the result of understanding or just a thoughtless calculation. Some arithmetical statements are considered "proven" by the machine. Say, a green light goes on, as in Section 4.3. We may not limit in advance the set of admissible Turing machines that can be proposed by the mechanist. We have to assume that Mr. A must respond to each consistent machine—that is, the machine whose arithmetical output (the set of "proven" statements) is consistent. What happens when an inconsistent machine is proposed is irrelevant: Mr. A either responds or disregards it. Inconsistency, according to Lucas and all who adopt his approach, makes the machine unsuitable as a model of our mind's capacity—and, certainly, of his own mind, as he assumes his consistency as obvious. In other words, that response is needed in relation to Case I from Section 4.4; in Case II, meanwhile, anything is allowed. Thus we assume:

> **Condition 2.** The anti-mechanist must respond to every (arithmetically) consistent machine.

The response to the supposition that the proposed machine is equivalent to the human mind, at least in the realm of arithmetic, must consist in the presentation of an arithmetical statement that is not “provable” by the machine. Normally, we would assume that the presented statement must be true. This is how Lucas’s procedure, or any similar procedure based on Gödel’s theorem, works. Let us, however, be much more charitable to Mr. A and demand nothing as regards the truth of the statement. He may present a false statement as long as inconsistency is avoided. This is conceivable. After all, we can’t assume that true sentences are known to us as being true. The Gödel-Rosser theorem gives examples of independent sentences, each of which could be chosen. The liberalized demand regarding the response of Mr. A makes his life much easier; in particular, he can ignore problems with equivocation, with establishing the truth of Gödel’s formula, and all the problems concerning the relation of the theory to metatheory that usually appear in discussions of Gödel’s construction. For Lucas, it was essential that we could “see” the truth of G (Lucas, 1996, p. 103). While his approach is allowed by our conditions, we permit many more responses, since we do not require any use or mention of the notion of truth. The sentence presented as the response to the machine need not be provable in any system. Therefore, we ignore the problem of whether the construction of the Gödelian formula from the code of the machine is practical, and also whether Mr. A must be a logician. Our condition is minimal:

Condition 3. The anti-mechanist’s response to an (arithmetically) consistent machine consists in presenting a statement that is not “provable” by the machine.

For procedures closer to the original out-Gödeling, we could assume that the statement given in response is—as with Gödelian formulas—not derivable using the usual logic from the sentences “provable” by the machine, or even from those sentences together with basic arithmetic.

There is, however, one important limitation that we must impose on Mr. A: namely, that his response must not be arbitrary; it has to be systematic, which here means effective. Moreover, we adopt Church’s Thesis, and assume that the procedure underlying the response must be recursive. Otherwise, we would be allowing a non-mechanical, because non-recursive, procedure, which would mean that Mr. A has non-mechanical powers. This would be exactly the thesis he wants to demonstrate, and such circularity is clearly unacceptable. A random response is not acceptable, because we would not know how to make sure that the proposed sentence is not “provable” by the machine. It must also be assumed that the response is fully determined and not dependent on additional external circumstances. For example, if Mr. A could demand that his opponent propose only consistent machines—as Lucas himself has proposed in some later publications—we would again fall into the trap of assuming non-mechanical human powers—this time those of the mechanist; this follows from the fact that the set of consistent machines is a non-recursive subset of all machines (cf. the Fact, in Section 4.6). In order to avoid circularity, we assume:

Condition 4. The response to the machine is effectively determined in advance.

The requirement of effectiveness must refer to the number (code) of the appropriate Turing machine, in accordance with Condition 1, because it is unclear what could be used if a machine were to be proposed empirically. Thus, first the number of the Turing machine must be found in an effective way, and then a predetermined response can be given, depending solely on this number.

Let me remark that some people have been dissatisfied with the last condition. If we believe that qua humans we are non-mechanical, they say, why should we assume that an effectively determined answer is given? In response to this, it is important to realize that Lucas, Penrose and all who have used the Gödel-based anti-mechanist argument always refer to some form of Gödel's theorem. Their answer is effective, known in advance, expressed as a recursive function of the number (code) of the machine. So Condition 4 fits their strategy. In addition, we allow other strategies as long as they are predetermined and effective. If we dropped this requirement, we would be allowing Mr. A to use his alleged non-mechanical powers, and the whole argument would be superfluous. Therefore, Condition 4 is justified. Together with the other conditions, it turns out, it implies the inconsistency of the anti-mechanist.

5.2. The Theorem Concerning Lucas's Inconsistency

The above conditions can be translated into the terms of mathematical logic. We may assume that all Turing machines are listed in an effective way: M_1, M_2, …, M_n, … Let us further assume that a Lucas-style method is given—that is to say, a method showing the non-mechanical character of the human mind in a way that satisfies Conditions 1 through 4. As explained above, we are dealing with a "dialectical" procedure, and due to Condition 1, we can assume that when applied to the n-th Turing machine M_n it shows that the mind is not equivalent to M_n. This means we have a function F such that for each n, its value, $F(n)$, is sufficient to demonstrate that the mind is not equivalent to M_n. More specifically, in accordance with Condition 3, $F(n)$ is an arithmetical formula not "provable" by M_n. Using "$\mathrm{S}(M_n)$" to denote the set of sentences "provable" by M_n, we get: $F(n) \notin \mathrm{S}(M_n)$. This is assumed for n's with consistent $\mathrm{S}(M_n)$ (briefly, when machine M_n is consistent), because to such machines Mr. A must respond. This is exactly what is stated by Condition 3.

While the scheme is similar to the use of Gödel's theorem, many aspects of Gödel's formula are ignored. Nothing is assumed about the complexity of $F(n)$, and no understanding of the formula is required, on whatever level this might be. As was mentioned before, we do not require that $F(n)$ be true, even though its truth is essential to Lucas's original argument, as is the demonstrability of the

Gödelian formula in a stronger theory. In the present framework, false $F(n)$'s are allowed, which admits many more out-Gödeling procedures. The only assumption is that $F(n)$ is not in $\mathrm{S}(M_n)$, if the latter set is consistent. This is a modest requirement of non-equivalence for the mind and the given machine.

Now we have to decide to what machines the generalization of the out-Gödeling procedure must be applied. The natural stipulation, that it be applicable to all consistent machines, must not be weakened, because no consistent machine may be *a priori* excluded as a simulation of the mind.[7] No restriction on the formula $F(n)$ is imposed for inconsistent M_n. The only limitation is global. As was shown before, consistency is a non-recursive condition—in other words, the set of consistent machines is not decidable: $\mathrm{C} = \{n\colon \mathrm{S}(M_n) \text{ is a consistent theory}\}$ is non-recursive.

This means that we may not assume that F is defined only on C. Were we to do so, we would be assuming Mr. A's power to flawlessly decide whether n belongs to C or not, which would mean his non-mechanical competence—which is precisely the thesis he wants to demonstrate using the hypothetical procedure F. Circularity must be avoided. Fortunately, we do not need to decide in advance what the domain of F is. The only assumption needed to satisfy Condition 2 is that F be a partial function defined at least for consistent machines: $\mathrm{C} \subseteq \mathrm{dom}(F)$.

As explained above, the most important assumption, that of the effectiveness of any hypothetical out-Gödeling procedure, is necessary to avoid circularity, or the assumption that at the very beginning Mr. A's mind is non-mechanical. This means that we assume that F is a partial recursive function, which obviously satisfies Condition 4 if Church's Thesis is accepted. If not, then some effective methods could exist that are not captured by recursive functions.

To sum up, what we must do here is deal with every function F defined for some natural numbers (considered as indices of Turing machines listed in some recursive way) with values that are (Gödel numbers of) arithmetical formulas, so that:

(i) F is partial recursive

(ii) $\mathrm{C} \subseteq \mathrm{dom}(F)$,

(iii) For each $n \in \mathrm{C}$: $F(n) \notin \mathrm{S}(M_n)$.

These assumptions are very weak, but sufficient to prove the following unexpected theorem:

The Inconsistency Theorem. Under the above assumptions, the set of values of F is inconsistent.

[7] The situation differs in Penrose's argument; see below, Section 6.

Proof: Assume that the set of F's values, A = $\{F(n)\colon n \in \mathrm{dom}(F)\}$, is consistent. It is recursively enumerable, due to (i), so it can be enumerated by a Turing machine. We may assume that for some k, A = $\mathrm{S}(M_k)$. By assumption, A is consistent, so $k \in \mathrm{C}$, and due to (ii), $F(k)$ is defined. By (iii), $F(k) \notin \mathrm{S}(M_k)$; that is, $F(k) \notin \mathrm{A}$, which contradicts the definition of A. The contradiction shows that A is inconsistent.

The above theorem is a far-reaching strengthening of the observation that C is non-recursive, and that there is therefore no effective way to distinguish between Case I and Case II in the Lucas procedure. This observation was made in Wang (1974, p. 317), while the set of Gödelian formulas for theories $\mathrm{S}(M_n)$ was considered in Webb (1980). Then, Bowie (1982) showed that an analysis of the set was enough to demonstrate that Lucas was inconsistent. The generalization to include other possible Lucas-style procedures was mentioned in Krajewski (1983), and the general sufficient conditions (i), (ii), (iii) were formulated in (Krajewski, 1988; 1993).

Some further features of the above proof are worth mentioning:

a) The proof shows that even the most sophisticated possible modifications of the "out-Gödeling" procedure, including those that would not use Gödel's theorem but another, perhaps still unknown independence result, all fall into the trap of global inconsistency. The latter is global, because while the set A is inconsistent, we cannot necessarily tell which of its finite subsets is. Moreover, the global inconsistency implies that some $F(n)$'s are false. This by itself need not be fatal in a general case, in contrast to the cases where Gödelian formulas themselves are used. In those cases, a single false response entails contradiction: when $F(n)$ is the Gödel formula for some $n \notin \mathrm{C}$, a specific contradiction is implied; that is to say, the false Gödel formula—let us now call it "G_n"—is provable (precisely because it says it isn't); thus, there exists a formal proof for it in the theory $\mathrm{T}(M_n)$. If k codes this proof, then the arithmetical statement "the number k is the proof of G_n in $\mathrm{T}(M_n)$" has only restricted quantifiers and is true. It is provable in basic arithmetic, so $\mathrm{T}(M_n) \vdash \mathit{Prf}(\mathrm{S}^{(k)}, \ulcorner\varphi\urcorner\ulcorner\mathrm{G}_n\urcorner)$, and this contradicts the provability of G_n, as on account of the definition of G_n, $\mathrm{T}(M_n) \vdash \neg(\exists x)\, \mathit{Prf}(x,\ulcorner\mathrm{G}_n\urcorner)$.

b) The assumption (ii) does not exclude *a priori* the equality of C and dom(F), or that F is defined just for $n \in \mathrm{C}$. That this is impossible, since C is not recursive, and not even recursively enumerable, must be demonstrated independently (as was done above, in Section 4.6).

c) It is worth mentioning that in Condition 1, the phrase "one such machine" cannot be replaced by, for example, the first such machine (in the given listing). If this were to be required, we would fall into a subtle trap. The function $\mathrm{m}(n) = \min\,\{k\colon \mathrm{S}(M_k) = \mathrm{S}(M_n)\}$ is not recursive. Hence, requesting the first appropriate Turing machine would amount to assuming in advance a non-mechanical power with respect to the mechanist.

d) One could conceivably question assumption (ii), the global applicability of the hypothetical procedure. Its dialectical character would then mean that a response is required only in the few cases where the mechanist really proposes a machine *M*. In that case, we would not consider an arbitrary procedure satisfying general conditions; we should restrict our attention to the original out-Gödeling, as advocated by Lucas—that is, the Gödelian formula as the response. Then, as mentioned in a) above, offering even one Gödelian formula in response to an inconsistent machine implies inconsistency.

e) Instead of assumption (iii), we could require something stronger, $\mathrm{S}(M_n)$ non $\vdash F(n)$, as I did in my early papers on the subject. This is in fact satisfied by the original out-Gödeling in which the Gödelian formula is given in response.

The Inconsistency Theorem is so general that we can be sure that not only Lucas, but everyone attempting some systematic version of out-Gödeling, necessarily falls into a contradiction. It is ironic that someone who is otherwise consistent (or, to put it more precisely, for whom the set of arithmetical statements they are ready to accept is consistent) automatically becomes inconsistent as soon as they decide to adopt some Lucas-style procedure. Hence, it seems to have been demonstrated—leaving aside questions about the consistency of women and politicians—that the class of inconsistent humans encompasses at the very least the philosophers who believe in the Gödel-based proof of their superiority over machines.

5.3. Possible Relations between the Mind and Machines: Robot Luke

While the anti-mechanist cannot prove his point by some sort of out-Gödeling, he can still be right. And he can still attempt out-Gödeling. Let us see what possible relations between the mind and machines are not excluded by the previous considerations, and how they could arise. Actually, all the possibilities were mentioned or alluded to by Gödel, especially in the remark quoted below in Section 7. Later, they were described by Putnam, Benacerraf, and others.

If the mind is not mechanical, which is the thesis that was obvious to everyone only a few decades ago and is still believed by most of us—and not just by Lucas, Penrose and of course Gödel—then, if faced with a machine (claimed to be equivalent to the mind), the mind either cannot find its number (Gödel, Putnam, Benacerraf) or it can, and in this case would present the machine's Gödelian formula. The formula will either be true and will serve as an example of the difference between the mind and that machine (Lucas), or it will be false, which would be the case if the machine was inconsistent but we were unable to know this (Putnam).

If the mind is mechanical or computational, and is equivalent to a machine *M*, then either it is (arithmetically) consistent or it is not so. If not, then our mind is an inconsistent machine, and the presentation of the Gödel formula as true only confirms our inconsistency. If *M* is consistent, then we cannot find its number, or code, or program. This was admitted as a possibility by Gödel, and then by Benacerraf, Putnam and, for example, Kripke, who said that there is nothing paradoxical about the impossibility of finding the program of *M*, because if it was found we would be able to distinguish "what I can really prove (absolutely) from what I merely think I can prove" (Chihara, 1972, p. 524). If, however, the number of *M* could be found, we would not be able to prove that the Gödelian formula is true. We couldn't exclude its falsity. The only situation excluded by Gödel's theorem is this: our mind is equivalent to a consistent machine, and we can prove the (Gödelian) formula expressing that consistency.

To put it even more informally, either (a) the mind is not a machine, and there are no Gödelian limitations on it, or (b) the mind is a machine and is inconsistent, and then no limitation based on Gödel's theorem applies, or (c) the mind is a machine and is consistent, and it cannot then prove the Gödelian formula for the machine—that is to say, for itself. This description is close to Gödel's Disjunction (see Section 7).

Assuming that a machine equivalent to the mind is possible, how can it come into being? To manufacture it, a laboratory unimaginably better than anything that is now available would be needed. There is another possibility, however: evolution. It was shown by von Neumann that a machine can replicate itself or produce a more complicated machine. He proposed that we imagine some evolution caused by natural selection (Von Neumann, 1966, Part II, Point 1.8; see also Smart, 1959; Anderson, 1964, p. 104). Random mutations could also take place. Scriven suggested imagining representatives of a robot civilization from another planet.[8] Rudy Rucker develops more fully fantasies about a civilization of robots on the Moon (Rucker, 1982, p. 181). Such a civilization could be initiated by us, humans, and then undergo a Darwinian evolution. Let us imagine that after many generations a robot is born—call him Luke—whose mathematical capabilities are exactly equivalent to those of Lucas. What would then happen?

First of all, we would not know the number of the machine on the list of all Turing machines. We would have no doubt that it is a Turing machine, but even if we could meet it, or even talk with it, we would not be able to analyze its program and make it transparent to us. No description would be available, as it would be too intricate—even if its distant ancestor had been fully described and given a specific number on the list of machines. Second, there would be no way to detect the equivalence of Luke with Lucas. A hypothetical super-mind could do that, if it could analyze and understand human mathematical powers, but the super-mind would not be able to demonstrate the equivalence in a way comprehensible to Lucas or the robot. Third, it would not be excluded that both Lucas and Luke are inconsistent, even if they do their best to fix any malfunctioning.

Now if Lucas really wanted to overcome each contradiction, he ought, in view of the Inconsistency Theorem and its consequences, to abandon any attempt to out-Gödel Luke. Maybe Lucas would still want to maintain that if Luke is consistent, then the Gödelian formula for Luke, which exists somewhere out there in the wide world, is true. However, Lucas would not be able to establish

[8] This appears in a text from 1953; see (Anderson, 1964, p. 38).

the consistency of Luke. Actually, Luke could say exactly the same: if he, Luke, is consistent, his Gödelian formula is true. What is more, Luke could say the same about Lucas! And there is little doubt that Luke would be tempted to try to out-Gödel Lucas. He would be convinced that he is better than Lucas and any human mind. Only it is rather unclear what Luke would say about the inconsistency of female robots and lunar robot politicians.

6. Penrose's "Unsoundness"

Roger Penrose, in books (1989; 1994) and articles (notably 1996),[9] has proposed a new version of the Lucas argument. The point remains the same, even if he is speaking about the non-algorithmic, rather than the non-mechanical, character of our mind or thinking, and even if he uses Turing's theorem on the undecidability of the halting problem rather than Gödel's theorem. Penrose is a well-known mathematician and theoretical physicist who writes with ease; he has presented his version of out-Gödeling in a more comprehensive way than Lucas, and has done so in part as entertaining literature. Both the attractive form of his writing and his scientific authority have made many readers think that a new kind of conclusion has been drawn from the incompleteness theorems.

Penrose attacks both AI and the idea that the mind cannot be grasped scientifically. According to him, conscious processes are different from what goes on in computers. Consciousness does not, however, go beyond the laws of physics—though it may go beyond the physical laws known to us. His speculations on the role of quantum effects and microtubules have met with criticism. Whatever one may think about it, the logical part of Penrose's argument calls for analysis as much as that constructed by Lucas. On it rests everything else, so if it is wrong, everything else becomes doubtful, independently of direct criticism of the physical and biological aspects.

6.1. Penrose's Argument

The logical ingredient of Penrose's work is a variant of the Lucas argument. He commits some mathematical errors: for example by presenting the Gödel sentence as if it were meant to express ω-consistency. Even so, if the ω-consistency schema is expressed as a single sentence, it is Π_3 rather than Π_1, and 1-consistency can be expressed as a Π_2 sentence. Responding to the criticism in Feferman (1995), Penrose not only agrees, but admits that the introduction of "$\Omega(F)$" was "essentially a red herring. In fact, the presentation in Shadows would have been usefully simplified if ω-consistency had not even been mentioned" (Penrose, 1996, paragraph 2.2). Feferman lists more errors in the field of mathematical logic: the

[9] This is an online article that gives a long and detailed reply to important criticisms put forward by David Chalmers, Solomon Feferman, Daryl McCullogh, Drew McDermott, and others in the same issue of *PSYCHE*.

lack of any distinction between the full soundness of a theory (1994, pp. 90–92) and the soundness for Π_1 sentences (1994, pp. 74–75); the substitution of the cases where consistency is needed with those needing ω-consistency; stating a false theorem that for every system *F*, its consistency implies the consistency of $F + Cons_F$ (1994, p. 108), and other inaccuracies.[10] Other errors are made in references to the literature of the subject, and in historical comments. It is hard not to ask the question whether the lack of competence demonstrated makes the whole argument of negligible significance. Well, I do not think so, because all those mistakes can be corrected, and the basic point remains—says Penrose: there is no reason to give up.

His first book, *The Emperor's New Mind* (1989), is less logically advanced, and contains none of the logic-related errors mentioned above. It reads very well, but fails for reasons mentioned earlier here in the analysis of Lucas's argument in Sections 4 and 5: the out-Gödeling procedure is algorithmic, and it depends on the consistency of the relevant theory. The way out would be to assume the consistency or a non-algorithmic insight, but that would amount to a circularity in reasoning. Interestingly, Penrose mentions the idea of "natural selection of algorithms", but rejects it because of the practical improbability of such evolution, as "the slightest 'mutation' of an algorithm […] would tend to render it totally useless" (Penrose, 1989, p. 415). Granted, but what we are dealing with is logical possibility rather than practical probability.

In *Shadows of the Mind* (1994), Penrose reasserted all his opinions, and gave a comprehensive reply to the critics of his first book. "I believe that my form of presentation is better able to withstand the different criticisms that have been raised against the Lucas argument, and to show up their various inadequacies" (p. 49). In one of his papers (1996), Penrose attempts to defend himself against the next wave of criticism. Generally speaking, he is more cautious in his later writings than at the beginning. His aim is to give "a very clear-cut argument for a non-computational ingredient in our conscious thinking" (*ibid.*).

Penrose takes into consideration the main aspects of the criticisms of the Lucas argument and the statements made by Gödel himself—especially Gödel's Disjunction (see Section 7), according to which we cannot rule out our being a machine. If we were, we would be able neither to ascertain the fact nor to detect the consistency of the machine. Schematically, assuming that a machine, algorithm or formal theory *T* is equivalent to the human mind as far as mathematical thinking is concerned, there are three possible cases, I, II, and III, as follows:

[10] See (Feferman, 1995, Part 3). Only for 1-consistent theory *F* does its consistency guarantee the consistency of $F + Cons_F$.

I. *T* is knowable,[11] and its equivalence to the mind is knowable.

II. *T* is knowable, but the equivalence is not.

III. *T* is not knowable.

We can say that III refers to Luke on the moon, and II to Luke carefully analyzed in a human laboratory. Both options are rejected (1994, Chapter 3), and Penrose claims that we are left with case I, the situation of complete knowledge. After an investigation of possible errors or contradictions, he rejects the cases in which *T* is unsound, and then is able, invoking Gödel's Theorem, to conclude that there exists no "knowably sound" system equivalent to the mind (in the realm of Π_1 sentences). Now, this conclusion seems justified. No knowable system—that is, no such system transparent to us and demonstrably consistent—can be equivalent to us. And since Penrose believes himself to have rejected II and III, he can claim that there exists no *T*.

Penrose works under more or less the same assumptions as Lucas, and it would seem that the Inconsistency Theorem applies to Penrose as well as to Lucas: after all, he does seem to accept Conditions 1 through 4 (of Section 5.1). However, in the course of his reasoning, Penrose argues that he would have to respond only to semantically adequate machines. This means that assumption (ii) of the Inconsistency Theorem, the requirement to respond to each consistent machine, is too strong. That is why a new version of the theorem is needed.

6.2. The Theorem Concerning Unsoundness

Let us assume that we have to deal with Lucas-style procedures that are to be applied to semantically adequate, or sound, machines or theories. To recall, an arithmetical theory is sound if all its theorems are true under the standard interpretation in the natural numbers. This is a condition of semantic adequacy. A Turing machine will be called sound if its arithmetical output is sound. Let us put $S = \{n\colon \mathrm{S}(M_n) \text{ is a sound theory}\}$.

Obviously, $S \subseteq C$. If we suppose, after Penrose, that Mr. A must only respond to sound machines, we arrive at the following assumptions:

(i) F is partially recursive,

(ii') $S \subseteq \mathrm{dom}(F)$,

(iii') For each $n \in S$: $F(n) \notin \mathrm{S}(M_n)$.

[11] Cf. (Penrose, 1994, pp. 130–131). I put "knowable" where the original has "consciously knowable" for brevity, and also because it is not clear what unconscious knowledge could mean.

These assumptions[12] are even weaker than before, but they suffice to prove a theorem with a somewhat weaker but similarly unexpected and equally devastating thesis:

The Unsoundness Theorem. *Under the above assumptions, the set of values of F is unsound.*

Proof: Assume that the set of F's values, A = $\{F(n): n \in \mathrm{dom}(F)\}$, is sound. It is recursively enumerable, due to (i), so it can be enumerated by a Turing machine. We may assume that for some k, A = S(M_k). A is sound by assumption, so $k \in$ S, and due to (ii'), $F(k)$ is defined. By (iii'), $F(k) \notin$ S(M_k), that is, $F(k) \notin$ A, which contradicts the definition of A. The contradiction shows that A is unsound.

The set A can be *a priori* consistent even if, being unsound, it contains a false sentence. The unsoundness is sufficient to defeat Penrose's claims, because it means that using his method, or any similar one, he is unsound, as he must accept a false arithmetical statement. His belief in the demonstration of the non-algorithmic character of the mind was based on the conviction that the methods used by him and other mathematicians are fundamentally adequate. Ultimately, no false statement is accepted, he maintains. This belief, coupled with out-Gödeling, results in something that is in contradiction with this very belief. The answer to the question "Do mathematicians unwittingly use an unsound algorithm?" that serves as the title of Section 3.4 in (Penrose, 1994) seems to be "Sometimes yes; for example, Penrose himself".

Thus, as soon as Penrose applies some Gödel-based method of refuting mechanism and algorithmism, he in fact contradicts his belief in the adequacy of the methods of proof he is ready to admit. Having shown his "unsoundness" we could stop here, but let us examine in more detail how the rejection of II and III goes, and why Putnam reproached Penrose for having ignored a possible Case IV.

6.3. The Missed Case, and How to Save Penrose

As has been stated above, the thesis that "we do not ascertain mathematical truth by means of knowably sound" (Penrose, 1994, p. 86) and, let us add, knowable, algorithms is justified, but it is still not excluded that there is a program that does what we do, but where we are not aware of this equivalence because of the program's complication and lack of transparency. Think of Luke.

Next, Penrose maintains that if we used an unsound rule that could produce a false theorem, then this would be fundamentally dubious, since we believe in our soundness. This takes care of Case I.

Penrose assumes that the system underlying our mathematical understanding "is supposed to be simple enough that we are able, at least in principle, to appreciate it in a perfectly conscious way" (Penrose, 1994, p. 132). Here, according to

[12] In (Krajewski, 2003), a slightly stronger assumption (iii') is adopted: $S(M_n)$ non $\vdash$ F(n).

Putnam, Penrose commits the same mistake as Lucas. Before explaining why, let us see how this assumption is used to eliminate Case II. The point here is that this case is said to be very implausible because, first, the algorithm T must be correctable, and therefore sound (1994, Point 3.4), and, second, if the axioms and rules are knowably sound, then all theorems are seen as true, including the Gödelian formula, which is not possible. It must be admitted, however, that Penrose is careful not to say too much; he admits, quoting a remark made by Gödel, that there is "no clear way of ruling out Case II on rigorous logical grounds alone" (1994, p. 133). Penrose also rejects Case III, the unknowable T equivalent to the mind. The main reason is that AI works with knowable programs and, in addition, that Case III would reduce to II or I anyway (1994, p. 144.). This is unsatisfactory, as what is at stake here is the theoretical possibility, and not the practical implementations, of AI. The most important element lacking in Penrose's considerations—to come back to Putnam's point—is the lack of awareness that there might be a program that cannot be understood by us. This would be Case IV. Imagine Luke's program being investigated by human computer scientists. They would never be able to tell what the program does. Actually, this lack of certainty is routine with respect to real-life large programs, which comprise numerous separate subprograms, as well as bugs.

It is worth indicating more explicitly how Case IV can arise. After all, Cases I to III seem to encompass all contingencies. To simplify the formulation as much as possible, let us see what can happen: I. T is known and we know $T \equiv$ mind; II. T is known and we do not know $T \equiv$ mind; III. T is not known. Indeed, nothing else is possible. However, the lacuna emerges when we note that in II it is tacitly assumed that if T is known, then T must be fully graspable. But no: we can, in fact, be faced with a complete description of a program and still have no idea what it does. If it is not "perspicuous" enough, we may be unable to say anything plausible about its consistency. This makes for Case IV.

According to Putnam, Penrose, who indirectly admits the possibility of Case IV,[13] is wrong in claiming that it reduces to Case III. In Penrose's book, Case III applies when we have no knowledge of the program. Therefore, "to reject the possibility that such a formal system might simulate the output of an idealized mathematician (as involving something 'somewhat miraculous' or 'essentially dubious') is to give no argument at all" (Putnam, 1995, p. 372). Putnam concludes that despite the book's strong points, he "regards its appearance as a sad episode in our current intellectual life".

Despite all the criticisms, Penrose maintains that his argument works. He tries to overcome the objections in two ways. One is to limit the possibilities of doing mathematics to familiar ways, while the other is to refer to the so-called

[13] In a letter to the *New York Times* of January 15th, 1995, which is a response to the review of Penrose (1994) by Putnam (*New York Times Book Review* of November 20th, 1994), on which (1995) is based.

“new argument”—considered below, in Section 6.4. For now, let us consider the former, which reveals whence Penrose’s conviction comes.[14]

In his first book, Penrose takes into account the hypothesis (first formulated by Gödel, though Penrose was clearly unaware of that) that our mathematical capabilities are equivalent to an algorithm that is “so complicated or obscure that its very validity can never be known to us”. Penrose’s reply is that “this flies in the face of what mathematics is all about!” (1989, p. 418). This naïve response comes easily if one makes the assumption, as Penrose does, that the putative algorithm is the one actually used by mathematicians. Then we may refer to the fact that mathematics is built from “simple and obvious ingredients”. What is disregarded is any possibility of a hidden algorithm. We are not talking about algorithms taught or acquired at universities, but about, say, the program of Luke.

The existence of Luke, or another complex, intractable formal system equivalent to the human mind, cannot be disproved. On the other hand, from a mathematician’s—as opposed to a logician’s—standpoint the considerations offered by Penrose seem convincing. The reason, mentioned in his first book as a remark on “what mathematics is all about”, was actually expressed by him during the discussion at a conference in Kraków in May 2010. It is that he seems to believe that a mathematical theory of a very different character than the ones we know would be “essentially dubious”, and the emergence of Luke’s mathematical power would be too “miraculous” to really take it into account. This is a perfectly natural attitude for a mathematician, even if it looks somewhat naïve from the logician’s—and perhaps also the computer scientist’s—perspective. The restriction of the range of theories to the “natural” ones does offer a way to overcome the controversy between Penrose and, to use Putnam’s phrase again, “the logical community” (Putnam, 1995, p. 370).

As long as we view mathematical theories, or algorithms, as fundamentally similar to what we know as mathematics, we tend to assume that all the theories that are encompassing our knowledge of the natural numbers must, in principle, be based on a series of transparent basic truths (axioms) and be developed due to the applications of known, correct logical rules. If so, every such theory, if presented to us, must be fully understood, or at least understandable. And this full understanding implies our knowledge of its consistency and, presumably, also soundness. Therefore, out-Gödeling is, indeed, possible.

Thus the “natural” view of the nature of mathematics—which Penrose seems to consider the only admissible one—can serve as an assumption that implies anti-computationalism when added to Gödel’s results. This is by no means a great discovery. Even so, when one is aware of it and, in addition, of Gödel’s Unknowability Thesis (see below, Section 8.1), many of the disputes about out-Gödeling become understandable as being based essentially on misunderstanding.

[14] This section is based on (Krajewski, 2015).

6.4. The “New” Argument

In Chalmers (1995), David Chalmers wrote that a “novel” argument was proposed, or rather “deeply buried”, in Chapter 3 of Penrose’s second book. Penrose (1996) welcomed this unexpected praise with obvious pleasure. While he expressed disappointment that the point was taken note of by almost nobody, and in particular was missed by Putnam, Penrose’s words suggest that the new argument was not really even noted by the author himself!

This “new” argument is supposed to demonstrate that mathematicians cannot consistently believe (know) that their capabilities are algorithmically describable, or even that the set of humanly provable Π_1-sentences is recursively enumerable. In other words, what Penrose really wants us to believe is a thesis stronger than the one he argued for in his book: namely, that “Human mathematicians are not using a knowably sound algorithm in order to ascertain mathematical truth” (1994, p. 76). Later (in Sections 3.16 and 3.23 of [Penrose, 1994], and more explicitly in Section 3 of [Penrose, 1996]) he dropped the adverb “knowably” in order to claim that “Human mathematicians are not using a sound algorithm in order to ascertain mathematical truth; and, obviously, they cannot use an unsound one”. Criticisms of this argument in (Chalmers, 1995; Lindström, 2001; 2006; Shapiro, 2003), and the writings of others, have not prevented Penrose from defending it (as he did in [Penrose, 1996] and, for example, at the 2006 Gödel Centenary Conference in Vienna, as reported in [Feferman, 2007], or in [Penrose, 2011].)

The novelty is that the argument does not depend on the claim that we are able to see that *T* is sound. Rather, the soundness of *T* is derived. That is to say, if we know that the mind is equivalent to *T*—in short, “the mind ≡ *T*”—and that the mind is sound (that is, proves only true statements), where this is something that is supposedly obvious to all of us and was taken for granted by Gödel and Penrose, then we can conclude that *T* is sound. That, according to Chalmers (1995, paragraph 3.2),[15] means the argument goes as follows:

(1) it is known that the mind ≡ *T*,

(2) it is known that the mind is sound,

(3) so *T* is sound;

(4) hence *T*’ = (*T* + “the mind ≡ *T*”) is sound,

(5) whence *Cons*(*T*’) is true, but *T*’ does not prove that (by Gödel’s Theorem);

(6) we know that *Cons*(*T*’) is true,

(7) a contradiction, because if we know that the mind ≡ *T* then *T* proves *Cons*(*T*’).

[15] Chalmers “decodes” the reasoning from a dialogue in (Penrose, 1994, 3.23). Here, I further simplify its formulation.

Having accepted the above proof of contradiction, how can we conclude that there exists no T equivalent to the mind? To reject (1) is not enough, as it only says that while we do not know the equivalence, it can in fact be true. “This is still a strong conclusion”, says Chalmers (1995, paragraph 3.3), “threatening to the prospects of AI”. Well, but rather than reject (1), we could reject (6): that is, we could admit that we do not know that the consistency statement is true. Moreover, we could reject (2). In fact, as Chalmers himself wrote, the assumption (2) by itself leads to contradiction: if we know—unassailably—that we are consistent, we get a contradiction very similar to the way in which it can be argued that our consistency is not provable (see below, Section 8.1). Chalmers (1995, paragraph 3.14) concludes that “perhaps we are sound, but we cannot know unassailably that we are sound”.

Penrose (1996, paragraph 3.4) replies that it is enough to replace (1) with a weaker assumption, the mind $\equiv T$. He also claims that the contradiction pointed out by Chalmers would be avoided if we took into account only the arithmetical Π_1 sentences. Penrose is, however, wrong. The argument sketched above can be further simplified even if the weaker assumption is also considered.

(1’) the mind $\equiv T$; (This is the weaker assumption postulated by Penrose.)

Let us define A as the set of all humanly provable arithmetical Π_1 sentences. By (1’) A is recursively enumerable, since it consists of the sentences provable by T.

(1) we know that the mind $\equiv T$; (The previous assumption.)

If (1), then we know that A consists of Π_1 sentences that are accessible to the mind—i.e. unassailably provable.[16] Further, we put

(2) we know that the mind is sound (at least for Π_1 sentences);

(2’) we know that T is sound in the sense that A consists of true sentences;

(G) as stated by Gödel’s theorem, the Gödelian formula for a consistent (*a fortiori*, sound) r.e. set of arithmetical sentences, is Π_1, true, and outside the set.

Claim: Whether we assume (1), (2), (G) or (1’), (2’), (G), we get a contradiction.

Proof: By (1’) A is r.e., and by (2’) A is sound. Due to (G) the Gödelian formula G is well defined and outside A. We know, however—because we know Gödel’s proof—that G is a true Π_1 sentence. The mind has demonstrated it, so G is in A,

[16] If $\neg$(1) and (1’), then A is equal to the set of provable sentences but possibly we do not know it.

a contradiction. If (1) and (2) are assumed, we have the weaker (1') too, and we get (2'), so we can refer to the previous case.

To avoid the contradiction resulting from (1') $\wedge$ (2'), we can either reject (1'), as Penrose originally wanted, or, going against him, reject (2')—that is, assume our lack of knowledge concerning the soundness of *T*. The contradiction does not follow so simply from (1') $\wedge$ (2). This analysis fits Putnam's criticism. Assuming that (1) $\wedge$ (2) corresponds to Case I (presented above, in Sections 6.1–6.3), the assumption (1') $\wedge$ (2') corresponds to Case II as it was understood by Penrose. And further, the apparently safer assumption (1') $\wedge$ (2) corresponds to Case IV; it does not involve (2'), our understanding of the algorithm *T*.

While (1') $\wedge$ (2) seems safer, we should remember Chalmers's warning, going back to Gödel himself, that (2) itself is problematic, independently of any assumptions concerning T, and independently even of the very existence of *T*. This will be our next topic—see Section 8.1.

7. Gödel's Disjunction

In 1951, in his Gibbs Lecture entitled "Some Basic Theorems on the Foundations of Mathematics and their Implications", Gödel presented the philosophical consequences of his incompleteness theorem, including the problem of mechanism. He believed that over the previous twenty years the philosophical implications of his results had not been understood deeply enough. Since then, his views have been in the process of being disseminated, very slowly, amongst wider professional circles. That progress has been due mostly to the efforts of Wang, Putnam and Benacerraf, and ultimately to Feferman and other editors of his collected works, with his lecture from 1951 being eventually published in 1995. As of now, his views are well-known, but it is still worth summarizing them.

Gödel firmly believed that the mind is not a machine, and he wanted to support this thesis using his formal results. He came to the conclusion, however, that his theorem alone was insufficient for this purpose. The theorem allows a weaker thesis to be demonstrated—what is known as "Gödel's disjunction". When one tries to understand Gödel's views, it is essential to remember that he was certain that we are fundamentally consistent. What is more, he believed that we prove objectively true theorems, at least at times. He distinguished objective from subjectively human mathematics. Proper mathematics in the objective sense consists of all (objectively) true propositions; in the subjective sense it is comprised of all demonstrable propositions, or propositions provable by humans by whatever methods. This is the distinction between, so to say, mathematics in itself and mathematics for us. It is conceivable that the mathematics accessible to humans, not only at a given moment but also potentially, forms just a fragment of the absolute, objective mathematics.

According to Gödel, his theorem implies that mathematics in the objective sense cannot be determined by a well-defined (recursive) system of axioms, which means that it cannot be produced by a Turing machine. And yet it is not excluded that mathematics in the subjective sense could be. In that case, everything that can be proved by humans could be produced by a "finite rule"—that is, by a Turing machine. "However, if such a rule exists, we with our human understanding could certainly never know it to be such". Also, "we could never know with mathematical certainty that all propositions it produces are correct" (Gödel, 1995, p. 309). To put it in other terms, the human mind, at least in the realm of mathematics, would be "equivalent to a finite machine that, however, is unable to understand completely its own functioning" (p. 310). Here, "understanding" means, in particular, the ability to "see" or detect consistency. Gödel later told Wang that one cannot exclude the existence of a machine with powers equivalent to our intuition, and, as quoted in Section 3.2, that such a machine could "even be empirically discoverable" (Wang, 1996, p. 184). This is the source of all later speculations about robot mathematicians, including our friend Luke. Thus, either there exists no Luke, or it (he? she?) can exist, and this produces a Diophantine problem absolutely unsolvable (by us). This is the sense of Gödel's famous Disjunctive Conclusion, a statement that seems to him to be "of great philosophical interest". To quote:

> Either mathematics is incompletable in this sense, that its evident axioms can never be comprised in a finite rule, that is to say, the human mind (even within the realm of pure mathematics) infinitely surpasses the powers of any finite machine, or else there exist absolutely unsolvable diophantine problems. (Gödel, 1995, p. 310)

Here, "absolutely" means "by any mathematical proof the human mind can conceive". Gödel described a simpler formulation of the disjunction to Wang: "Either subjective mathematics surpasses the capability of all computers, or else objective mathematics surpasses subjective mathematics, or both alternatives may be true" (1996, p. 186, quotation 6.1.4).

The last clause reveals that the thesis is meant as a non-exclusive disjunction. However, Gödel did not believe that both are true. He—independently of his theorem—was deeply convinced that the second clause is false, meaning that there is, to use Hilbert's dictum, no *ignorabimus* in mathematics, and that the first clause holds, meaning that the mind goes beyond the mechanical, the algorithmic, and indeed the material. He wanted to establish this claim no less passionately than Lucas, Penrose and many others amongst us. He did not, however, want to accept logically flawed arguments.

In the present paper, the phrase "we know that…" has been treated until now in an informal way. The development of epistemic arithmetic—that is, formalized arithmetic extended by the addition of a predicate K, where K(x) means "x is known"—was initiated by William Reinhardt (1986), and further examined by Shapiro and others, especially Peter Koellner. This last, in (2016) and the accompanying papers (2018a; 2018b), showed that in a natural epistemic arithmetic Gödel's disjunction is provable. Furthermore, using such a framework he demonstrated that strict counterparts of Penrose's and Lucas's arguments fail, as does Penrose's "new" argument. An earlier classic argument in this style is presented below, in Section 8.1.

8. On What Does Follow from Gödel's Theorem

There are various philosophical consequences of Gödel's incompleteness results and the technique utilized in their proofs: for example, the creative role of formalization and the equally unexpected—before Gödel—power of elementary arithmetic. Here it seems appropriate only to consider the consequences directly related to anti-mechanist arguments.

8.1. A Warranted Conclusion: Our Consistency is Not Provable

Gödel's Second Theorem implies that we cannot unassailably prove our consistency. That is to say, whatever the mind is, if we could establish our consistency in a completely precise, undeniable way, *more geometrico*, the proof would be formalizable; this means that it could be simulated on an appropriate machine containing a part of our abilities, i.e. the part that was used in the proof. Such a machine, being weaker than the mind, would be able to prove its own consistency. According to Gödel's results, it would be inconsistent. If it, or rather the formal system corresponding to it, were inconsistent, a larger system—that corresponding to the whole mind, even if not formal—would also be inconsistent. Thus, if we assume the strict provability of our consistency, we arrive at the provability of our inconsistency. This argument *ad absurdum* proves a philosophical thesis. It is that even if we are consistent, we cannot prove this in a precise mathematical way!

The first person to realize this curious limitation was Gödel himself.[17] Later, many philosophers repeated the thesis in one way or another, not always with a full awareness of the history of this statement. I think it deserves a name, such as "Gödel's Thesis of the Undemonstrability of our Consistency", or, more succinctly, "Gödel's Unknowability Thesis" (it being assumed that what is meant here is knowability achievable through rigorous, mathematical-like demonstration).

Gödel's Unknowability Thesis. We cannot unassailably demonstrate our own consistency (let alone soundness).

(NB: Our consistency/soundness is assumed here.)

[17] Even though the thesis was not stated explicitly in (Gödel, 1951), it is certain that the idea comes from him. Cf., however, a fragment in (Gödel, 1995, p. 309), and the notes made by Wang (1974, p. 319) after conversations with him. The thesis is stated in (Wang, 1974, p. 324), and later on in (Wang, 1993, p. 119).

So, one can only conclude that we feel we are consistent, but cannot prove it. Of course, the thesis is not as simple as it looks. As Wang noted, in (1974), it is even unclear whether it is possible to formulate the statement "I am consistent" in terms suited to a mathematical-like demonstration. Shapiro (1998) and Feferman (2007), meanwhile, point to other assumptions needed to make the above sketch work. Things become clearer and stricter when we operate within a more formal framework. In that case, another, more abstract version of the thesis is possible, modeled on the proof of Gödel's second theorem from Löb's derivability conditions. Within the framework of the debate about out-Gödeling and, more specifically, Penrose's new argument, this version was invoked by Chalmers (1995, Section 3). Let knowability be denoted by "$B(.)$", and unconditional (and unassailable) provability (which, of course, implies knowability) by "$\vdash$". The difference between the two is that whereas knowability is something potential, "$\vdash$" means something stronger—namely, that we actually have a proof. Now, assuming three natural conditions, one can directly derive inconsistency and knowledge of inconsistency.

The Abstract Form of the Unknowability Thesis. Assuming $\vdash$ *Cons*, which means, to be specific, $\vdash \neg B(\ulcorner 0 = 1 \urcorner)$, and the conditions

(1) if $\vdash \varphi$ then $\vdash B(\ulcorner \varphi \urcorner)$,
(2) $\vdash B(\ulcorner \varphi \urcorner) \wedge B(\ulcorner \varphi \rightarrow \psi \urcorner) \rightarrow B(\ulcorner \psi \urcorner)$,
(3) $\vdash B(\ulcorner \varphi \urcorner) \rightarrow B(\ulcorner B(\ulcorner \varphi \urcorner) \urcorner)$,

one can derive $\vdash$ Inconsistency.

Proof sketch: Using the diagonal lemma, one can construct Gödel's sentence G (equivalent to $\neg B(\ulcorner G \urcorner)$), and then, from (1), (2) and (3), derive $\vdash (\textit{Cons} \rightarrow G)$. From $\vdash$ *Cons* it follows that $\vdash$ G, so, by (1), $\vdash B(\ulcorner G \urcorner)$, but at the same time, by construction, $\vdash \neg B(\ulcorner G \urcorner)$.

Thus, if we can prove our consistency we are forced to believe a direct contradiction! Many considerations, including also those made by or in relation to Lucas or Penrose, become more transparent once the above thesis is clearly grasped. That is to say, there is a major point of confusion, often encountered in connection with out-Gödeling arguments, that reflects a lack of awareness of it. Hence, the contradiction derived from (Gödel's theorem and) the existence of a machine/program equivalent to the mind is interpreted as furnishing a refutation of the possibility of the existence of such a machine, while the contradiction can already follow from the very assumption that we (unassailably) know our consistency.

In addition to Gödel's results, at least two assumptions that are not self-evident are used in the above reasoning. First, that every exact proof of our consistency can be formalized, and second, that it is possible to express “our consistency”. The first point results from a general principle: complete precision means formalizability. This principle cannot be irrefutably proved, but it makes sense as it is related to Church’s Thesis, and because the thesis is so well grounded the principle seems difficult to refute. If this is accepted, one could question the second point: it is not clear at all how one can express “our consistency”. Basically, there are two options for doing so: either (i) by the common sense statement “I am consistent”, or (ii) by a formal counterpart to this statement. Let us consider them in turn.

In (i), we refer to a common sense statement that has no connection to formal considerations. Wang reflected on just this statement (1974, pp. 317–320),[18] and believed it not provable. The justification for this stance is independent of the reasoning presented above; instead, a more general reason is given: we do not know how to make formal derivations that would lead to a statement about “us”. If the statement “I am consistent” were provable, it would represent provability in a non-formal sense. If that were possible, it would mean that we are not machines, or that we are not even equivalent to machines in the realm of proof-generating reasoning. We certainly may believe that, but it is no more than a general feeling.

In (ii), we consider the formal counterpart to a loose statement expressing consistency; the counterpart cannot be about “me” or “us”, but must rather concern a theory *S* that corresponds to my (or our) mathematical abilities. In that case, we are dealing with a formula that is a formal expression of, say, “S_{ar} is consistent”. The reasoning in question demonstrates that the formula is not provable if S is consistent (that is, I am). It is, however, rather doubtful if a sentence of the type $Cons_S$ is a proper rendering of the statement “I am consistent”. The usual meaning of the statement refers to the will to avoid contradictions, the reliability of our vision of the world, and the claim that the methods used by mathematicians are unfailing. The sentence *Cons*, or any other similar arithmetical formula, is rather far from those ideas. Thus, while something is strictly proved, it is unclear to what extent the conclusion conveys our consistency.

8.2. We Cannot Define the Natural Numbers

The point is that we cannot define numbers. The concept of natural numbers seems perfectly natural. When we consider only the successor function, which seems to define the numbers, the resulting theory is complete and decidable. Adding addition does not change the situation, as was shown by Presburger (1929). Introducing multiplication changes everything, as we have known since Gödel (1931): the resulting theory is incomplete, as are its recursive extensions. They are also undecidable. This is surprising—even, I guess, for those who have been used to the fact and know how to prove it. This phenomenon deserves to be

[18] The sentence “I am consistent” is denoted there by “A”.

called "mathematical emergence" (Krajewski, 2012a). As soon as we have both addition and multiplication, the natural numbers turn out to be extremely complicated. They seem simple, but their structure is objectively complex. At the same time, it seems that we know what numbers are, and that we should be able to define them. The Peano axioms constituted such an attempt but, as we have seen thanks to Gödel, they are not exhaustive. Second-order axioms give a complete theory, but their foundation, the concept of a set of natural numbers, is not completely defined, so the incompleteness reemerges. This means that our axioms define numbers only when taken together with some background knowledge or apparatus that makes possible our intuitive grasp of numbers. We all seem to develop this intuition at some point, if we have normal intellectual capacities. Whatever mechanism is responsible for this development—and we should not pretend that we know it—we can conclude that a complete description of this intuition is impossible. If so, no computer can be taught our concept of a number.

This conclusion is striking, and can be seen as actually another variant of the position defended by Lucas and Penrose. It essentially says that we are better than any machine. If so, we should beware: there must be present here the same subtlety that plagues the arguments of Lucas and Penrose: namely, that the indescribability of the concept of natural numbers means there is no complete description known to us. However, this does not exclude the possibility of a full recursive description of our concept of a number—that is, to use Gödel's term, of subjective arithmetic. This description can be buried in the program of Luke, but we would not be able to formulate it. If presented with the program, we would not know that it does the job, and we would not be able to show that it defines a consistent concept, let alone a sound one. All the limitations treated in the previous sections apply here, as well. Still, the fact that we cannot give a definition of the natural numbers as we understand them is of interest. I suspect that this fact encompasses most of the attractive aspects of Gödel's discoveries so vigorously defended by adherents of the Gödel-based argument for human superiority over machines/programs/robots.

Because no algorithm that we can produce can be known to include our understanding of numbers, we can be sure that creativity is necessary in arithmetic. On the other hand, this conclusion seems certain independently of Gödel, was obvious in the past, and remains convincing to everyone—apart, that is, from some of those who have become believers in the full success of the AI program.

8.3. The Doubtful Impact of the Gödel-Based Anti-Mechanist Argument

Our attitude toward the arguments of Lucas, Penrose, and others is shaped mostly by our general vision of machines and minds, where this in turn must adjust to civilizational changes. For the youth of today, if I may judge from listening to my students, our computerized world makes it easier to accept the idea that anything is mechanizable—including the mind. Now, if the basic assumptions are more important than proofs—which is typically the case where philosophical views are concerned, anyway—it should be expected that the anti-Lucas argument presented here will hardly convince anyone. Moreover, when pointing out contradiction or circularity in Lucas-style arguments, I am not claiming that a proof can be offered—either of the thesis concerning the mechanical character of the mind, or of its contradictory. Generally, I share the opinions of Penrose about the need for intuition and insight in mathematics, and in thinking overall. Nevertheless, I believe that Gödel's results furnish only limited support—though they certainly do offer some: they eliminate the naïve belief in a system of mathematics or an algorithm that is all-encompassing, created by us, and fully understood up to and including the insight of it being contradiction-free**.**

One can doubt the value of the whole anti-mechanist endeavor by noting that no mathematical result can decide a philosophical issue. Shapiro expresses the concerns of many when he states that the problems with the alleged refutations of the mechanist thesis lie "in the idealizations we need in order to make sense of the issues and then apply the incompleteness theorem" (Shapiro, 2016, p. 189). A major problem is caused by the circumstance that the set of knowable, unassailably provable arithmetical sentences seems to have no sharp boundaries. The notion of ideal (available in principle) human (arithmetical) abilities has no clear meaning. Even if we assume, as with machines, the presence of unlimited lifespans, energy and memory, and an absence of mistakes—ideas that are very strange when applied to humans—this is not enough: we need to consider arithmetical sentences that have "an adequate backing", and this is a vague concept; in addition, it seems that we have no adequate backing for the claim that the set of sentences that have an adequate backing is consistent (Shapiro, 2016, p. 199). Further problems with the idealization of the human mind are indicated in Koellner (2018b, Section 5). For example, in science, idealizations involve attributing to some parameters an extreme value, which is often zero; when we consider the "idealized" mind, this is hardly the case. In what principled sense can humans, even on an idealized construal, perform calculations longer than the number of particles in the known physical universe? Such arguments lead Koellner to a disjunctive conclusion:

> Either the statements that "the mind can be mechanized" and "there are absolutely undecidable statements" are indefinite (as the philosophical critique maintains) or they are definite and […] are about as good examples of "absolutely undecidable" propositions as one might find. (2018b, p. 477)

The vagueness of the concepts used in the Lucas-Penrose arguments is a reason to question the whole procedure of demonstrating the superiority of the mind over machines. Still, it makes sense to assume an interpretation that is charitable (to the proponents of the arguments): that is, to accept the possibility of procedures of the kind deployed by Lucas and Penrose. And the present paper then provides a refutation of these procedures, due to the inevitable inconsistency or unsoundness produced by that very reliance on them.

The Lucas-style or Penrose-style argument does not seem to have converted anyone. Those who believe in the fundamentally non-mechanical or non-algorithmic nature of the mind may be glad to witness a mathematical proof of their belief, but such proof will not convince those who posit that a machine can be equivalent to our mind. If pressed, Lucas would, I imagine, say the following: “If I were a machine, then, I am sure, the sentence *Cons* made for me would be true. Whence do I know that? Because I know I am consistent. How do I know? I just know; I feel it. How can the consistency be proved? Well, I feel it; so I am not a machine after all!” Circularity is unavoidable. And, on the other hand, if someone believes that deep down we are complicated machines of some sort, then—even granting the consistency—it is not surprising that we may be unable to prove this consistency. After all, we are not an omniscient machine! As should be clear from the preceding sections, a subtle algorithm, such as Luke’s program, is not logically impossible. Indeed, much the same position has been expressed by Feferman when he writes that

> Even though I am convinced of the extreme implausibility of a computational model of the mind, Penrose’s Gödelian argument does nothing for me personally to bolster that point of view, and I suspect the same will be in general true of readers with similar convictions. On the other hand, I’m sure that those whose sympathies lie in the direction of the computational model of mind will find reasons to dismiss the Gödelian argument quickly. (Feferman, 1995, Part 1.2)

REFERENCES

Anderson, A. R. (Ed.). (1964). *Minds and Machines*. Englewood Cliffs, NJ: Prentice-Hall.

Boolos, G. (1995). Introductory Note to *1951. In: S. Feferman et al. (Eds.), *Collected Works III, Unpublished Essays and Lectures* (pp. 290–304). Oxford: Oxford University Press.

Boolos, G. (1998). *Logic, Logic, and Logic*. Cambridge, MA: Harvard University Press.

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

Berto, F. (2009). *There’s something about Gödel*. Hoboken, New Jersey: Wiley-Blackwell.

Bowie, G. L. (1982). Lucas’ Number is Finally Up. *Journal of Philosophical Logic*, *11*, 279–285.

Brockman, J. (Ed.). (1995). *The Third Culture*. New York: Simon & Schuster.

Burgess, J. (1998). Introduction to Part III. In: G. Boolos, *Logic, Logic, and Logic* (pp. 345–353). Cambridge, MA: Harvard University Press.

Byers, W. (2007). *How Mathematicians Think: Using Ambiguity, Contradiction, and Paradox to Create Mathematics*. Princeton, NJ: Princeton University Press.

Chalmers, D. (1995). Minds, Machines, and Mathematics. *PSYCHE*, *2*(9).

Chihara, C. (1972). On Alleged Refutations of Mechanism Using Gödel's Incompleteness Results. *Journal of Philosophy*, *69*(17), 507–526.

Craig, W. (1953). On Axiomatizability Within a System. *Journal of Symbolic Logic*, *18*, 30–32.

Davis, M. (Ed.). (1965). *The Undecidable*. New York: Raven Press.

Descartes, R. (1637). *Discourse on the Method*. Leiden. Retrieved from: http://www.gutenberg.org/files/59/59-h/59-h.htm

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

Feferman, S. (1962). Transfinite Recursive Progressions of Axiomatic Theories. *Journal of Symbolic Logic*, *27*, 259–316.

Feferman, S. (1984). Kurt Gödel: Conviction and Caution. In: P. Weingartner, C. Puhringer (Eds.), *Philosophy of Science—History of Science. A Selection of Contributed Papers of the 7th International Congress of Logic, Methodology and Philosophy of Science*. Salzburg: Anton Hain, Meisenheim/Glan.

Feferman, S. (1988). Turing in the Land of O(z). In: R. Herken (Ed.), *The Universal Turing Machine. A Half-Century Survey* (pp. 113–147). Oxford: Oxfrod University Press.

Feferman, S. (1995), Penrose's Gödelian Argument, *PSYCHE*, *2*(7).

Feferman, S. (2006). Are There Absolutely Unsolvable Problems? Gödel's Dichotomy. *Philosophia Mathematica*, *14*(2), 134–152.

Feferman, S. (2006a). The Nature and Significance of Gödel's Incompleteness Theorems (Lecture in Princeton, 2006). Retrieved from: http://math.stanford.edu/~feferman/papers/Godel-IAS.pdf

Feferman, S. (2007). *Gödel, Nagel, Minds and Machines* [Ernest Nagel Lecture]. Retrieved from: Columbia University, http://math.stanford.edu/~feferman/papers/godelnagel.pdf

Feigenbaum, E. A., & Feldman, J. (Eds.). (1995). *Computers and Thought*. New York: McGraw-Hill.

Franzén, T. (2004). *Inexhaustibility, A Non-Exhaustive Treatment*, Wellesley, MA: A K Peters.

Franzén, T. (2004a). Transfinite Progressions: A Second Look at Completeness. *Bull. Symb. Log.*, *10*(3), 367–389.

Franzén, T. (2005). *Gödel's Theorem: An Incomplete Guide to Its Use and Abuse*. Wellesley, MA: A K Peters.

Goldstein, R. (2005). *Incompleteness: The Proof and Paradox of Kurt Gödel (Great Discoveries)*. New York: W. W. Norton & Company.

Good, I. J. (1967). Human and Machine Logic. *British Journal for the Philosophy of Science*, *18*, 144–147.

Good, I. J. (1969). Gödel's Theorem is a Red Herring. *British Journal for the Philosophy of Science*, *19*, 357–358.

Gödel, K. (1931). Über formal unentscheidbare Sätze der *Principia mathematica* und verwandter Systeme I [On Formally Undecidable Propositions of Principia Mathematica and Related Systems I]. *Monatshefte für Mathematik und Physik*, *38*, 173–198.

Gödel, K. (1951). Some Basic Theorems on the Foundations of Mathematics and Their Implications. In: S. Feferman et al. (Eds.), *Collected Works III, Unpublished Essays and Lectures* (pp. 304–323). Oxford: Oxford University Press.

Gödel, K. (1986). *Collected Works, Volume I: Publications 1929–1936*. New York: Oxford University Press.

Gödel, K. (1995). *Collected Works III, Unpublished Essays and Lectures*. Oxford: Oxford University Press.

Hofstadter, D. R. (1979). *Gödel, Escher, Bach, an Eternal Golden Braid*. New York: Basic Books.

Horsten, L., & Welch, P. (Eds.). (2016). *Gödel's Disjunction. The Scope and Limits of Mathematical Knowledge*, Oxford: Oxford University Press.

Kemeny, J. G. (1959). *A Philosopher Looks at Science*. Princeton, NJ: D. Van Nostrand.

Koellner, P. (2016). Gödel's Disjunction. In: L. Horsten, P. Welch (Eds.), *Gödel's Disjunction. The Scope and Limits of Mathematical Knowledge* (pp. 148–188). Oxford: Oxford University Press.

Koellner, P. (2018a). On the Question of Whether the Mind can be Mechanized I: From Gödel to Penrose. *The Journal of Philosophy*, *115*(7), 337–360.

Koellner, P. (2018b). On the Question of Whether the Mind can be Mechanized II: Penrose's New Argument. *The Journal of Philosophy*, *115*(9), 453–484.

Krajewski, S. (1983). Philosophical Consequences of Gödel's Theorem. *Bulletin of the Section of Logic*, *12*, 157–164.

Krajewski, S. (1988). Twierdzenie Gödla a filozofia. *Studia Filozoficzne*, *6–7*(271–272), 157–177.

Krajewski, S. (1993). Did Gödel Prove That We Are Not Machines? In: Z. W. Wolkowski (Ed.), *First International Symposium on Gödel's Theorems* (pp. 39–49). Singapore: World Scientific Publishing Co.

Krajewski, S. (2003). *Twierdzenie Gödla i jego interpretacje filozoficzne – od mechanicyzmu do postmodernizmu*. Warsaw: IFiS PAN.

Krajewski, S. (2004). Gödel's Theorem and Its Philosophical Interpretations: From Mechanism to Postmodernism (A Book Summary). *Bulletin of Advanced Reasoning and Knowledge*, *2*, 103–108.

Krajewski, S. (2007). On Gödel's Theorem and Mechanism: Inconsistency or Unsoundness is Unavoidable in Any Attempt to 'Out-Gödel' the Mechanist. *Fundamenta Informaticae*, *81*(1–3), 173–181.

Krajewski, S. (2012). Umysł a metalogika. In: M. Miłkowski, R. Poczobut (Eds.), *Przewodnik po filozofii umysłu* (pp. 619–647). Kraków: WAM.

Krajewski, S. (2012a). Emergence in Mathematics? *Studies in Logic, Grammar and Rhetoric*, *27*(40), 95–105.

Krajewski, S. (2015). Penrose's Metalogical Argument Is Unsound. In: J. Ladyman et al. (Eds.), *Road to Reality with Roger Penrose* (pp. 87–104). Kraków: Copernicus Center Press.

La Mettrie, J. O. de (1748). *L'homme-machine*, Leiden.

Lewis, D. (1969). Lucas Against Mechanism. *Philosophy*, *44*, 231–233.

Lewis, D. (1979). Lucas Against Mechanism II. *Canadian Journal of Philosophy*, *9*(3), 373–376.

Lindström, P. (2001). Penrose's New Argument. *Journal of Philosophical Logic*, *30*, 241–250.

Lindström, P. (2006). Remarks on Penrose's "New Argument". *Journal of Philosophical Logic*, *35*, 231–237.

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

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

Lucas, J. R. (1970). *The Freedom of the Will*. Oxford: Oxford University Press.

Lucas, J. R. (1970a). Mechanism: A Rejoinder. *Philosophy*, *45*, 149–151.

Lucas, J. R. (1996). Minds, Machines and Gödel: A Retrospect. In: P. Millican, A. Clark (Eds.), *Machines and Thought* (pp. 103–124). Oxford: Oxford University Press.

Lucas, J. R. (1997). The Gödelian Argument. *Truth Journal*. Retrieved from: http://www.leaderu.com/truth/2truth08.html

Lucas, J. R. (1998). The Implications of Gödel's Theorem [talk given to the Sigma Club]. Retrieved from:
http://users.ox.ac.uk/~jrlucas/Godel/goedhand.html

Lucas, J. R. (2000). *The Conceptual Roots of Mathematics. An Essay on the Philosophy of Mathematics*. London, New York: Routledge.

Matiyasevich, Y. V. (1993). *Hilbert's Tenth Problem*, Cambridge, MA: MIT Press.

Nagel, E., & Newman, J. R. (1989). *Gödel's Proof*. New York: New York University Press.

Nagel, E., & Newman, J. R. (1961). Answer to Putnam (1960a). *Philosophy of Science*, *28*, 209–211.

Neumann, J. von (1966). *Theory of Self-Reproducing Automata*. Urbana: University of Illinois Press.

Penrose, R. (1989). *Emperor's New Mind*. Oxford: Oxford University Press.

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

Penrose, R. (1996). Beyond the Doubting of a Shadow. *PSYCHE: An Interdisciplinary Journal of Research on Consciousness*, *2*(23).

Penrose, R. (1997). *The Large, the Small and the Human Mind*. Cambridge: Cambridge University Press.

Penrose, R. (2006). Lecture at Gödel Centenary Conference. Vienna.

Penrose, R. (2011). Gödel, the Mind and the Laws of Physics. In: M. Baaz, Ch. H. Papadimitriou, H. Putnam, D. Scott, Ch. Harper (Eds.), *Kurt Gödel and the Foundations of Mathematic: Horizons of Truth* (pp. 339–358). Cambridge: Cambridge University Press.

Post, E. (1941). Absolutely Unsolvable Problems and Relatively Undecidable Propositions—Account of an Anticipation. In: M. Davis, *The Undecidable* (pp. 340–433). New York: Raven Press.

Presburger, M. (1929). Über die Vollständigkeit eines gewissen Systems der Arithmetik ganzer Zahlen, in welchem die Addition als einzige Operation hervortritt. In: *Comptes Rendus de 1er Congrès des Mathématiciens des Pays Slaves* (pp. 92–101). Warsaw.

Putnam, H. (1960). Minds and Machines. In: S. Hook (Ed.), *Dimensions of Mind: A Symposium* (pp. 138–164). New York: New York University Press.

Putnam, H. (1960a). Review: Nagel and Newman, *Gödel's Proof. Philosophy of Science*, *27*, 205–207.

Putnam, H. (1995). Review of *The Shadows of the Mind. Bulletin of the American Mathematical Society*, *32*(3), 370–373.

Raatikainen, P. (2005). On the Philosophical Relevance of Gödel's Incompleteness Theorems. *Revue Internationale de Philosophie*, *59*(4), 513–534.

Reinhardt, W. N. (1986). Epistemic Theories and the Interpretation of Gödel's Incompleteness Theorems. *Journal of Philosophical Logic*, *15*, 427–474.

Rodriguez-Consuegra, F. A. (1995). *Kurt Gödel. Unpublished Philosophical Essays*. Boston: Birkhauser Verlag.

Rosenbloom, P. (1950). *Elements of Mathematical Logic*. New York: Dover.

Rucker, R. von (1982). *Infinity and the Mind*. Boston: Birkhäuser.

Searle, J. R. (1990). Is the Brain's Mind a Computer Program? *Scientific American*, *1*, 26–31.

Shanker, S. G. (Ed.). (1988). *Gödel's Theorem in Focus*. London: Croom Helm.

Shapiro, S. (1996). *The Limits of Logic*. Aldershot: Dartmouth.

Shapiro, S. (1998). Incompleteness, Mechanism, and Optimism. *Journal of Philosophical Logic*, *4*, 273–302.

Shapiro, S. (2003). Mechanism, Truth, and Penrose's New Argument. *Journal of Philosophical Logic*, *32*, 19–42.

Shapiro, S. (2016). Idealization, Mechanism, and Knowability. In: L. Horsten, P. Welch (Eds.), *Gödel's Disjunction. The Scope and Limits of Mathematical Knowledge* (pp. 189–207). Oxford: Oxford University Press.

Slezak, P. (1982). Gödel's Theorem and the Mind. *British Journal for the Philosophy of Science*, *33*, 41–52.

Smart, J. J. C. (1959). Professor Ziff on Robots. *Analysis*, *19*, 117–118.

Smart, J. J. C. (1961). Gödel's Theorem, Church's Thesis, and Mechanism. *Synthese*, *13*, 105–110.

Turing, A. (1937). On Computable Numbers, With an Application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society*, *s2–42*(1), 230–265.

Turing, A. (1939). Systems of Logic Based on Ordinals. *Proceedings of the London Mathematical Society*, *s2–45*, 161–228.

Turing, A. (1950). Computing Machinery and Intelligence. *Mind*, *59*, 433–460.

Wang, H. (1974). *From Mathematics to Philosophy*. New York: Routledge and Kegan Paul.

Wang, H. (1993). On Physicalism and Algorithmism: Can Machines Think? *Philosophia Mathematica*, *1*, 97–138.

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

Webb, J. C. (1980). *Mechanism, Mentalism, and Metamathematics*. Dordrecht: Reidel.

STUDIA SEMIOTYCZNE (SEMIOTIC STUDIES), t. XXXIV, nr 1 (2020), s. 57–80
ISSN 0137-6608, e-ISSN 2544-073X
DOI: 10.26333/sts.xxxiv1.03

Essay

WILFRIED SIEG *

GÖDEL'S PHILOSOPHICAL CHALLENGE (TO TURING)[1]

SUMMARY: The incompleteness theorems constitute the mathematical core of Gödel's philosophical challenge. They are given in their "most satisfactory form", as Gödel saw it, when the formality of theories to which they apply is characterized via Turing machines. These machines codify human mechanical procedures that can be carried out without appealing to higher cognitive capacities. The question naturally arises, whether the theorems justify the claim that the human mind has mathematical abilities that are not shared by any machine. Turing admits that non-mechanical steps of intuition are needed to transcend particular formal theories. Thus, there is a substantive point in comparing Turing's views with Gödel's that is expressed by the assertion, "The human mind infinitely surpasses any finite machine". The parallelisms and tensions between their views are taken as an inspiration for beginning to explore, computationally, the capacities of the human mathematical mind.[2]

KEYWORDS: computability, Church's Thesis, Turing's Thesis, incompleteness, undecidability, Post production systems, computable dynamical systems.

* Carnegie Mellon University. E-mail: sieg@cmu.edu. ORCID: 0000-0002-7130-0524.

[1] This essay by Wilfried Sieg, "Gödel's Philosophical Challenge (to Turing)" was previously published in Copeland, B. Jack, Carl J. Posy, and Oron Shagrir, eds., *Computability—Turing, Gödel, Church, and Beyond*, © 2013 Massachusetts Institute of Technology; it is reprinted here by permission of The MIT Press. The author has added a new Postscriptum in June 2020.

[2] In Milan Kundera's *Ignorance* (2002) one finds on page 124, "We won't understand a thing about human life if we persist in avoiding the most obvious fact: that a reality no longer is what it was when it was; it cannot be reconstructed". These remarks of Kundera, born in Gödel's hometown Brno, apply even to attempts of understanding and reconstructing a limited aspect of past intellectual life.

Introduction

“To Turing” is flanked by parentheses in the title, as the philosophical challenge issued by Gödel’s mathematical results, the incompleteness theorems, was not only a challenge to Turing but also to Gödel himself; it certainly should be taken up by us. At issue is the question whether there is a rigorous argument from these results to the claim that machines can never replace mathematicians or, more generally, that the human mind infinitely surpasses any finite machine. Gödel made the former claim already in 1939; the latter assertion was central in his Gibbs Lecture of 1951. In his note of 1972, Gödel tried to argue for that assertion with greater emphasis on subtle aspects of mathematical experience in set theory. He explored, in particular, the possibility of a humanly effective, but non-mechanical process for presenting a sequence of ever-stronger axioms of infinity.

To understand the claims in their broad intellectual context, one is almost forced to review the emergence of a rigorous notion of computability in the early part of the twentieth century. Gödel’s role in that emergence is “dichotomous”, as John Dawson noted in his lecture (2006). There are crucial impulses, like the definition of general recursive functions in the 1934 Princeton Lectures. This definition was the starting point for Kleene’s work in recursion theory and served as the rigorous mathematical notion in Church’s first published formulation of his “thesis” in (1935). However, there is neither a systematic body of recursion theoretic work nor an isolated central theorem that is associated with Gödel’s name. The reason for that is clear: Gödel was not interested in developing the theory, but rather in securing its conceptual foundation. He needed such a foundation for two central and related purposes, namely, (i) to formulate the incompleteness theorems in mathematical generality for all formal theories (containing arithmetic) and (ii) to articulate and sharpen philosophical consequences of the undecidability and incompleteness results.

The philosophical consequences, as I indicated, are concerned with the claimed superiority of the human mind over machines in mathematics. This takes for granted that a convincing solution to the issue indicated under (i) has been found and that such a solution involves suitably characterized machines. The first two parts of this essay, entitled *Primitive & General Recursions* and *Finite Machines & Computors*, present the general foundational context. It is only then that the central philosophical issue is addressed in the third part, *Beyond Mechanisms & Discipline*. Gödel’s and Turing’s views on mind are usually seen in sharp opposition to each other. Indeed, Gödel himself claimed to have found a “philosophical error in Turing’s work”; his argument for such an error rests on the (incorrect) assumption that Turing tried to establish in (1936) that mental procedures do not go beyond mechanical ones. If one focuses on the real challenge presented by the incompleteness theorems, then one finds that Gödel and Turing pursue parallel approaches with complementary programmatic goals, but dramatically different methodological perspectives. Concrete work to elucidate the situation is suggested in the last part of the essay, *Finding Proofs (With Ingenuity)*.

I. Primitive & General Recursions

It was of course Kronecker who articulated in the 1870s forcefully the requirement that mathematical objects should be finitely presented, that mathematical notions should be decidable, and that the values of functions should be calculable in finitely many steps. And it was of course Dedekind who formulated in his essay *Was sind und was sollen die Zahlen?* the general schema of primitive recursion. At the turn from the nineteenth to the twentieth century, Hilbert transferred Kronecker's normative requirements from mathematics to the frameworks in which mathematical considerations were to be presented, i.e., to axiomatic theories. This shift was accompanied by a methodologically sound call for proofs to establish the theories as consistent.[3] A syntactic and, in Hilbert's view, first "direct" consistency proof was given in his (1905) for a purely equational theory. The approach was criticized fairly by Poincaré and, for a long time, not pursued further by Hilbert. Only in 1921 did Hilbert come back to this particular argument and used it as the starting point of novel proof theoretic investigations, now with a finitist foundation that included recursion equations for all primitive recursive functions as basic principles.[4]

In order to carry out the proof theoretic arguments, functions in formal theories have to be calculable, indeed, calculable from a finitist perspective. That is clear from even a rough outline of the consistency proof Hilbert and Bernays obtained in early 1922. It was presented in (Hilbert, 1923) and concerns the quantifier-free theory we call primitive recursive arithmetic (PRA) and proceeds as follows. The linear proofs are first transformed into tree structures; then all variables are systematically replaced by numerals resulting in a configuration of purely numeric statements that all turn out to be true and, consequently, cannot contain a contradiction. Yet to recognize the truth of the numeric formulae one has to calculate, from a finitist perspective, the value of functions applied to numerals.[5] This was a significant test of the new proof theoretic techniques, but the result had one drawback: a consistency proof for the finitist system PRA was not needed according to the programmatic objectives, but a treatment of quantifiers was required. Following Hilbert's *Ansatz* of eliminating quantifiers in favor of ε-terms, Ackermann carried out the considerations for "transfinite" theories, i.e., for the first-order extension of PRA (correctly, as it turned out, only with just

[3] This is in the logicist tradition of Dedekind (cf. Sieg & Schlimm, 2005; Sieg, 2009a).

[4] For the development of Hilbert's foundational investigations, it has to be mentioned that the Göttingen group had in the meantime assimilated Whitehead and Russell's *Principia Mathematica*; that is clear from the carefully worked out lecture notes from the winter term 1917–1918; cf. (Sieg, 1999).

[5] That was done in (Hilbert & Bernays, 1921/2); a summary is found in Section II of (Ackermann, 1925), entitled *The Consistency Proof Before the Addition of the Transfinite Axioms*. Ackermann does not treat the induction rule, but that can easily be incorporated into the argument following Hilbert and Bernays. The presentation of these early proof theoretic results is refined and extended in (Hilbert & Bernays, 1934).

quantifier-free induction). Herbrand obtained in 1931 the result for essentially the same system, but with recursion equations for a larger class of finitistically calculable functions; that is how Herbrand described the relation of his result to that of Ackermann in a letter of 7 April, 1931, to Bernays.

As to the calculability of functions, Hilbert and Bernays had already emphasized in their lectures from 1921–1922, "For every single such definition by recursion it has to be determined that the application of the recursion formula indeed yields a number sign as function value—for each set of arguments". Such a determination was taken for granted for primitive recursive definitions. We find here, in a rough form, Herbrand's way of characterizing broader classes of finitistically calculable functions according to the schema in his 1931 letter to Gödel:

> In arithmetic, we have other functions as well, for example functions defined by recursion, which I will define by means of the following axioms. Let us assume that we want to define all the functions $f_n(x_1, x_2, \ldots, x_{pn})$ of a certain finite or infinite set F. Each $f_n(x_1, \ldots)$ will have certain defining axioms; I will call these axioms (3F). These axioms will satisfy the following conditions:
>
> (i) The defining axioms for f_n contain, besides f_n, only functions of lesser index.
>
> (ii) These axioms contain only constants and free variables.
>
> (iii) We must be able to show, by means of intuitionistic proofs, that with these axioms it is possible to compute the value of the functions univocally for each specified system of values of their arguments. (This letter is found in [Gödel, 2003].)

Having given this schema, Herbrand mentions that the non-primitive recursive Ackermann function falls under it. Recall that Herbrand, as well as Bernays and von Neumann at the time, used "intuitionistic" as synonymous with "finitist".

In two letters from early 1931, Herbrand and Gödel discussed the impact of the incompleteness theorems on Hilbert's Program. Gödel claimed that some finitist arguments might not be formalizable even in the full system of *Principia Mathematica*; in particular, he conjectured that the finitist considerations required for guaranteeing the unicity of the recursion axioms are among them. In late 1933, Gödel gave a lecture in Cambridge (Massachusetts) and surveyed the status of foundational investigations; see (Gödel, 1933). This fascinating lecture describes finitist mathematics and reveals a number of mind changes: (i) when discussing calculable functions, Gödel emphasizes their recursive definability, but no longer the finitist provability requirement, and (ii) when discussing Hilbert's Program, Gödel asserts that all finitist considerations can be formalized in elementary number theory. He supports his view by saying that finitist considerations use only the proof and definition principle of complete induction; the class of functions definable in this way includes all those given by Herbrand's schema. I take Gödel's deliberate decision to disregard the provability condition as a first and very significant step toward the next major definition, i.e., that of general recursive functions.

A few months after his lecture in Cambridge, Gödel was presented with Church's proposal of identifying the calculability of number-theoretic functions with their λ-definability. Gödel, according to Church in a letter of 29 November, 1935, to Kleene, viewed the proposal as "thoroughly unsatisfactory" and proposed "to state a set of axioms which would embody the generally accepted properties of this notion [i.e., effective calculability], and to do something on that basis" (in Sieg, 1997, p. 463). However, instead of formulating axioms for that notion in his 1934 Princeton lectures, Gödel took a second important step in further modifying Herbrand's definition. He considered as general recursive those total number theoretic functions whose values can be computed in an equational calculus, starting with general recursion equations and proceeding with very elementary replacement rules. In a 1964 letter to van Heijenoort, Gödel asserted, "… it was exactly by specifying the rules of computation that a mathematically workable and fruitful concept was obtained".[6]

Gödel had obviously defined a broad class of calculable functions, but at the time he did not think of general recursiveness as a rigorous explication of calculability.[7] Only in late 1935 did it become plausible to him, as he put it on 1 May, 1968, in a letter to Kreisel, "that my [incompleteness] results were valid for all formal systems". The plausibility of this claim rested on an observation concerning computability in the Postscriptum to his 1936-note, *On the Length of Proofs*. Here is the observation for systems S_i of i-th order arithmetic, $i > 0$.

> It can, moreover, be shown that a function computable in one of the systems S_i, or even in a system of transfinite order, is computable already in S_1. Thus, the notion "computable" is in a certain sense "absolute", while almost all metamathematical notions otherwise known (for example, provable, definable, and so on) quite essentially depend upon the system adopted. (Gödel, 1936, p. 399)

Ten years later, in his contribution to the Princeton Bicentennial Conference, Gödel formulated the absoluteness claim not just for higher-type extensions of arithmetic, but for any formal system containing arithmetic, in particular, for set theory. The philosophical significance of general recursiveness is almost exclusively attributed to its absoluteness. Connecting his remarks to a previous lecture given by Tarski, Gödel started his talk with:

> Tarski has stressed in his lecture (and I think justly) the great importance of the concept of general recursiveness (or Turing's computability). It seems to me that this importance is largely due to the fact that with this concept one has for the first time succeeded in giving an absolute definition of an interesting epistemological notion, i.e., one not depending on the formalism chosen. (Gödel, 1946, p. 150)

[6] For brief descriptions of the equational calculus see Gödel's (1934, pp. 368–369) or his (193?, pp. 166–168).

[7] Cf. his letter to Martin Davis quoted in (Davis, 1982, p. 9).

In 1965, Gödel added a footnote to this remark clarifying the precise nature of the absoluteness claim:

> To be more precise: a function of integers is computable in any formal system containing arithmetic if and only if it is computable in arithmetic, where a function *f* is called computable in *S* if there is a computable term representing *f*.

The metamathematical absoluteness claim as formulated in 1936 can readily be established for the specific theories of higher-order arithmetic. However, in order to prove the claim that functions computable in any formal system containing arithmetic are general recursive, the formal nature of the systems has to be rigorously characterized and then exploited. One can do that, for example, by imposing on such systems the recursiveness conditions of Hilbert and Bernays that were formulated in Supplement II of the second volume of their *Grundlagen der Mathematik*. When proceeding in this way one commits, however, a subtle circularity in case one simultaneously insists that the general recursive functions allow the proper mathematical characterization of formality.[8]

In Gödel's 1946 Princeton remark, "Turing's computability" is mentioned, but is listed parenthetically behind general recursiveness without any emphasis that it might play a special role. That notion becomes a focal point in Gödel's reflections only in the 1951 Gibbs Lecture where he explores the implications of the incompleteness theorems, not in their original formulation, but rather in a "much more satisfactory form" that is "due to the work of various mathematicians". He stresses, "The greatest improvement was made possible through the precise definition of the concept of finite procedure, which plays such a decisive role in these results".[9] Gödel points out that there are different ways of arriving at a precise definition of finite procedure, which all lead to exactly the same concept. However, and here is the observation on Turing,

> The most satisfactory way … [of arriving at such a definition] is that of reducing the concept of finite procedure to that of a machine with a finite number of parts, as has been done by the British mathematician Turing. (Gödel, 1951, pp. 304–305)

Gödel does not expand on this brief remark; in particular, he gives no hint of how reduction is to be understood. He also does not explain, why such a reduction is "the most satisfactory way" of getting to a precise definition or, for

[8] This is analyzed in section 2 of (Sieg, 1994) and with an illuminating Churchian perspective, in section 4 of (Sieg, 1997).

[9] In a footnote Gödel explains that the concept of "finite procedure" is considered to be equivalent to the concept of a "computable function of integers", i.e., a function *f* "whose definition makes it possible actually to compute *f*(*n*) for each integer *n*". The reason why that can be done is formulated as follows: "The procedures to be considered do not operate on integers but on formulas, but because of the enumeration of the formulas in question, they can always be reduced to procedures operating on integers".

that matter, why the concept of a machine with a finite number of parts is equivalent to that of a Turing machine. At this point, it seems, the ultimate justification lies in the pure and perhaps rather crude fact that finite procedures can be effected by finite machines.[10]

Gödel claims in the Gibbs Lecture (1951, p. 311) that the state of philosophy "in our days" is to be faulted for not being able to draw in a mathematically rigorous way the philosophical implications of the "mathematical aspect of the situation", i.e., the situation created by the incompleteness results. I have argued that not even the mathematical aspect had been clarified in a convincing way; after all, it crucially depended on very problematic considerations concerning a precise notion of computability.

II. Finite Machines & Computors

To bring out very clearly that the appeal to a reduction is a most significant step for Gödel, let me go back to the informative manuscript (Gödel, 193?) from the late 1930s. In it, Gödel examines general recursiveness and Turing computability, but under a methodological perspective that is completely different from the one found in the Gibbs Lecture. After having given a perspicuous presentation of his equational calculus, Gödel claims outright that it provides "the correct definition of a computable function". Thus, he seems to be fully endorsing Church's Thesis concerning general recursive functions. He adds a remark on Turing asserting, "That this really is the correct definition of mechanical computability was established beyond any doubt by Turing". How did Turing establish this claim? Here is Gödel's answer:

> [Turing] has shown that the computable functions defined in this way [via the equational calculus] are exactly those for which you can construct a machine with a finite number of parts which will do the following thing. If you write down any number $n_1, \ldots, n_r$ on a slip of paper and put the slip of paper into the machine and turn the crank, then after a finite number of turns the machine will stop and the value of the function for the argument $n_1, \ldots, n_r$ will be printed on the paper. (Gödel, 193?, p. 168)

The mathematical theorem stating the equivalence of Turing computability and general recursiveness plays the pivotal role at this time: Gödel does not yet focus

[10] In his (1933, p. 45) Gödel describes the constructivity requirements on theories and explicates the purely formal character of inference rules. The latter "refer only to the outward structure of the formulas, not to their meaning, so that they could be applied by someone who knew nothing about mathematics, or by a machine". He also asserts there, "thus the highest possible degree of exactness is obtained".

on Turing's analysis as being the basis for a reduction of mechanical calculability to (Turing) machine computability.[11]

The appreciation of Turing's work indicated in the Gibbs Lecture for the first time is deepened in other writings of Gödel. Perhaps, it would be better to say that Turing's work appears as a topic of perceptive, but also quite aphoristic remarks. Indeed, there are only three such remarks that were published during Gödel's lifetime after 1951: (i) the *Postscriptum* to the 1931 incompleteness paper, (ii) the *Postscriptum* to the 1934 Princeton Lecture Notes, and (iii) the 1972 note *A Philosophical Error in Turing's Work*. The latter note appeared in a slightly different version in Wang's book from 1974. In the sequel, I will refer to the "1972-note" and the "1974-note", though I am convinced that the first note is the later one.

The brief *Postscriptum* added to (Gödel, 1931) in 1963 emphasizes the centrality of Turing's work for both incompleteness theorems; here is the text:

> In consequence of later advances, in particular of the fact that due to A. M. Turing's work a precise and unquestionably adequate definition of the general notion of formal system can now be given, a completely general version of Theorems VI and XI is now possible. That is, it can be proved rigorously that in every consistent formal system that contains a certain amount of finitary number theory there exist undecidable arithmetic propositions and that, moreover, the consistency of any such system cannot be proved in the system. (Gödel, 1931, p. 195)

In the more extended *Postscriptum* written a year later for his Princeton Lecture Notes, Gödel repeats this remark almost verbatim, but then states a reason why Turing's work provides the basis for a "precise and unquestionably adequate definition of the general concept of formal system": "Turing's work gives an analysis of the concept of 'mechanical procedure' (alias 'algorithm' or 'computation procedure' or 'finite combinatorial procedure'). This concept is shown to be equivalent with that of a 'Turing machine'" (Gödel, 1934, pp. 369–370).

In a footnote attached to the last sentence Gödel refers to (Turing, 1936) and points to its ninth section, where Turing argues for the adequacy of his machine concept. Gödel emphasizes that previous equivalent definitions of computability, including general recursiveness and λ-definability, "are much less suitable for our purposes". However, he does not elucidate the special character of Turing computability in this context or any other context I am aware of, and he does not indicate either, how he thought an analysis proceeded or how the equivalence

[11] In the spring of 1939, Gödel gave a logic course at the University of Notre Dame and argued for the superiority of the human mind over machines via the undecidability of the decision problem for predicate logic; the latter is put into the historical context of Leibniz's *Calculemus*! He claims: "So here already one can prove that Leibnitzens [sic!] program of the *calculemus* cannot be carried through, i.e. one knows that the human mind will never be able to be replaced by a machine already for this comparatively simple question to decide whether a formula is a tautology or not". The conception of machine is as in (193?)—an office calculator with a crank.

between the (analyzed) concept and Turing computability could be shown. In the next paragraph, I will give a very condensed version of Turing's important argument, though I note right away that Turing did not view it as proving an equivalence result of the sort Gödel described.[12]

Call a human computing agent who proceeds mechanically a computor; such a computor operates deterministically on finite, possibly two-dimensional configurations of symbols when performing a calculation.[13] Turing aims to isolate the most basic steps taken in calculations, i.e., steps that need not be further subdivided. This goal requires that the configurations on which the computor operates be immediately recognizable. Joining this demand with the evident limitation of the computor's sensory apparatus leads to the "boundedness" of configurations and the "locality" of operations:

(B) There is a fixed finite bound on the number of configurations a computor can immediately recognize; and

(L) A computor can change only immediately recognizable (sub-) configurations.

As Turing considers the two-dimensional character of configurations as inessential for mechanical procedures, the calculations of the computor, satisfying the boundedness and locality restrictions, are directly captured by Turing machines operating on strings; the latter can provably be mimicked by ordinary two-letter Turing machines.[14]

So, it seems we are naturally and convincingly led from calculations of a computor on two-dimensional paper to computations of a Turing machine on a linear tape. Are these machines in the end, as Turing's student Gandy put it, nothing but codifications of computors? Is Gandy right when claiming in (1980, p. 124) that Turing's considerations provide (the outline of) a proof for the claim, "What can be calculated by an abstract human being working in a routine way is computable?" Does Turing's argument thus secure the conclusiveness and generality of the limitative mathematical results, respect their broad intellectual

[12] I have analyzed Turing's argument in other papers (e.g., 1994; 2002). My subsequent discussion takes Turing machines in the way in which Post defined them in (1947), namely, as production systems. That has the consequence that states of mind are physically represented, quite in Turing's spirit; cf. part III of section 9 in his paper (1936) and the marvelous discussion in (Turing, 1954).

[13] That captures exactly the intellectual problematic and context: the *Entscheidungsproblem* was to be solved mechanically by us; formal systems were to guarantee intersubjectivity on a minimal, mechanically verifiable level between us.

[14] It should be noted that step-by-step calculations in the equational calculus cannot be carried out by a computor satisfying these restrictive conditions: arbitrarily large numerals have to be recognized and arbitrarily complex terms have to be replaced by their numerical values—in one step.

context and appeal only to mechanical procedures that are carried out by humans without the use of higher cognitive capacities?

Turing himself found his considerations mathematically unsatisfactory. Indeed, he took two problematic steps by (i) starting the analysis with calculations on two-dimensional paper (this is problematic as possibly more general configurations and procedures should be considered) and (ii) dismissing, without argument, the two-dimensional character of paper as "no essential of computation". However, a restricted result is rigorously established by Turing's considerations: Turing machines can carry out the calculations of computors—as long as computors not only satisfy (*B*) and (*L*), but also operate on linear configurations; this result can be extended to extremely general configurations, K-graphs.[15] But even then, there is no proof of Turing's Thesis.

The methodological difficulties can be avoided by taking an alternative approach, namely, to characterize a Turing Computor axiomatically as a discrete dynamical system and to show that any system satisfying the axioms is computationally reducible to a Turing machine (Sieg, 2002; 2009a). No appeal to a thesis is needed; rather, that appeal has been replaced by the task of recognizing the correctness of axioms for an intended notion. This way of extracting from Turing's analysis clear axiomatic conditions and then establishing a representation theorem seems to follow Gödel's suggestion to Church in 1934; it also seems to fall, in a way, under the description Gödel gave of Turing's work, when arguing that it analyzes the concept "mechanical procedure" and that "this concept is shown to be equivalent with that of a Turing machine".[16]

With the conceptual foundations in place, we can examine how Gödel and Turing thought about the fact that humans transcend the limitations of any particular Turing machine (with respect to the first incompleteness theorem). They chose quite different paths: Gödel was led to argue for the existence of humanly effective, non-mechanical procedures and continued to identify finite machines with Turing machines; thus, he "established" our topical claim that the human mind infinitely surpasses any finite machine. Turing, by contrast, was led to the more modest demand of releasing computors and machines from the strict discipline of carrying out procedures mechanically and providing them with room for initiative. Let us see what that amounts to.

III. Beyond Mechanisms & Discipline

Gödel's paper (193?) begins by referring to Hilbert's famous words, "for any precisely formulated mathematical question a unique answer can be found".

[15] The underlying methodological matters are discussed in (Sieg & Byrnes, 1996), where K-graphs were introduced as a generalization of the graphical structures considered in (Kolmogorov & Uspenski, 1963).

[16] In (Martin, 2005), a particular (and insightful) interpretation of Gödel's view on mathematical concepts is given. It is developed with special attention to the concept of set, but it seems to be adaptable to the concept of computability. Cf. the summary on pp. 223–224.

Those words are taken to assert that for any mathematical proposition A there is a proof of either A or not-A, "where by 'proof' is meant something which starts from evident axioms and proceeds by evident inferences". He argues that the incompleteness theorems show that something is lost when one takes the step from this notion of proof to a formalized one:

> [I]t is not possible to formalise (sic!) mathematical evidence even in the domain of number theory, but the conviction about which Hilbert speaks remains entirely untouched. Another way of putting the result is this: it is not possible to mechanise (sic!) mathematical reasoning […]. (Gödel, 193?)

And that means for Gödel that "it will never be possible to replace the mathematician by a machine, even if you confine yourself to number-theoretic problems" (pp. 164–165). Gödel took this deeply rationalist and optimistic perspective still in the early 1970s: Wang reports that Gödel rejected the possibility that there are number theoretic problems undecidable for the human mind (Wang, 1974, pp. 324–325).[17]

Gödel's claim that it is impossible to mechanize mathematical reasoning is supported in the Gibbs Lecture by an argument that relies primarily on the second incompleteness theorem; see the detailed analyses in (Feferman, 2006a) and (Sieg, 2007, Section 2). This claim raises immediately the question, "What aspects of mathematical reasoning or experience defy formalization?" In his 1974-note, Gödel points to two "vaguely defined" processes that may be sharpened to systematic and effective, but non-mechanical procedures; namely, the process of defining recursive well-orderings of integers for larger and larger ordinals of the second number class and that of formulating stronger and stronger axioms of infinity. The point is reiterated in the modified formulation of the 1972-note, where Gödel, on p. 305, considers another version of his first theorem that may be taken "as an indication for the existence of mathematical yes or no questions undecidable for the human mind". However, he points to a fact that in his view weighs against such an interpretation: "There do exist unexplored series of axioms which are analytic in the sense that they only explicate the concepts occurring in them". As an example, he again presents axioms of infinity, "which only explicate the content of the general concept of set". These reflections on axioms of infinity and their impact on provability are foreshadowed in (Gödel, 1947, p. 182), where Gödel asserts that the current axioms of set theory "can be supplemented without arbitrariness by new axioms which are only the natural continuation of the series of those [axioms of infinity] set up so far". So, there may be a completeness theorem stating, "every proposition expressible in set theory is decidable from the present axioms plus some true assertion about the largeness of the universe of all sets".

[17] For a broad discussion of Gödel's reflections on "absolutely unsolvable problems", cf. (Feferman, 2006a; Kennedy, van Atten, 2004; 2009).

Though Gödel calls the existence of an unexplored series of axioms a fact, he asserts also that the process of forming such a series does not yet form a "well-defined procedure which could actually be carried out (and would yield a non-recursive number-theoretic function)", because it would require "a substantial advance in our understanding of the basic concepts of mathematics" (Gödel, 1972, p. 306). A *prima facie* startlingly different reason for not yet having a precise definition of such a procedure is given in the 1974-note, p. 325: it would require "a substantial deepening of our understanding of the basic operations of the mind". That is only prima facie different, as Gödel's 1972-note connects such a procedure with the dynamic development of the human mind.

> [M]ind, in its use, is not static, but constantly developing, i.e., that we understand abstract terms more and more precisely as we go on using them, and that more and more abstract terms enter the sphere of our understanding. (Gödel, 1972, p. 306)[18]

Gödel continues:

> There may exist systematic methods of actualizing this development, which could form part of the procedure. Therefore, although at each stage the number and precision of the abstract terms at our disposal may be finite, both [...] may converge toward infinity in the course of the application of the procedure.

The procedure mentioned as a plausible candidate for satisfying this description is again the process of forming ever stronger axioms of infinity.

The notes (1972) and (1974) are very closely connected, but there is a subtle and yet, it seems to me, substantive difference. In the 1974-note the claim that the number of possible states of mind may converge to infinity is a consequence of the dynamic development of mind. That claim is followed by a remark that begins in a superficially similar way as the first sentence of the above quotation, but ends with a quite different observation: "Now there may exist systematic methods of accelerating, specializing, and uniquely determining this development, e.g. by asking the right questions on the basis of a mechanical procedure" (Gödel 1974, p. 325).

[18] Gödel's brief exploration of the issue of defining a non-mechanical, but effective procedure is preceded in this note by a severe critique of Turing. He *assumes* that Turing's argument in the 1936 paper was to show that "mental procedures cannot go beyond mechanical procedures" and considers it as inconclusive, because Turing neglects the dynamic nature of mind. However, simply carrying out a mechanical procedure does not, and indeed should not, involve an expansion of our understanding. Turing viewed the restricted use of mind in computations undoubtedly as static. I leave that misunderstanding out of the systematic considerations in the main text. The appeal to finiteness of states of mind when comparing Gödel's and Turing's perspectives is also pushed into the background as it is not crucial at all for the central issues under discussion: there does not seem to be any disagreement.

I do not fully understand these enigmatic observations, but three points can be made. First, mathematical experience has to be invoked when asking the right questions; second, aspects of that experience may be codified in a mechanical procedure and serve as the basis for asking the right questions; third, the answers may involve abstract terms that are introduced by the non-mechanical mental procedure. We should not dismiss or disregard Gödel's methodological remark that "asking the right questions on the basis of a mechanical procedure" may be part of a systematic method to push forward the development of mind.[19] Even this very limited understanding allows us to see that Gödel's reflections overlap with Turing's proposal for investigating matters in a more empirical and directly computational manner.

Much of Turing's work of the late 1940s and early 1950s explicitly deals with mental processes. But nowhere is it claimed that the latter cannot go beyond mechanical ones. Mechanical processes are still made precise as Turing machine computations; in contrast, machines that might exhibit intelligence have a more complex structure than Turing machines and, most importantly, interact with their environment. Conceptual idealization and empirical adequacy are now being sought for quite different purposes, and one might even say that Turing is actually trying to capture what Gödel described when searching for a broader concept of humanly effective calculability, namely, "… that mind, in its use, is not static, but constantly developing". In his paper *Intelligent Machinery*, Turing states:

> If the untrained infant's mind is to become an intelligent one, it must acquire both discipline and initiative. So far we have been considering only discipline [via the universal machine]. […] But discipline is certainly not enough in itself to produce intelligence. That which is required in addition we call initiative. This statement will have to serve as a definition. Our task is to discover the nature of this residue as it occurs in man, and to try and copy it in machines. (Turing, 1948, p. 21)[20]

How, in particular, can we transcend discipline when doing mathematics? Turing provided a hint already in his 1939-paper, where ordinal logics are introduced to expand formal theories in a systematic way; (cf. Feferman, 1988; 2006b) for informative discussions. In that paper, his Ph.D. thesis written under the di-

[19] There seems to be also a connection to remarks in his (1947, pp. 182–183), where Gödel points out that there may be "another way" (apart from judging its intrinsic necessity) to decide the truth of a new axiom. This other way consists in inductively studying its success, "that is, its fruitfulness in consequences and in particular in 'verifiable' consequences, i.e., consequences demonstrable without the new axiom, whose proofs by means of the new axiom, however, are considerably simpler and easier to discover, and make it possible to condense into one proof many different proofs".

[20] In his (1950, p. 459), Turing points out, in a similar spirit: "Intelligent behaviour presumably consists in a departure from the completely disciplined behaviour involved in computation, but a rather slight one, which does not give rise to random behaviour, or to pointless repetitive loops".

rection of Church, Turing distinguishes between ingenuity and intuition. He observes that in formal logics their respective roles take on a greater definiteness. Intuition is used for "setting down formal rules for inferences which are always intuitively valid", whereas ingenuity is to "determine which steps are the more profitable for the purpose of proving a particular proposition". He notes:

> In pre-Gödel times it was thought by some that it would be possible to carry this programme to such a point that all the intuitive judgements of mathematics could be replaced by a finite number of these rules. The necessity for intuition would then be entirely eliminated. (Turing, 1939, p. 209)

That intuition cannot be eliminated, on account of the first incompleteness theorem, is emphasized in Turing's letters to Max Newman from around 1940 that have been reprinted in (Copeland, 2004, pp. 211–216). After all, one can determine the truth of the Gödel sentence, say, for ZF set theory, despite the fact that it is independent of ZF. Providing a general reason for such a determination, Turing writes, "… there is a fairly definite idea of a true formula which is quite different from the idea of a provable one" (p. 215). Eight years later, in his (1948, p. 107), Turing formulated at the very outset reasons given by some for asserting, "it is not possible for machinery to show intelligent behaviour [sic!]". One of the reasons is directly related to the limitative theorems. They are assumed to show that when machines are used for "determining the truth or falsity of mathematical theorems […] then any given machine will in some cases be unable to give an answer at all". This inability of any particular machine is contrasted with human intelligence that "seems to be able to find methods of ever-increasing power for dealing with such problems 'transcending' the methods available to machines" (Turing, 1948, p. 108).

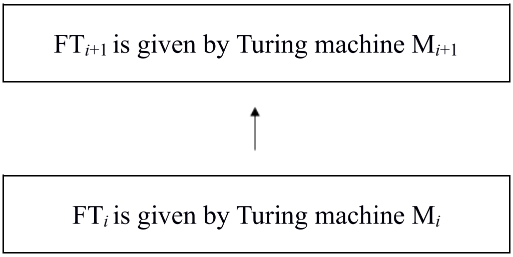
It is thus not surprising that Turing takes in his paper (1950, pp. 444–445) the mathematical objection to his view quite seriously. He considers the objection as based on the limitative results, in particular Gödel's theorems, which are understood by some as proving "a disability of machines to which the human intellect is not subject". Turing gives two responses. The short one states that the objection takes for granted, without any sort of proof, that the human intellect is not subject to the limitations to which machines provably are. However, Turing thinks that the objection cannot be dismissed quite so lightly and proceeds to a second response. It acknowledges the superiority of the human intellect with respect to a single machine (we can recognize the truth of "its" Gödel sentence), but Turing views that as a petty triumph. The reason for this is formulated succinctly as follows: "There would be no question of triumphing simultaneously over all machines. In short, then, there might be men cleverer than any given machine, but then there might be other machines cleverer again, and so on" (Turing, 1950, p. 445).

Turing does not offer a proof of the claim that there is "no question of triumphing simultaneously over all machines". It is precisely here that Gödel's "fact" concerning a humanly effective, but non-mechanical procedure seems to be in conflict with Turing’s assertion.[21] If the “fact” were a fact, then it would sustain the objection successfully. Can one go beyond claim and counterclaim? Or, even better, can one use the tension as an inspiration for concrete work that elucidates the situation?

IV. Finding Proofs (With Ingenuity)

Let us return, as a first positive step towards bridging the gap between claim and counterclaim, to Turing’s distinction between ingenuity and intuition. Intuition is explicitly linked to the incompleteness of formal theories and provides an entry point to exploiting, through computational work, a certain parallelism between Turing’s and Gödel’s considerations, when the latter are based on mechanical procedures. Copying the residue in machines is the common task at hand. It is a difficult one in the case of mathematical thinking, and Gödel would argue an impossible one, if machines are particular Turing machines. Turing would agree, of course. Before we can start copying, we have to discover partially the nature of the residue; one might hope to begin doing that through proposals for finding proofs in mathematics.

In his lecture to the London Mathematical Society and in *Intelligent Machinery*, Turing calls for heuristically guided intellectual searches and for initiative that includes, in the context of mathematics, proposing new intuitive steps. Such searches and the discovery of novel intuitive steps would be at the center of “research into intelligence of machinery”. Let me draw a diagram: the formal theory FT$_i$ has been expanded to the proof theoretically stronger theory FT$_{i+1}$; the theories are presented via Turing machines M$_i$ and M$_{i+1}$, respectively.


FT$_{i+1}$ is given by Turing machine M$_{i+1}$
FT$_i$ is given by Turing machine M$_i$

[21] “Seems”, as Turing pits individual men against particular machines, whereas Gödel pits the “human mind” against machines. This aspect is also briefly discussed in the first letter to Newman in (Copeland, 2004, p. 215): if one moves away from considering a particular machine and allows machines with different sets of proofs, then “by choosing a suitable machine one can approximate ‘truth’ by ‘provability’ better than with a less suitable machine, and can in a sense approximate it as well as you please”.

The transition from one theory to the next and, correspondingly, from one Turing machine to the next is non-mechanical for Gödel as well as for Turing. In Gödel's case, unfolding the explication of the concept of set by a non-mechanical method is the basis for a humanly effective procedure. Even if Gödel's method would take into account a mechanical procedure of the character described above, in the end, it would present a new and stronger axiom of infinity; it is in this sense that the method could be viewed as uniform. For Turing, it seems, the addition of intuitive steps (outside of his ordinal logics) is principally based on the analysis of machine learning and computer experimentation.[22] It would be closely tied to the particulars of a situation without the connecting thread of Gödel's method and, thus, it would not be uniform. In addition, Turing emphasizes at a number of places that a random element be introduced into the development of machines, thus providing an additional feature that releases them from strict discipline and facilitates a step from M_i to M_{i+1}.

What is striking is that both Gödel and Turing make "completeness claims": at the end of the second paragraph of section III, I quoted Gödel's remark from his 1947-paper that every set theoretic statement is decidable from the current axioms together with "a true assertion about the largeness of the universe of all sets"; in note 20, Turing's remark is quoted that by choosing a suitable machine one can approximate "truth" by "provability" and "in a sense approximate it [truth] as well as you please". That is highly speculative in both cases; slightly less speculatively, Turing conjectured:

> As regards mathematical philosophy, since the machines will be doing more and more mathematics themselves, the centre of gravity of the human interest will be driven further and further into philosophical questions of what can in principle be done etc. (1947, p. 103)

This expectation has not been borne out yet, and Gödel would not be surprised. However, he could have cooperated with Turing on the "philosophical questions of what can in principle be done" and, to begin with, they could have agreed terminologically that there is a human mind whose working is not reducible to the working of any particular brain. They could have explored and, possibly argued about, Turing's contention in his (1951, p. 472) "that machines can be constructed which will simulate the behaviour (sic!) of the human mind very closely". Indeed, Turing had taken a step toward a concept of human mind, when he emphasizes at the end of *Intelligent Machinery*, "the isolated man does not develop any intellectual power", and then argues:

[22] Copeland, in his (2006), gives much the same interpretation. He remarks on p. 168: "In his post-war writing on mind and intelligence […] the term "intuition" drops from view and what comes to the fore is the closely related idea of *learning*—in the sense of devising and discovering—new methods of proof".

> It is necessary for him to be immersed in an environment of other men, whose techniques he absorbs during the first twenty years of his life. He may then perhaps do a little research of his own and make a very few discoveries which are passed on to other men. From this point of view the search for new techniques must be regarded as carried out by the human community as a whole, rather than by individuals. (p. 127)

Turing calls this, appropriately enough, a cultural search in contrast to the more limited intellectual searches possible for individual men or machines. To build machines that think serves also another purpose as Turing explained in a 1951 radio broadcast: "The whole thinking process is still rather mysterious to us, but I believe that the attempt to make a thinking machine will help us greatly in finding out how we think ourselves" (Turing, 1951b, p. 486).

For the study of human thinking mathematics is a marvelous place to start. Where else do we find an equally rich body of rigorously organized knowledge that is structured for both intelligibility and discovery? Turing, as we saw above, had high expectations for machines' progress in doing mathematics; but it is still extremely difficult for them to "mathematize" on their own. Newman, in a radio debate with Braithwaite, Jefferson, and Turing, put the general problem very well:

> Even if we stick to the reasoning side of thinking, it is a long way from solving chess problems to the invention of new mathematical concepts or making a generalisation (sic!) that takes in ideas that were current before, but had never been brought together as instances of a single general notion. (Turing, 1952, p. 498)

The important question is whether we can gain, by closely studying mathematical practice, a deeper understanding of fundamental concepts, techniques and methods of mathematics and, in that way, advance our understanding of the capacities of the mathematical mind as well as of basic operations of the mind. This question motivates a more modest goal, namely, formulating strategies for an automated search: not for proofs of new results, but for proofs that reflect logical and mathematical understanding; proofs that reveal their intelligibility and that force us to make explicit the ingenuity required for a successful search.[23] The logical framework for such studies must include a structural

[23] This involves undoubtedly reactions to Turing's remarks and impatient questions in a letter to Newman: "In proofs there is actually an enormous amount of sheer slogging, a certain amount of ingenuity, while in most cases the actual 'methods of proof' are quite well known. Cannot we make it clearer where the slogging comes in, where there is ingenuity involved, and what are the methods of proof"? (Copeland, 2004, p. 213). Abramson, in his (2008), emphasizes insightfully the significance of Lady Lovelace's objection. In the context here, his emphasis pointed out to me that Turing (1950, p. 451), views "the mere working out of consequences from data and general principles" as a "virtue" and as a "source for surprises". Turing articulates that important perspective after having called "false" the assumption that "as soon as a fact is presented to a mind all consequences of the fact spring into the mind simultaneously with it".

theory of proofs that extends proof theory through (i) articulating structural features of derivations and (ii) exploiting the meaning of abstract concepts; both aspects are crucial for finding humanly intelligible proofs.[24] We will hopefully find out what kind of broad strategies and heuristic ideas will emerge, what is the necessary ingenuity. In this way, we will begin to uncover part of Turing's residue and part of what Gödel considered as humanly effective, but not mechanical, in each case "by asking the right questions on the basis of a mechanical procedure" (Gödel, 1974, p. 325).

The very last remark in (Turing, 1954) comes back, in a certain sense, to the mathematical objection. Turing views the limitative results as being "mainly of a negative character, setting bounds to what we can hope to achieve purely by reasoning". Characterizing in a new way the residue that has to be discovered and implemented to construct intelligent machinery, Turing continues, "These, and some other results of mathematical logic may be regarded as going some way towards a demonstration, within mathematics itself, of the inadequacy of 'reason' unsupported by common sense". This is as close as Turing could come to agree with Gödel's dictum "The human mind infinitely surpasses any finite machine", if "finite machine" is identified with "Turing machine".

Acknowledgments

Versions of this essay were read at the *Computability in Europe* conference in 2006, as the Gödel Lecture at the *Colloquium Logicum 2006*, and at the Logic Seminar of the Mathematics Department, University of Lisbon. It is closely based on previous papers of mine (Sieg, 2006; 2007), but I argue here for a parallelism between Gödel's and Turing's considerations for transcending purely mechanical processes in mathematics. I revised and completed the manuscript in March 2009, while I was a Fellow at the Swedish Collegium for Advanced Study in Uppsala, but made some stylistic changes two years later. Finally, remarks by Jack Copeland and Oron Shagrir stimulated a sharpening and expansion of points in the last section; that version served as the basis for my talk at the Studia Logica Conference "Church's Thesis: Logic, Mind and Nature" in Krakow (June 3–5, 2011).

[24] I have been pursuing a form of such a structural proof theory for quite a number of years. Central considerations and results are presented in (Sieg, 2010); there I also pointed out connections with Greek mathematics and the radical transformation of mathematics in the nineteenth century, as described in (Stein, 1988). A fully automated proof search method for (classical) first-order logic has been implemented in the AProS system. The overall project, addressing strategic search and dynamic tutoring, is being extended now also to elementary set theory; it is described at http://www.phil.cmu.edu/projects/apros/, and AProS is downloadable from that site.

REFERENCES

Abramson, D. (2008). Turing's Responses to Two Objections. *Minds & Machines*, *18*(2), 147–167.

Ackermann, W. (1924). Begründung des 'tertium non datur' mittels der Hilbertschen Theorie der Widerspruchsfreiheit. *Mathematische Annalen*, *93*, 1–36.

Church, A. (1935). An Unsolvable Problem of Elementary Number Theory. *Bulletin of the AMS*, *41*, 332–333.

Copeland, J. (2004). *The Essential Turing: The Ideas That Gave Birth to the Computer Age*. Oxford: Clarendon Press.

Copeland, J. (2006). Turing's Thesis. In: A. Olszewski, J. Wolenski, R. Janusz (Eds.), *Church's Thesis After 70 Years* (pp. 147–174). Frankfurt: Ontos Verlag.

Davis, M. (1982). Why Gödel Did Not Have Church's Thesis. *Information and Control*, *54*, 3–24.

Dawson, John W. (2006). *Gödel and the Origin of Computer Science*. Lecture at CiE 2006, Swansea.

Dedekind, R. (1888). *Was sind und was sollen die Zahlen?* Braunschweig: Vieweg.

Feferman, S. (1988). Turing in the Land of O(z). In: R. Herken (Ed.), *The Universal Turing Machine—A Half-Century Survey* (pp. 113–147). Oxford: Oxford University Press.

Feferman, S. (2006a). Are There Absolutely Unsolvable Problems? Gödel's Dichotomy. *Philosophia Mathematica*, *14*(2), 134–152.

Feferman, S. (2006b). Turing's Thesis. *Notices of the AMS*, *53*(10), 1200–1205.

Gandy, R. (1980). Church's Thesis and Principles for Mechanisms. In: J. Barwise, H. J. Keisler, K. Kunen (Eds.), *The Kleene Symposium* (pp. 123–148), Amsterdam: North-Holland Publishing Company.

Gödel, K. (1931). Über formal unentscheidbare Sätze der *Principia Mathematica* und verwandter Systeme I [On Formally Undecidable Propositions of Principia Mathematica and Related Systems I]. *Monatshefte für Mathematik und Physik*, *38*, 173–198.

Gödel, K. (1933). The Present Situation in the Foundations of Mathematics. In: K. Gödel, *Collected Works III* (pp. 45–53). Oxford: Oxford University Press.

Gödel, K. (1934). On Undecidable Propositions of Formal Mathematical Systems. In: K. Gödel, *Collected Works I* (pp. 346–371). Oxford: Oxford University Press.

Gödel, K. (1936). Über die Länge von Beweisen. *Ergebnisse eines mathematischen Kolloquiums*, Heft 7, 23–24.

Gödel, K. (1939). Finding Aid (Notre Dame Lecture on Logic, Spring 1939). In: K. Gödel, *Collected Works IV–V* (pp. 527–528). Oxford: Oxford University Press.

Gödel, K. (193?). Undecidable Diophantine Propositions. In: K. Gödel, *Collected Works III* (pp. 164–175). Oxford: Oxford University Press.

Gödel, K. (1946). Remarks Before the Princeton Bicentennial Conference on Problems in Mathematics. In: K. Gödel, *Collected Works II* (pp. 150–153). Oxford: Oxford University Press.

Gödel, K. (1947). What is Cantor's Continuum Problem? In: K. Gödel, *Collected Works II* (pp. 167–187). Oxford: Oxford University Press.

Gödel, K. (1951). Some Basic Theorems on the Foundations of Mathematics and Their Implications. In: K. Gödel, *Collected Works III* (pp. 304–323). Oxford: Oxford University Press.

Gödel, K. (1964). Postscriptum for [1934]. In: K. Gödel, *Collected Works I* (pp. 369–371). Oxford: Oxford University Press.

Gödel, K. (1972). Some Remarks on the Undecidability Results. In: K. Gödel, *Collected Works II* (pp. 305–306). Oxford: Oxford University Press.

Gödel, K. (1974). Note. In: H. Wang, *From Mathematics to Philosophy* (pp. 325–326). London: Routledge & Kegan Paul.

Gödel, K. (1986). *Collected Works I*. Oxford: Oxford University Press.

Gödel, K. (1990). *Collected Works II*. Oxford: Oxford University Press.

Gödel, K. (1995). *Collected Works III*. Oxford: Oxford University Press.

Gödel, K. (2003). *Collected Works IV–V*. Oxford: Oxford University Press.

Herbrand, J. (1931). On the Consistency of Arithmetic. In: W. Goldfarb (Ed.), *Jacques Herbrand, Logical Writings* (pp. 282–298). Cambridge, Mass.: Harvard University Press.

Herken, R. (Ed.). (1988). *The Universal Turing Machine—A Half-Century Survey*. Oxford: Oxford University Press.

Hilbert, D. (1900). Über den Zahlbegriff. *Jahresbericht der Deutschen Mathematiker Vereinigung*, *8*, 180–194.

Hilbert, D. (1905). Über die Grundlagen der Logik und der Arithmetik. In: A. Krazer (Ed.), *Verhandlungen des 3. Internationalen Mathematiker-Kongresses: in Heidelberg vom 8. bis 13* (pp. 174–185). Leipzig: Teubner.

Hilbert, D. (1923). Die logischen Grundlagen der Mathematik. *Mathematische Annalen*, *88*, 151–165.

Hilbert, D., Bernays, P. (1921/2). Grundlagen der Mathematik. in: W. Ewald, W. Sieg (Eds.), *David Hilbert's Lectures on the Foundations of Arithmetic and Logic 1917–1933* (pp. 431–521). Springer.

Hilbert, D., Bernays, P. (1934). *Grundlagen der Mathematik I*. Berlin: Springer Verlag.

Hilbert, D., Bernays, P. (1939). *Grundlagen der Mathematik II*. Berlin: Springer Verlag.

Kennedy, J., van Atten, M. (2004). Gödel's Modernism: On Set-Theoretic Incompleteness. *Graduate Faculty Philosophy Journal*, *25*(2), 289–349.

Kennedy, J., van Atten, M. (2009). "Gödel's Modernism: On Set-Theoretic Incompleteness", Revisited. In: S. Lindström et al. (Eds.), *Logicism, Intuitionism, and Formalism—What Has Become of Them?* (Synthese Library 341, pp. 303–355). Springer.

Kolmogorov, A., Uspenski, V. (1963). On the Definition of an Algorithm. *AMS Translations*, *21*(2), 217–245.

Martin, D. A. (2005). Gödel's Conceptual Realism. *Bulletin of Symbolic Logic*, *11*(2), 207–224.

Post, E. (1947). Recursive Unsolvability of a Problem of Thue. *J. Symbolic Logic*, *12*, 1–11.

Sieg, W. (1994). Mechanical Procedures and Mathematical Experience. In: A. George (Ed.), *Mathematics and Mind* (pp. 71–117). Oxford: Oxford University Press.

Sieg, W. (1997). Step by Recursive Step: Church's Analysis of Effective Calculability. *Bulletin of Symbolic Logic*, *3*(2), 154–180.

Sieg, W. (1999). Hilbert's Programs: 1917–1922. *Bulletin of Symbolic Logic*, *11*(2), 1–44.

Sieg, W. (2002). Calculations by Man and Machine: Conceptual Analysis. In: W. Sieg, R. Sommer, C. Talcott (Eds.), *Reflections on the Foundations of Mathematics* (Lecture Notes in Logic 15, pp. 390–409). Cambridge University Press.

Sieg, W. (2006). Gödel on Computability. *Philosophia Mathematica*, *14*(2), 189–207.

Sieg, W. (2007). On Mind & Turing's Machines. *Natural Computing*, *6*, 187–205.

Sieg, W. (2009a). On Computability. In: A. D. Irvine (Ed.), *Philosophy of Mathematics* (pp. 535–630). Amsterdam: North Holland.

Sieg, W. (2009b). Hilbert's Proof Theory. In: D. M. Gabbay, J. Woods (Eds.), *Handbook of the History of Logic. Volume 5: Logic from Russell to Church* (pp. 321–384). Amsterdam: North Holland.

Sieg, W. (2010). Searching for Proofs (And Uncovering Capacities of the Mathematical Mind). In: S. Feferman, W. Sieg (Eds.), *Proofs, Categories and Computations* (pp. 189–215). London: College Publications.

Sieg, W., Byrnes, J. (1996). K-Graph Machines: Generalizing Turing's Machines and Arguments. In: P. Hájek (Ed.), *Godel '96: Logical Foundations of Mathematics, Computer Science and Physics* (Lecture Notes in Logic 6, pp. 98–119). Berlin: Springer Verlag.

Sieg, W., and D. Schlimm (2005). Dedekind's Analysis of Number: Systems and Axioms. *Synthese*, *147*, 121–170.

Stein, H. (1988). Logos, Logic, and Logistiké: Some Philosophical Remarks on Nineteenth-Century Transformation of Mathematics. In: W. Aspray, P. Kitcher (Eds.), *History and Philosophy of Modern Mathematics* (vol. XI of Minnesota Studies in the Philosophy of Science, pp. 238–259). Minneapolis: University of Minnesota Press.

Turing, A. M. (1936). On Computable Numbers, With an Application to the Entscheidungsproblem. *Proc. London Math. Soc.* (series 2), *42*, 230–265.

Turing, A. M. (1939). Systems of Logic Based on Ordinals. *Proc. London Math. Soc.* (series 2), *45*, 161–228.

Turing, A. M. (1947). Lecture to the London Mathematical Society on 20 February 1947. In: D. C. Ince (Ed.), *Collected Works of A.M. Turing—Mechanical Intelligence* (87–105). Amsterdam: North Holland.

Turing, A. M. (1948). Intelligent Machinery. In: D. C. Ince (Ed.), *Collected Works of A.M. Turing—Mechanical Intelligence* (107–127). Amsterdam: North Holland.

Turing, A. M. (1950). Computing Machinery and Intelligence. *Mind*, *59*, 433–460.

Turing, A. M. (1951a). Intelligent Machinery, A Heretical Theory [Radio broadcast]. Printed in J. Copeland, *The Essential Turing: The Ideas That Gave Birth to the Computer Age* (pp. 472–475). Oxford: Clarendon Press.

Turing, A. M. (1951b). Can Digital Computers Think? [Radio broadcast]. Printed in J. Copeland, *The Essential Turing: The Ideas That Gave Birth to the Computer Age* (pp. 482–486). Oxford: Clarendon Press.

Turing, A. M. (1952). Can Automatic Calculating Machines Be Said to Think? [Radio discussion of A. Turing, R. Braithwaite, G. Jefferson, and M. Newman. Printed in J. Copeland, *The Essential Turing: The Ideas That Gave Birth to the Computer Age* (pp. 494–506). Oxford: Clarendon Press.

Turing, A. M. (1954). Solvable and Unsolvable Problems. *Science News*, *31*, 7–23.

van Heijenoort, J. (1968). *From Frege to Gödel—A Source Book in Mathematical Logic, 1879–1931*. Cambridge, Mass.: Harvard University Press.

Wang, H. (1974*). From Mathematics to Philosophy*. London: Routledge & Kegan Paul.

NEW REFERENCES (FOR THE POSTSCIPTUM)

Bernays, P. (1930). Die Philosophie der Mathematik und die Hilbertsche Beweistheorie. In: P. Bernays, *Abhandlungen zur Philosophie der Mathematik* (pp. 17–61). Darmstadt: Wissenschaftliche Buchgesellschaft.

Davis, M., Sieg, W. (2015). Conceptual Confluence in 1936: Post and Turing. In: G. Sommaruga, T. Strahm (Eds.), *Turing's Revolution* (pp. 3–27). Berlin: Springer.

Gentzen, G. (1936). Die Widerspruchsfreiheit der reinen Zahlentheorie. *Mathematische Annalen*, *112*(1), 493–565.

Hilbert, D. (1918). Axiomatisches Denken. *Mathematische Annalen*, *78*, 405–415.

Hilbert, D., Ackermann, W. (1928). *Grundzüge der theoretischen Logik*. Berlin: Springer.

Hilbert, D., Bernays, P. (1917–18). Prinzipien der Mathematik. In: W. Ewald, W. Sieg (Eds.), *David Hilbert's Lectures on the Foundations of Arithmetic and Logic 1917–1933* (pp. 59–221). Berlin: Springer.

Sieg, W. (2013). Hilbert's Programs and Beyond. Oxford: Oxford University Press.

Sieg, W. (2014). The Ways of Hilbert's Axiomatics: Structural and Formal. *Perspectives on Science*, *22*(1), 133–157.

Sieg, W. (2018). What is the *Concept* of Computation? In: F. Manea, R.G. Miller, D. Nowotka (Eds.), *Sailing Routes in the World of Computation*, *Proceedings of CiE 2018* (pp. 386–396). Springer LNCS 10936.

Sieg, W., Morris, R. (2018). Dedekind's Structuralism: Creating Concepts and Deriving Theorems. In: E. H. Reck (Ed.), *Logic, Philosophy of Mathematics, and Their History: Essays in Honor of W.W. Tait* (pp. 251–301). London: College Publications.

Sieg, W., Walsh, P. (2019). Natural Formalization: Deriving the Cantor-Bernstein Theorem in ZF. *The Review of Symbolic Logic*. doi:10.1017/S175502031900056X

Sieg, W., Derakhshan, F. (2020). Human-Centered Automated Proof Search. Manuscript submitted for publication.

Sieg, W., Szabo, M., McLaughlin, D. (2016). Why Post Did [Not] Have Turing's Thesis. In: E. Omodeo, A. Policriti (Eds.), *Martin Davis on Computability, Computational Logic, and Mathematical Foundations* (pp. 175–208). Berlin: Springer.

Postscriptum

This essay was originally published in the volume *Computability—Turing, Gödel, Church, and Beyond*, MIT Press 2013. It is reprinted here with the permission of MIT Press. The current version is not literally the same essay, as I made a few minor stylistic changes. Three developments in my own thinking, since the completion of the essay in 2011, are worthwhile to point out and to describe briefly in this Postscriptum. The first provides a stronger connection to the past, the second is a further deepening of the analysis of the concept of computability, and the third yields a systematic connection to the future from the perspective of 2011.

There is then, first of all, a deeper historical understanding of the methodological basis for the investigations of Gödel and Turing. The crucial building blocks for that basis were provided by the radically new structuralist conception of mathematics in the work of Dedekind and Hilbert and the dramatically expanded reach of logic primarily through Frege's efforts; (Sieg & Morris, 2018). The mathematical work and the logical work were hardly connected when they were created during the last thirty years of the nineteenth century. After Whitehead and Russell had reshaped logic through *Principia Mathematica*, the two building blocks were joined and received a rigorous mathematical description in (Hilbert & Bernays, 1917–1918). These lectures are the beginning of modern mathematical logic and opened the door for metamathematical investigations in the 1920s; they are also, via (Hilbert & Ackermann, 1928), the backdrop for Gödel and Turing. The emergence of metamathematics took place during the first thirty years of the twentieth century; it is incisively described in (Bernays, 1930). Many people have contributed to a deeper historical understanding that is reflected in the first half of my book *Hilbert's Programs and Beyond*. The shift from structural to formal axiomatics, absolutely central for Gödel and Turing, is elucidated in (Sieg, 2014). Book and paper contain, of course, references to the rich literature.

The second development is a sharpening of my structural axiomatic approach in order to characterize computability as an abstract mathematical concept. That is alluded to in this essay at the end of Section II. It has an historical component that brings out the significance of Post's work (Sieg, Szabo & McLaughlin, 2016); it also uncovers the deep conceptual confluence of Post's and Turing's work in 1936, presented in (Davis & Sieg, 2015). Finally, in a paper that was dedicated to Davis' ninetieth birthday (Sieg, 2018), I raised and sought to answer the key methodological question, "What is the concept of computation?" Drawing on my earlier work, the concise answer is given in terms of computable dynamical systems. This is done against the background of two classes of mathematical results generalizing the considerations of Section I (Gödel's Absoluteness) and of Section II (Turing's Reducibility). The set theoretic formulation of the abstract concept "computable dynamical system" is waiting for an illuminating category theoretic characterization.

We finally come to the third development since 2011. It concerns neither the historical background for Sections I and II nor the axiomatic sharpening of the concept of computation. It is rather connected to the comparative analysis of Gödel's and Turing's suggestions for transcending mechanical procedures in Sections III and IV. The goals of that development are described in broad strokes in the penultimate paragraph of the essay and have been pursued within my AProS Project that is mentioned in Note 23. The latter seeks to find strategies for the automated search for humanly intelligible proofs in constructive and classical logic, but also in meta-mathematics (Gödel's incompleteness theorems) and set theory (the Cantor-Bernstein Theorem). My views on "natural formalization within a foundational frame" and "human-centered automated proof search" are at the center of and operative in (Sieg & Walsh, 2019), respectively (Sieg & Derakhshan, 2020).

The relevant theoretical perspective is this: formalizing mathematical practice is central for the significance of proof theoretic investigations, be they concerned with the consistency problem of formal theories or with the "mining" of particular proofs. We use refined, conceptually organized formal frameworks to reflect deep structures of mathematical proofs. Thus, we aim for a theory of proofs in which "ordinary" proofs are treated as objects of investigation. That is in the spirit of the pioneers. Hilbert remarked in (1918), "[w]e must—that is my conviction—take the concept of the specifically mathematical proof as an object of investigation". In just this spirit, Gentzen thought in his (1936, p. 499) that one can obtain only through formalization a "rigorous treatment of proofs" and emphasized then most strongly, "[t]he objects of proof theory shall be the proofs carried out in mathematics proper".

STUDIA SEMIOTYCZNE (SEMIOTIC STUDIES), t. XXXIV, nr 1 (2020), s. 81–82
ISSN 0137-6608, e-ISSN 2544-073X
DOI: 10.26333/sts.xxxiv1.04

Note

RUDY RUCKER [*]

A NOTE ON THE LUCAS ARGUMENT

This note is derived from my books *Infinity and the Mind* (2005, Preface) and *The Lifebox, the Seashell, and the Soul* (2016, footnote 102).

We're talking about J. Anthony Lucas's classic argument that Gödel's Second Incompleteness Theorem rules out man-machine equivalence. This is an argument that Penrose revived and popularized in the 1990s. This fallacious argument is a thoroughly dead horse. But I'll give it another beating here. Do note that the Lucas-Penrose argument is a completely distinct issue from Penrose-Hameroff speculation that the brain can act as a coherent quantum computer. It's to Penrose's credit that he's associated with multiple controversial ideas!

Before continuing, I should explain the Gödel's Second Incompleteness Theorem is the result that if F is a consistent formal system as strong as arithmetic, then F cannot prove the sentence $Con(F)$. $Con(F)$ is the sentence that expresses the consistency of F by asserting that F will never prove, say, $0 = 1$. If we think of h as being the index of the Turing machine Mh, we can write $Con(h)$ as shorthand for $Con(Mh)$.

Suppose h is an integer that codes the program for a device Mh whose output is very much like a person's. Lucas and Penrose want to say the following

(1) After hanging around with Mh for a while, any reasonable person will feel like asserting $Tr(h)$, a sentence which says something like, "If I base a machine Mh on the algorithm coded by h I'll get a machine which only ouputs true sentences about mathematics".

(2) Having perceived the truth of $Tr(h)$, any reasonable person will also feel like asserting $Con(h)$, a sentence which says something like, "If I base

[*] San Jose State University, Department of Computer Science. E-mail: rudy@rudyrucker.com. ORCID: 0000-0001-7679-3025.

a machine *Mh* on the algorithm coded by *h* I'll get a machine which never generates any mathematical contradictions".

(3) Gödel's Second Incompleteness Theorem shows that *Mh* can't prove *Con*(*h*), so now it looks as if any reasonable person who hangs around with a human-like *Mh* will soon know something that the machine itself can't prove.

The philosopher Hilary Putnam formulated what remains the best counterargument in his 1960 essay, *Minds and Machines* (1964). For Lucas's ripostes to such objections, see his genial if unconvincing essay, *A Paper Read to the Turing Conference at Brighton on April 6th, 1990* (Lucas, 1990).

Putnam's point is simple. Even if you have seen *Mh* behaving sensibly for a period of time, you still don't have any firm basis for asserting either that *Mh* will always say only true things about mathematics or that *Mh* will never fall into an inconsistency. Now if you were to have a full understanding of how *Mh* operates, then perhaps you could prove that *Mh* is consistent. But, in the case where *h* is the mind recipe, the operation of the eventual *Mh* is incomprehensibly intricate, and we will never be in a position to legitimately claim to know the truth of the sentence *Con*(*h*) which asserts that *Mh* is consistent. This is, indeed, the content of Gödel's Second Incompleteness Theorem. Rather than ruling out man-machine equivalence, the theorem places limits on what we can know about machines equivalent to ourselves.

And, really, this shouldn't come as a surprise. You can share an office or a house with a person *P* for fifteen years, growing confident in the belief that *P* is consistent, and then one day, *P* begins saying and doing things that are completely insane. You imagined that you knew *Con(P)* to be true, but this was never the case at all. The only solid reason for asserting *Con*(*P*) would have been a systematic proof, but, given that you and *P* were of equivalent sophistication, this kind of proof remained always beyond your powers. All along, the very fact that *Con*(*P*) wasn't provable contained the possibility that it wasn't true. Like it or not, that's the zone we operate in when relating to other intelligent beings.

REFERENCES

Lucas, J. R. (1990). A Paper Read to the Turing Conference at Brighton on April 6th, 1990. Retrieved from: http://users.ox.ac.uk/~jrlucas/Godel/brighton.html

Putnam, H. (1964). Minds and Machines. In: R. Anderson, *Minds and Machines* (pp. 43–59). Upper Saddle River: Prentice-Hall.

Rucker, R. (2005). *Infinity and the Mind* (3rd ed.). Princeton: Princeton University Press.

Rucker, R. (2016). The Lifebox, the Seashell, and the Soul. Edinburgh: Transreal Fiction.

STUDIA SEMIOTYCZNE (SEMIOTIC STUDIES), t. XXXIV, nr 1 (2020), s. 83–108
ISSN 0137-6608, e-ISSN 2544-073X
DOI: 10.26333/sts.xxxiv1.05
Article

ARNON AVRON [*]

THE PROBLEMATIC NATURE OF GÖDEL'S DISJUNCTIONS AND LUCAS-PENROSE'S THESES

SUMMARY: We show that the name "Lucas-Penrose thesis" encompasses several different theses. All these theses refer to extremely vague concepts, and so are either practically meaningless, or obviously false. The arguments for the various theses, in turn, are based on confusions with regard to the meaning(s) of these vague notions, and on unjustified hidden assumptions concerning them. All these observations are true also for all interesting versions of the much weaker (and by far more widely accepted) thesis known as "Gödel disjunction". Our main conclusions are that pure mathematical theorems cannot decide alone any question which is not purely mathematical, and that an argument that cannot be fully formalized cannot be taken as a mathematical proof.

KEYWORDS: Gödel disjunction, Lucas-Penrose argument, mechanism, mind, computationalism.

1. Introduction

When I was invited to contribute to this special issue about the Lucas-Penrose argument (LP), I was hesitating whether there is any point of doing so. There were two reasons for that.

- The arguments of Lucas and Penrose have been totally refuted several times in the past. (This was done in more than one way, but this is not be-

[*] Tel Aviv University, School of Computer Science. E-mail: aa@cs.tau.ac.il. ORCID: 0000-0001-6831-3343.

cause it is not clear what is wrong with them, but because they contain several clear mistakes, not just one.) Nevertheless, the debate continues, and it seems that it will continue forever. The reason is that Lucas-Penrose "proofs" that humans are not machines belong to the class I call "proofs for the believers". (They resemble in this respect the well-known classical "proofs" of the existence of God.) What is characteristic of such "proofs" is that they have never actually convinced anybody to accept their conclusion. The only persons who have ever "accepted" the validity of "proofs" of this kind were people who had believed their conclusion already before that, and because of other reasons. Thus even Lucas and Penrose do not deny the fact that almost every logician who wrote something about their "proofs" rejected them as invalid. This fact itself should have been sufficient for them (according to their own views about the nature of a mathematical proof) to realize that their proofs cannot be mathematically valid. Nevertheless, they (and the few philosophers who support them) continue to maintain that their argument is valid.[1] It seems that somehow, when it comes to their arguments, even people who Lucas and Penrose otherwise respect as brilliant logicians (including Gödel himself) suddenly become extremely stupid, and just cannot see the light of their unshakable logical arguments… I believe that in situations like this it makes no sense to continue arguing with the believers. In the words of Penrose (1989; which were said about "very dogmatic formalists"): we should now simply ignore the supporters of the arguments of Lucas and Penrose.

- It seems to me that practically everything worth saying about LP has by now been said. Therefore I was not sure that I can do more than repeating arguments and points already made by others. And indeed, almost everything I write below can be found in some form or another somewhere in the existing literature. (See, in particular, Feferman, 2006; Franzén, 2005; Koellner, 2016; LaForte, Hayes, & Ford, 1998; Putnam, 2011; Shapiro, 1998; 2016.)

Nevertheless, after reading a great part of the related literature, I realized that there are still important aspects of the debate that have not got sufficient attention so far. Accordingly, the main goals of this paper is to explicitly state, and to provide strong evidence for, the following claims:

1. Pure mathematical theorems cannot decide alone any question which is not purely mathematical. For this reason it should have been clear from the start, that the "mathematical refutations" of the mechanistic thesis about the mind, given by Lucas and Penrose, cannot be sound. Any such refutation should depend also on some non-mathematical assumptions. This principle seems to me self-evident. Yet

[1] Or at least "is, in essence, correct", as Penrose wrote in (1994).

even Gödel has done in (1951), the logical mistake of attributing the honor of being a "mathematically established fact" to a disjunction of LP with another far-fetched thesis. This claim of Gödel about the human mind is now called "Gödel Disjunction" (GD) in (Horsten & Welch, 2016a), and "Gödel Dichotomy" in (Feferman, 2006). In (Horsten & Welch, 2016b, p. 3) it is stated that in contrast to Lucas-Penrose thesis, "Gödel's argument for his disjunctive thesis is highly compelling" and that "In the literature on the subject there is a consensus that Gödel's arguments for his disjunction are definitive".[2] Accordingly, this paper is mainly devoted to a critical discussion of GD rather than to LP. Needless to say, rejecting the former implies rejecting also the latter.

2. A crucial factor in the debate on LP that I have never seen explicitly stated, and is perhaps the main reason that it is such an Hydra, is that there is no single "Lucas-Penrose thesis", but there are several Lucas-Penrose theses. Different authors, or the same author in different places (frequently within one paper) provide different formulations of the thesis that (as we are going to argue) cannot be taken as equivalent. Since LP is one of the two disjuncts in GD, the situation with the latter is even worse. As we show in the sequel, we can even find in the literature purely mathematical formulations of it which indeed follow (trivially) from the theorems of Gödel and Tarski. Unfortunately, those formulations have very little interest for themselves. GD has of course also very interesting formulations, that try to say something significant on the nature of human beings. However, the more interesting a formulation is, the less clear is what it says, and the more doubtful are the non-mathematical assumptions that underlie it.
3. The arguments for the various Lucas-Penrose theses, as well those for the non-trivial versions of GD, are based on confusions concerning the terminology employed. Therefore those arguments include hidden, unjustified assumptions. In the words of Koellner in (2016, p. 1): "One problem with the discussion in the literature as it currently stands is that the background assumptions on the underlying concepts (like truth, absolute provability, and idealized human knowability) are seldom fully articulated".

2. Formulations of the Two Disjuncts

We start with a list of some formulations of the two disjuncts that have been given in the literature. The list is far from being exhaustive, but it is sufficiently

[2] I do not know on what basis this claim abut "consensus" is made. (Horsten & Welch, 2016b) is an introduction to (Horsten & Welch, 2016a), and in this book alone Gödel Disjunction is severely criticized in three different papers (Koellner, 2016; Shapiro, 2016; Williamson, 2016). Strong criticism of GD appeared also in (Boolos, 1995; Feferman, 2006; Franzén, 2005).

diverse to do for our purposes. From the discussions in the sequel it follows that no two of the formulations in it are really equivalent.

2.1. The First Disjunct ("Lucas-Penrose Theses")

1-Gödel-A The human mind cannot be reduced to the working of the brain. (Gödel, 1951)

1-Lucas The human mind is not equivalent to a (finite) machine. (Lucas, 1961)[3]

1-Krajewski The operation of the mind in the field of arithmetics cannot be simulated by a machine. (Krajewski, 2020)

1-Penrose-A The human mind is not a Turing machine. (Penrose, 1989; 1994)

1-Horsten-Welch-A There is no algorithm that can produce all the theorems that the human mind is capable of producing. (Horsten & Welch, 2016b)

1-Koellner-A The mathematical outputs of the idealized human mind cannot coincide with the mathematical outputs of an idealized finite machine. (Koellner, 2016; 2018a; 2018b)

1-Koellner-B The mathematical outputs of an idealized human mind cannot coincide with the mathematical outputs of any idealized finite machine. (Koellner, 2016; 2018a; 2018b)

1-Penrose-B Human understanding is something that cannot be reduced to computation. (Penrose, 2011)

1-Horsten-Welch-B The collection of humanly knowable theorems cannot be recursively axiomatized in some formal theory. (Horsten & Welch, 2016b)

1-Gödel-B No well-defined system of correct axioms can contain the system of all demonstrable mathematical propositions. (Gödel, 1951)

1-Charlesworth No computer program can accurately simulate the input-output properties of human mathematical reasoning. (Charlesworth, 2016)

1-Gödel-C Mathematics is incompletable in this sense, that its evident axioms can never be comprised in a finite rule. (Gödel, 1951)

1-Shipman Define "Human mathematics" as the collection of formalized sentence in the language of set theory which are logical consequences of statements that will eventually come to be accepted by a consensus of human mathematicians as "true". There is no r.e. consistent r.e. set which equals (or at least contains) Human mathematics. (From a message to FOM, August 2006)

[3] In (Godel, 1951, p. 310), this claim is formulated in stronger words: "The human mind (even within the realm of pure mathematics) infinitely surpasses the power of any finite machine".

2.2. The Second Disjunct

2-Koellner There are mathematical truths that cannot be proved by the idealized human mind. (Koellner, 2016; 2018a; 2018b)

2-Gödel There are absolutely undecidable [Diophantine] problems (Gödel, 1951).

Other, more or less equivalent versions of this thesis are:

- There are objective (mathematical) truths that can never be humanly demonstrated. (Feferman, 2006)
- Mathematical truth outstrips human reason. (Koellner, 2016; 2018a; 2018b)
- There exists a particular true arithmetic statement that is impossible for human mathematical reasoning to master. (Charlesworth, 2016)

2-Shipman There are mathematical truths that do not belong to "Human mathematics". (From the message to FOM cited above.)

3. The Mathematically Valid "Gödel Disjunction"

Let **T** be a second-order constant, to be interpreted as the set of the true sentences in the language $\mathcal{L}_{PA}$ of Peano arithmetics. Let F and S be second-order variables for sets of arithmetical sentences (not necessarily subsets of **T**!). Finally, let *formal*(S) be a second-order formula which says that S is the set of theorems of some formal system. Then Tarski's theorem implies:

$$\forall F(formal(F) \rightarrow \mathbf{T} \neq F)$$

This, in turn, is logically equivalent to:

$$\text{(MGD)} \qquad \forall S(S \neq \mathbf{T} \vee \forall F(formal(F) \rightarrow S \neq F))$$

(MGD) is the purely mathematical formulation of "Gödel Disjunction". Since it is just a trivial corollary of Tarski's theorem about the arithmetic undefinability of arithmetic truth, it is for itself not very interesting. However, Gödel and others add here one more step. Denoting by **K** "the system of all humanly demonstrable mathematical propositions", they infer from (MGD):

$$\forall F(formal(F) \rightarrow \mathbf{K} \neq F) \vee \mathbf{K} \neq \mathbf{T}$$

Getting by this the disjunction of [1-Gödel-B] and [2-Gödel]: either the set of humanly demonstrable theorems cannot be axiomatized by any effectively given formal system, or there are absolutely undecidable problems. However, the last inference is logically valid only provided that "the system of all demonstrable mathematical propositions" is well-defined. Personally, I do not see any reason to think so. In any case, this question is not a purely mathematical one. Therefore it cannot be "mathematically established", as Gödel has claimed. (Note that by using precisely the same argument, we can "demonstrate" other "disjunctions", by taking **K** to denote, e.g., "the system of all mechanically demonstrable mathematical propositions" or "the system of all mathematical propositions which can be proved in some sound formal system" or "the system of all mathematical propositions which can be proved in some sound and justified formal system", etc. All these disjunctions will be no less "valid" than the original one of Gödel.)

In the next sections it will be explained why Gödel's notion of "the system of all humanly demonstrable mathematical propositions" is ill-defined, so even the disjunction of [1-Gödel-B] and [2-Gödel] is extremely vague. We also show that even if we accept this particular "Gödel's disjunction", the other, more interesting formulations of that disjunction do not follow.

4. Mind(s)

From a philosophical point of view, the most interesting Gödel's disjunctions are those that refer to "the human mind". Thus these disjunctions might be relevant to classical problems like the mind-body problem, and the problem of free will (Lucas, 1961). However, it has already been pointed out by several authors that the use of this notion in the disjunctions is rather problematic: "It is certainly not obvious what it means to say that the human 'mind', or even the 'mind' of some human being, is a finite machine, e.g., a Turing machine" (Boolos, 1995, p. 293). "Hardly any mathematicians would ascribe mathematical clarity to the concept of 'the human mind'" (Feferman, 2006, p. 141). "Gödel's generic talk of 'the human mind' in his Gibbs talk is dangerously misleading" (Williamson, 2016, p. 249).

Because of this fuzzy notion that is used in many of the Gödel's disjunctions, their "mathematical proofs" (including Gödel's original one) rely on some crucial hidden assumptions. In what follows we reveal those assumptions, and show that it is extremely unclear what is meant by "human mind" (and by some related notions that appear in versions of GD and their "proofs").

4.1. "Turing Machines" and "Church Thesis"

First of all, the meaning of the word "mind" here is doubtful. It is clear that in the context in which this noun is used here, it is assumed that it denotes some object (unlike, e.g. when one uses in sentences nouns like "luck" or "fate"). But what is that object? The mechanist claims that there are really no objects that may be called "human minds"—there are only human brains. Hence the related disjunctions are meaningless, and so certainly cannot be "proved". The obvious (and justified) reply to this first objection is, of course, that the main point of the first disjunct is that just the activity of our brains cannot account for our mathematical capabilities, and so we should have something else, and this something else is what is called here "mind". But except for [1-Gödel-A], none of the other formulations above of the first disjunct even mentions the word "brain". Neither is "brain" mentioned in the proof that Gödel provided to his disjunction. Indeed, we have seen that the most this proof might show is the disjunction of [1-Gödel-B] and [2-Gödel]. Gödel then derives [1-Gödel-A] from [1-Gödel-B] as follows. First, by Church Thesis (CT), [1-Gödel-B] is equivalent to the claim that the set of humanly demonstrable theorems cannot be produced by any Turing machine. Then another application of CT yields that the set of humanly demonstrable theorems cannot be produced by any finite machine. Since the human brain is obviously a finite machine, 1-Gödel-A follows. However, these two applications of "Church Thesis" are in fact applications of two different theses. The first application relies on the mathematical thesis that a function $f\colon \mathcal{N} \rightarrow \mathcal{N}$ is computable by some uniform discrete algorithm iff it is recursive (or, according to a provably equivalent version, is computable by some particular "Turing machine"). The second application above of "Church Thesis" takes it to be claiming that if the values taken by some function $f\colon \mathcal{N} \rightarrow \mathcal{N}$ (for example: the characteristic function of the set of true arithmetic sentences) can all be somehow computed in one way or another by some machine (e.g., a human brain), then f is recursive (or computable by some particular "Turing machine"). Since "a machine" in general is not, and never has been, a mathematical notion, this is a much stronger, nonmathematical thesis. (In other words: despite the confusion that the use of a natural language causes here, "mechanically computable" and "computable by a machine" mean quite different things.) Unlike the mathematical (and original) version of CT, the stronger one is not supported by the evidence for CT that can be found in the literature, and a "proof" of GD that uses it is circular. Hence even if we accept the mathematical CT as an axiom, and in addition accept Gödel's proof of the disjunction of [1-Gödel-B] and [2-Gödel], we still cannot see the disjunction of [1-Gödel-A] and [2-Gödel] as a "mathematically established fact".

The question about the meaning and scope of CT seems to stand also behind the different views of Lucas and Penrose concerning what exactly their "Gödel argument" is showing. While Lucas (and Gödel) took it as refuting mechanism, that is: the thesis that the activity of the "human mind" can be reduced to the activity of the human brain and the laws of Physics, Penrose explicitly does not agree. He claims to refute only computationalism, that is: the thesis that the activity of the human "mind" can be reduced to computations. This very significant difference is reflected in the difference between [1-Lucas] and [1-Penrose]. Anyway, the questions what is exactly Church Thesis, and what version of it we are justified to accept, are complicated. Therefore we shall not enter deeper into them here. It will be done in a different paper. Accordingly, for the sake of argument we shall accept in what follows the identification of "finite machine" with "Turing machine".

Next we notice that even the use of the notion of a “Turing machine” is very ambiguous in the literature on GD and LP. When it is said that the “mind” is not a Turing machine, it is not always clear whether what is meant by the latter is a combination of hardware and software, that is: the idealized Turing’s device together with a specific program (i.e. a finite set of quadruples of a certain type), or just the hardware, i.e. the idealized device needed for running Turing-type programs on some input.[4] At first sight, the second interpretation seems more reasonable, since when we perceive a computer as a “machine”, we think about it as a device that can execute many programs, i.e. can simulate the activity of many Turing machines (even all, in case we are talking about an idealized computer). However, for reasons that are not fully clear to me, it seems that it is the first interpretation that most of the various authors have in mind in all of the formulations above. This is explicit, e.g., in both [1-Horsten-Welch-A] and [1-Horsten-Welch-B].

4.2. “The Human Mind”

A particularly problematic aspect of the formulations of the disjuncts that refer to “the mind” is the use of the definite article in the repeated talks on “the human mind”, and the frequent back-and-forth moves from “the human mind” to “a human mind” in the discussion of the theses. Koellner’s formulations above of the first disjunct provide a good example. In these formulations Koellner has tried (with certain amount of success) to provide a less vague versions of GD. However, there is from the start an obvious ambiguity in his formulation: sometimes he uses [1-Koellner-A], which is about the outputs of the human “mind”, and sometimes [1-Koellner-B], which is about the outputs of a human “mind”. It is remarkable that he has never used the formulation: “The mathematical outputs of the idealized human ‘mind’ cannot coincide with the mathematical outputs of the idealized finite machine”. This again shows how much prejudice and hidden assumptions are contained just in the formulations of LP and GD, to say nothing about their “proofs”. A similar phenomenon is encountered in most other papers on the subject. But are [1-Koellner-A] and [1-Koellner-B] (for example) really equivalent? There is just one case in which the answer to this question is positive: if we assume that (the mathematical thought of) all (idealized) human “minds” are essentially the same. (This seems to be the view of Penrose. See below.) In the words of Williamson: “Talk of ‘the human mind’ may work better within a conception on which all normal humans have the same intellectual competence, all differences coming from accidental limitations on performance” (Williamson, 2016, p. 250).

[4] Limiting the discussion to universal Turing machines does not eliminate the ambiguity: Instead of talking on combinations of a device and a program that wait for an input in order to run, in the case of universal Turing machines we talk on a combination of a device and a fixed part of the input, that wait for another part of the input in order to run.

This is of course an assumption that cannot be established mathematically, so using it (as Gödel might implicitly have done—he did not explain this point) already refutes the claim of "mathematically establishing" Gödel disjunction. But what reason do we have even to believe it? It is certainly false for actual human "minds". Most people on earth do not even understand Gödel's theorem and its proof, let alone would ever be able to discover and prove it themselves. I guess this is why participants in the discussions of the subject, including Penrose himself, rely on the activity (either actual or potential) of mathematicians. (By this they seem to leave open the possibility that the "minds" of people who cannot be worthy mathematicians are Turing machines…) Thus in Chapter 10 of (1989) Penrose argues:

> A mathematical argument that convinces one mathematician—providing that it contains no error—will also convince another, as soon as the argument has been fully grasped. […] Thus we are not talking about various obscure algorithms that might happen to be running around in different particular mathematicians' heads. We are talking about one universally employed formal system which is equivalent to all the different mathematicians' algorithms for judging mathematical truth. (pp. 539–540)

Even had this observation about mathematicians been true, this fact would have been no more than an empirical fact, not a mathematical one. But actually what Penrose says here is simply false. There have been, and there still are, many disagreements among mathematicians about validity of proofs. Here are few examples. Many more can be given.

- The debates on GD and LP provide good examples themselves. While Gödel believed that GD is a "mathematically established fact", Feferman (for example) did not accept his proof (Feferman, 2006). Similarly, while almost every mathematical logician rejects the proofs that Lucas and Penrose have given to their theses, Lucas and Penrose insist that they are ("essentially") correct. Obviously, the "minds" of Lucas and Penrose differ from those of the majority of the logicians…
- Gödel was a devoted platonist that saw no problem in using actual infinity in proofs (something that according to his own testimony has allowed him to prove his theorems). In contrast, the only infinity that was acceptable to Euclid was potential infinity. Indeed, in most of the history of mathematics, from the Greeks to Gauss, the use of actual infinity in proofs was rejected by almost all the mathematicians. Only in recent times its use is viewed as legitimate by the majority of them—and there are several respectable mathematicians who still reject it. Therefore I see no reason to think that the ("idealized" versions of the) "minds" of Gödel and Euclid (say) were identical.

- There is also a great disagreement between constructivists on one hand, and classical mathematicians on the other. As is well known, constructivists reject the general use of the law of excluded middle, while classical mathematicians use it freely. There are also many disagreements among the followers of various brands of constructivism: Intuitionism, Bishop's constructivism, Russian constructivism (in the tradition of Markov and others), and so on.
- Even among classical mathematicians who are not finitists or constructivists, there is a controversy about the acceptance of certain axioms. Thus there are mathematicians who believe that they can "see" that measurable cardinals exist (or at least that their existence is consistent with **ZFC**), while many other mathematicians (like me) totally lack this ability. Even Penrose himself admits in Chapter 4 of (1989) that

> When all the ramifications of set theory are considered, one comes across sets which are so wildly enormous and nebulously constructed, that even a fairly determined Platonist such as myself may begin to have doubts that their existence, or otherwise, is indeed an "absolute" matter. There may come a stage at which the sets have such convoluted and conceptually dubious definitions that the question of the truth or falsity on mathematical statements concerning them may begin to take on a somewhat "matter-of opinion" quality rather than a "god-given" one. (p. 147)

For fairness, I should note that Penrose did not completely ignore the difficulties to his thesis (about the "universal mathematician") that are caused by the different views that actual mathematicians have about mathematical truth and validity of proofs. In a footnote to Chapter 10 of (1989) he says:

> Some readers may be troubled by the fact that there are indeed different points of view among mathematicians. Recall the discussion given in Chapter 4. However the differences, where they exist, need not greatly concern us here. They refer only to esoteric questions concerning very large sets, whereas we can restrict our attention to propositions in arithmetic (with a finite number of existential and universal quantifiers) and the foregoing discussion will apply. (Perhaps this overstates the case somewhat, since a reflection principle referring to infinite sets can sometimes be used to derive propositions in arithmetic.) As to the very dogmatic Godel-immune formalist who claims not even to recognize that there is such a thing as mathematical truth, I shall simply ignore him, since he apparently does not possess the truth-divining quality that the discussion is all about! (pp. 581-582)

Here, as a side remark inside brackets within a footnote, Penrose is burying the point that decisively refutes what he is claiming. His case is not just "overstated" because of the fact noted in the brackets. That fact demolishes his case completely, because "the propositions in arithmetic that axioms of strong infinity are used for their proofs" are exactly of the type that Lucas and Penrose use in their arguments. Thus assume that Penrose has doubts about the strong infinity axiom *I*, While *W* is a mathematician who "sees" or somehow feels s/he knows that *I* is true. Then *W* also knows the truth of the Π_1^0-arithmetic proposition that states that **ZFC+I** is consistent—something that there seems to be no way for Penrose to know. So, according to Penrose's own argument, the "mind" of *W* "surpasses the power" of Penrose to prove Π_1^0-arithmetic propositions, and in particular—the "minds" of Penrose and *W* are different in an essential way.

Note 1 Gödel too did not ignore the problems that are caused to his disjunction by the the existence of different schools of mathematics. Therefore he did his best to make his argument for GD independent of a mathematician's philosophy of mathematics: "It is of great importance that at least this fact [i.e. that the disjunction is 'an established mathematical fact'] is entirely independent of the special standpoint taken toward the foundations of mathematics" (Godel, 1951, p. 310).

However, what is in question here is whether the formulation of GD is meaningful. Hence Gödel's care for the independence of his argument from philosophical views is irrelevant to the point we are making.

The upshot of this discussion is that [1-Koellner-A] and [1-Koellner-B] are not equivalent. What is more, it casts strong doubt on the meaning of the former. The only possibility that remains to try to give some meaning to it and to all the other formulations above that mention "the human mind", is to understand "the mathematical outputs of the (idealized) human mind" as referring to the totality (that is: the union) of the true mathematical outputs of the (idealized) human "minds".[5] This interpretation is examined in the next Section. Meanwhile we turn to a further examination of [1-Koellner-B].

4.3. "The Mind" of a Particular Mathematician

Let us turn to versions of GD that do not pretend to describe properties of the mythic "Human mind", but instead claim that some given specific "mind" "is not a machine". As is stated in [1-Koellner-B], and confirmed by Gödel and Penrose themselves, these versions do not really speak of the actual "mind" of someone like Gödel (say), but on the "mind" of an idealized Gödel, who lives for ever, and has other nice non-human qualities, but still is exactly like the real Gödel with respect to his mathematical abilities. Similarly, GD is not about any real finite machine, but about an idealized one. These facts, especially the first one, have been severely criticized in a very convincing way in (Feferman, 2006; Koellner, 2018b; Putnam, 2011), and especially in (Shapiro, 1998) and (Shapiro,

[5] As noted in (Feferman, 2006), an indication that this was not what Gödel himself had in mind is provided by what he said in a conversation with Hao Wang reported in p. 189 of (1996): "By mind I mean an individual mind of unlimited life span. This is still different from the collective mind of the species".

2016). I am not going to repeat the arguments given in these papers here. Instead, I want to emphasize the following points (several of them new, as far as I know):

- The mechanist and the computationalist theses are not about idealized human beings and idealized machines, but about real human beings and real machines. I have never seen any explanation (by either Gödel, Penrose, Lucas, or anybody else) how a claim like [1-Koellner-B] implies a claim like [1-Gödel-A], in case what is meant in the latter by "the human mind" is (say) "the mind of the real Gödel".
- It seems to me almost certain, and certainly possible, that an essential part of the permanent code that is built into any human machine ensures its mortality. Therefore the concept of an immortal human "mind" might well be an oxymoron!
- The idealization of "a human mind" that is involved in the picture that Gödel had of this notion, goes far beyond imagining it to be able to work for ever. It is actually based on a very naive view of a "mind", that for the task of doing mathematics is self-contained, and in principle independent of getting external output. I see no reason to believe in this romantic picture. Thus no matter how genius Archimedes has been, his abilities were limited by the culture in which he was active. Because of this culture, he was unable even to introduce the number zero. As for Gödel's theorems—they were not a part of the mathematics which was accessible to him. In fact, it seems to me very likely that even had Archimedes been immortal, as long as he would have worked in complete isolation from other mathematicians, he might have never discovered Gödel's theorems.
- Let us go one step further. We maintain that not only talks about "the human mind" in general, but also talks about the "mind" of a particular person like Gödel, are misleading. Is GD intended to tell us something about the "mind" of Gödel when he was four years old? Or even about his "mind" when he was 70 years old? Certainly not. The reason is that a person's "mind" is something dynamic. There is no single "mind of Gödel". There is at most "the mind of Gödel at a certain time of his life". The "mind" of any particular living person changes all the time by its interaction with the world and by learning new things (and forgetting others—this is also an essential component of the development of any actual "mind"). This, e.g. is the reason why it frequently happens that a problem one could not solve at one point of her life, she finds a solution to a few years later.

Note 2 A particularly interesting implication of the dynamic nature of a human "mind" is given by the following scenario: suppose a certain person who understands Gödel's incompleteness theorems and their proofs, e.g. Lucas, somehow learns at a certain time t_2 of his life that the set of true arithmetic propositions he could potentially have known at some previous time t_1, is identical to the set of theorems of the formal system $\mathcal{T}$. (This could happen if he is told so by "his creator"—a term used by Gödel in [1951]—or if he infers this with very high degree of certainty from new experimental data that he had meanwhile acquired.) This fact was not (and could not have been) a part of his knowledge at time t_1. Hence the "mind" of Lucas at time t_2 is different from his "mind" at time t_1. This fact makes it possible for him to know at time t_2 various Gödel's sentences for $\mathcal{T}$ that were not (and could not have been) known to him at time t_1.

Note 3 Interestingly, on another occasion Gödel himself noted the dynamic nature of a human "mind". In a note, which was prepared for publication but never actually published, he wrote:

> Turing gives an argument which is supposed to show that mental procedures cannot go beyond mechanical procedures. However, this argument is inconclusive. What Turing disregards completely is the fact that mind, in its use, is not static, but constantly developing. (Gödel, 1990, p. 306)

I wonder why Gödel has not noticed the crucial importance of this correct observation to his own disjunction. (Or maybe he did? After all, [Gödel, 1951] has never been published by Gödel himself.)

It follows from the discussion at the last item above that even in [1-Koellner-B] the first disjunct is very vague, and should be reformulated, e.g., as "The (realistic) potential mathematical outputs of a given person at a given point of time cannot coincide with the (realistic) potential mathematical outputs of any finite machine (at some point of time)". In my opinion, this formulation of the first disjunct is probably false. What is sure is that Gödel theorems have little to tell us about its truth value.

In connection with this, it should be noted that it seems that almost all the participants, from both sides, in the debates about GD and LP have followed Gödel and Lucas in ignoring the dynamic nature of human "minds", and so have discussed only the question whether it can be equivalent to some static Turing machine. The question should have been whether it can be equivalent to a robot whose "mind" (i.e. the combination of its hardware, software, and memory) continuously changed through learning (both from the experience it gets from its interaction with the neighborhood, and from direct teachers) and forgetting. Such robots already exist, and I do not see any "Gödel argument" that can prevent us from making in the future a robot that has even the same mathematical abilities that Gödel had when he was at his twenties. I suspect that the importance for the debate of the power of learning, and of the dynamic aspects of both "minds" and machines, was disregarded because of the continuing confusion noted above about what is meant by a "machine": Is it just the device (i.e. hardware), or is it something bigger, like the device together with (a part of) the software and memory?

5. “Knowable”, “Demonstrable”, “Certain”, “Evident”

In this section we examine the alternative interpretation (which was mentioned at the end of Section 4.2), of “the mathematical outputs of the (idealized) human mind” as referring to all the true mathematical facts that may be output by (idealized) human “minds”. This interpretation is explicitly reflected (with important amendments that Shipman has found necessary) only in [1-Shipman] and [2-Shipman]. However, it seems to stand also behind most (if not all) the formulations above that avoid the use of the notion of “human mind”, replacing it instead with some less ontologically committed notions, like: “human understanding”, “human mathematical reasoning”, “the collection of humanly knowable theorems”, and “all demonstrable mathematical propositions”. As was forcefully argued in (LaForte, Hayes, & Ford, 1998), it should be clear that in this form, GD and LP have no real relevance to the mechanist (or even the computationalist) thesis, because the claim that (“knowable”) mathematics is r.e. (i.e. is encapsulated by some formal system) is completely different from the claim that the (“knowable”) mathematics of any specific mathematician is r.e. Nevertheless, the corresponding theses still have interest and philosophical implications of their own. So let us examine them.

5.1. “Knowable” Versus “Demonstrable”

The notions of “human understanding”, and “human mathematical reasoning” are too broad and fuzzy to be used in a logico-mathematical discussion. So let us concentrate on the two collections of mathematical objects that are mentioned in the previous paragraph. To make it more plausible that they describe definite mathematical objects themselves, we shall restrict ourselves to two less general (but sufficiently rich) sub-collections: “the collection of humanly knowable arithmetic propositions” and “the collection of humanly demonstrable arithmetic propositions”.[6] Assuming, for the time being, that these two collections are well-defined, let us discuss first the question whether they are identical. The obvious answer should be that they are not. Here are two examples:

- Even children know that multiplication of natural numbers is commutative. In contrast, even the majority of the scientists do not know how to demonstrate this mathematically. Their knowledge of it is based on a mixture of personal experience with what is taught in school.
- A more subtle example is given by complexity theory. For all practical purposes, the computer scientists behave as if they know that $P \neq NP$. In fact, most of them feel that they indeed know this, even though none of them can mathematically demonstrate it.

[6] We may further restrict them by replacing “arithmetic” with “Π^0_1-arithmetic”.

The obvious reply to this objection that one can implicitly find in the literature on the subject is that what is meant here by both “knowable” and “demonstrable” is “knowable with mathematical certainty” (Godel, 1951) or “logically derivable from evident axioms” (Godel, 1951, again), or “perceivable by mathematicians as unassailably true” (Penrose, 1994), or “demonstrably true by human reason and insight” (Penrose, 2011), or “knowable with unassailable mathematical certainty, via full mathematical rigor” (Shapiro, 2016). The use here of several different formulations (and several others can be found in the literature), employing different words which have similar but not identical meanings, is already suspicious. True, when we need to express ourselves precisely, it is often helpful to have in our language different words whose meaning is close but not identical. However, this fact also makes it possible to obscure things by switching from one word to another. This is indeed what repeatedly happens in the papers on the subject, especially in papers that try to support LP. However, here I would like to give an example from an argument of an opponent: Stewart Shapiro. Usually, Shapiro is very careful in distinguishing between different concepts, and he uses this repeatedly and convincingly in order to show that there is no sufficiently precise mechanistic thesis that is undermined by Gödel’s theorems (Shapiro, 1998; 2016). However, when he discusses the candidacy of **ZFC** as a formal system that encapsulates all “unassailably true arithmetic propositions” he is less careful. He writes: “Moreover, is Zermelo-Fraenkel set theory sufficient for all unassailable mathematical knowledge? If so, the mechanist wins. But **ZFC** clearly isn’t sufficient. Don’t forget the Gödel sentence for **ZFC**. I presume we do know that” (Shapiro, 2016, p. 198).

Notice that Shapiro does not write that he is presuming that the Gödel sentence for **ZFC** belongs to our “unassailable mathematical knowledge”—he is careful to presume only that we know it. By this he is taking advantage of the crucial difference between “knowing” and “mathematically demonstrating” noted above. Thus I, for one, feel that I know with very high degree of confidence (which is as least as high as my knowledge that all men are mortal, or that the sun will rise tomorrow), that **ZFC** is consistent. The reason is simple: I am convinced that had it been inconsistent then this would have been discovered by now (more than a century after the best mathematicians in the world start to extensively investigate and use it).[7] Moreover: even though I am not a platonist, I admit that the picture of the “Von Neumann universe” provides strong intuitive support to the belief in the consistency of **ZFC**, even though this support is not absolutely conclusive. Still, I definitely cannot demonstrate, or claim to know with “absolute mathematical certainty”, that **ZFC** is consistent.[8]

[7] Gödel himself notes in (1951) the possibility of *empirical* certainty that the brain works like a computer, or that the mathematical human “mind” is equivalent to a Turing machine.

[8] Actually, Shapiro himself observed in (1998) that given a system S, “for each axiom ψ of S, we can have good reason to think that ψ is true without having good reason to

5.2. Degrees of Certainty

The discussion above shows that it is anything but clear what exactly is claimed in each of the above vague formulations of the first disjunction in case it is not (or may not be) about the "mind" of a single person, or whether they all say the same thing. In order to give some chance for a Gödel's disjunction to mean something which is not just a trivial reformulation of Tarski's theorem, and may follow from Gödel incompleteness theorems, we shall henceforth assume that all of these formulations indeed try to make the same claim: that the set of Π^0_1-arithmetic propositions which are "provable with unassailable mathematical certainty" differs from the set of Π^0_1-arithmetic theorems of any formal system. Does at least this formulation express a unique meaningful claim? Not really. The reason is that the notion of "unassailable mathematical certainty" does not have a determined unique meaning. The main problem with it was formulated in (Koellner, 2018b, p. 473) as follows: "justification and evidence in mathematics come in degrees". In other words: there are different levels of mathematical certainty. They are mainly characterized by the role that infinity is allowed to have in proofs. Here are the most important groups of levels. (The reason why we speak here about groups of levels is explained in the sequel.)

Finitistic mathematics. Here references to infinite objects and quantification over an infinite collection of objects are strictly forbidden in propositions and proofs. According to Hilbert, only the use of finitistic methods of proof provides absolute mathematical certainty. However, this position is shared now by very few mathematicians. Still, it should be noted that in (Ye, 2011) it is shown that Finitistic mathematics is quite rich and its power is far bigger than what one might have expected.

Predicative mathematics (Feferman, 2005)**.** Here potentially infinite objects are allowed. As noted above, this was the way infinity was viewed by most of the mathematicians throughout almost the whole history of mathematics; the change came only at the second half of the 19th century. The modern predicativist program was initiated by Poincaré (1906; 1909), in his follow up on (Richard, 1905). Its viability was demonstrated by Hermann Weyl, who seriously developed it for the first time in his famous small book *Das Kontinuum* (1918; 1987). After Weyl, the predicativist program was extensively pursued by Feferman, who in a series of papers (see, e.g., 1964; 1998; 2005) developed proof systems for predicative mathematics. Weyl and Feferman have shown that a very large part of classical analysis can be developed within their systems.

Feferman further argued that predicative mathematics in fact suffices for developing all the mathematics that is actually indispensable to present-day natural

think that S is consistent". Now take S to be **ZFC**, where by "good reason" we understand provable with unassailable mathematical certainty...

sciences. Allow me to add to that my personal opinion (Avron, 2020): I believe that predicative mathematics is exactly the part of mathematics that deserves being called "absolutely certain".

For the predicativist program, the following well-known fact about Π^0_1-sentences is very important: if ψ is such a sentence, and $T \vdash \psi$ (where T is some formal theory), then **PA** + $Con_{\mathrm{T}} \vdash \psi$, where **PA** is first-order Peano's Arithmetics. Since **PA** is a part of predicative mathematics, it follows that no matter how strong and large a formal theory T is, and to what extent it goes beyond predicatively acceptable mathematics, as far as Π^0_1-sentences are concerned, the use of T in proofs is equivalent to the use in predicative mathematics of the single arithmetic sentence that expresses the fact that T is consistent. In other words: the degree of certainty, that a proof of a Π^0_1-sentence ψ in a given formal theory T gives us about the truth of ψ, is identical to the degree of certainty that we have in the consistency of T.

ZF(C). ZFC is the canonical system in which almost all of mathematics is officially developed. What is more: it is safe to say that the axioms of **ZF** include all the axioms of set theory that the great majority of the mathematicians in the world are ready to accept as uncontroversial (although there might be different opinions about what it means to say that they are "true"). It seems that nowadays most mathematicians think that the axiom of choice is true too. However, historically many great mathematicians have strongly objected to the use of that axiom. The fact that this situation has been changed might reflect cultural environment—hardly what justifies seeing something as "obviously true". Luckily, since the consistency of **ZFC** follows in **PA** from the consistency of **ZF**, **ZFC** is as good as **ZF** for justifying the acceptance of the truth of Π^0_1-sentences. Things are different with respect to other axioms of **ZF** that some mathematicians find dubious, like replacement or powerset. In any case, it seems to me that only few mathematicians would deny that proofs in **PA** of Π^0_1-sentences provide higher degree of certainty than proofs in **ZFC**.

Extensions of ZFC. Many set theorists feel that there is no reason to stop at **ZFC**, especially since the latter cannot prove its own consistency (which should be taken for granted by anybody who uses **ZFC** for showing the truth of some Π^0_1-sentence). The natural direction of going beyond **ZFC** is to add to it stronger and stronger axioms of strong infinity. Thus in (1946) Gödel proposed provability with regard to extensions of **ZFC** with true large cardinal axioms as a plausible concept of absolute demonstrability. Similarly, in (2005), Franzén wrote that **ZFC**+some infinity axiom may represent exactly the "human demonstrated mathematics". Unfortunately, "The case for the axioms gets harder and more involved as one ascends to higher and higher reaches". (Koellner, 2018b, p. 473). (Recall what Penrose himself has said about this in [1989, Section 4.2].) The situation with respect to the "absolute certainty" of large cardinal axioms was best described by Feferman as follows:

> I don't know of anyone who says that we can be assured that all the large-cardinal axioms that have been considered to date lead only to mathematical truths, let alone that they are "evident" as required by Gödel in his disjunctive formulation. (2006, p. 149)

This state of affairs is obviously the reason why Shipman has turned to acceptance of set-theoretical statements not on the basis of their being evident, or "knowable with unassailable mathematical certainty", but on the basis of future consensus. To see how vague is his notion of "human mathematics" it is enough to follow him word by word and define "machine mathematics" as the collection of formalized sentences in the language of set theory which are logical consequences of statements that will eventually come to be accepted by a consensus of machine mathematicians as "true". What can we infer from Gödel theorems about this "machine mathematics"? Actually, there might be reasons to believe that it includes all true arithmetical sentences: Call any machine which produces arithmetical sentences "a machine mathematician" iff all the arithmetical sentences it produces are true. Let an arithmetical sentence be "accepted by a consensus of machine mathematicians" once 1000 machine mathematicians have produced it. Then obviously all true arithmetical sentences belong to "machine mathematics" according to these definitions. Shipman might object, of course, that these are not good definitions or characterizations of "mathematicians" or "consensus". I would agree, but I cannot see what better ones he might be able to offer.

Another aspect of Shipman's definition is its dependence on time ("eventually"). Similarly, on many occasions H. Friedman has expressed his belief that the use of strong cardinal axioms will necessarily become a part of humane mathematics. So he too is speaking about the future. Why? Because nobody can claim that such axioms are "a part of humane mathematics" at present. It seems therefore that what the "human mind" can prove with "unassailable mathematical certainty" depends on time, consensus, etc. How can such a concept be connected with Gödel's theorems?

Note 4 As was noted already in Note 1, Gödel was aware of the difficulties that are caused to his disjunctive thesis by the existence of different views about what is evident and what is not. Therefore he explicitly tried to make his argument for his thesis independent of one's views on the matter. In other words, he claimed that his argument should be acceptable not only to platonists, but also to finitists, constructivists, predicativists, etc. The difference, he wrote, between the various schools would be with respect to the truth-values of the two disjuncts; not with respect to the truth-value of their disjunction. However, Gödel missed the real problems here. First, it might be that because they all use the same vague, informal language, they all would accept a certain formulation of the disjunction—but each one would understand by this a completely different thesis. Since each group above includes many variants and non-identical theses, the number of theses here would be almost the same as the number of people who are interested in the subject. Second, as we have emphasized in Note 1, no matter what school one is associated with, in most cases the main words involved in the formulations of the disjunction would be extremely vague. (And again, the disjunction is trivial and totally uninteresting in the few cases in which its formulation can be taken as meaningful.)

5.3. On Geometric Reasoning

The discussion so far concentrated on the degree of certainty that can be achieved using formal reasoning about abstract notions like numbers and sets. What about geometric reasoning? Until the 19th century, it had a central part in mathematical reasoning (and for long periods—it was its main rigorous part). The invention/discovery of non-Euclidean geometries has changed this situation. Nowadays geometric reasoning is still taken to be useful for getting intuitive understanding of theorems in analysis, and for providing hints how they may be rigorously proved. However, direct use of them in proofs of arithmetical propositions is usually considered to be illegitimate. This approach may be questioned. It might be argued that geometric arguments do provide some degree of certainty. Thus Penrose gave in (1994) the (Euclidean) geometric proof that $a \times b = b \times a$ as an elementary example of geometrical reasoning, and said that it is "a perfectly good proof, though not a formal one" of a general property of natural numbers. However, on another occasion he described Euclidean geometry as inaccurate:

> The most ancient of the SUPERB theories is the Euclidean geometry that we learn something of at school. The ancients may not have regarded it as a physical theory at all, but that is indeed what it was: a sublime and superbly accurate theory of physical space—and of the geometry of rigid bodies. Why do I refer to Euclidean geometry as a physical theory rather than a branch of mathematics? Ironically, one of the clearest reasons for taking that view is that we now know that Euclidean geometry is not entirely accurate as a description of the physical space that we actually inhabit! (Penrose, 1989, p. 197)

The reason that Euclidean geometry is described by Penrose as "inaccurate" (Popper would have simply said "false") is that according to Einstein's general relativity theory, the real geometry of our universe is actually a non-Euclidean one. Nevertheless, when he is talking about applying geometrical reasoning in demonstrating properties of the natural numbers, Penrose has only Euclidean geometry in mind:[9]

> The study of non-Euclidean geometries is something mathematically interesting, with important applications […] but when the term "geometry" is used in ordinary

[9] Also in Chapter 3 of (1989), where Penrose describes with fascination the amazing geometric properties of Mandelbrot set, saying then (p. 125) that "Like Mount Everest, the Mandelbrot set is just *there*!", the set he is talking about exists in the *Euclidean plane*. So if it has a platonic existence, then necessarily so does the Euclidean plane itself.

> language (as distinct from when a mathematician or theoretical physicist might use that term), we do indeed mean the ordinary geometry of Euclid. (1994, p. 111)

These incoherent views on the role of geometry in mathematics, all of them in the "mind" of just one, a particularly brilliant mathematician, shows how uncertain is what the degree of certainty that the use of geometrical reasoning provides is. It also gives further strong evidence that there are several different levels of "mathematical certainty".

6. Some Remarks on Lucas-Penrose's Theses

What we did above is to question the meaningfulness of the various formulations of the Gödel's disjunction in general, and of the various Lucas-Penrose theses in particular. For completeness, in this section we assume, for the sake of argument, that at least one of the latter makes sense, and briefly describe the two main mistakes (that is: unjustified hidden assumptions) that were noted in the literature in its alleged "proof".

1. The assumption that the (or a) "human mind" is consistent.
2. The assumption that in any case that we realize that the (or a) "human mind" is equivalent to a Turing machine, we should know this with mathematical certainty.

Unlike what is sometimes argued (partially even in [Krajewski, 2020]), there is no conflict between those that have emphasized the first assumption, and those that have emphasized the second one. Actually, there are good reasons to seriously take into account the possibility that our "mathematical mind" is based on a theory which is inconsistent, and we do not know this fact!

Let us start with some reasons that were given in the literature to doubt the truth (to say nothing about the certainty) of the first assumption, that is: the consistency of the mathematical "human mind":

Putnam: An actual mathematician makes mistakes, and her outputs contains inconsistencies (Putnam, 2011).

Davis: Great logicians (Frege, Curry, Church, Quine, Rosser) have managed to propose quite serious systems of logic which later have turned out to be inconsistent. "Insight" didn't help (Davis, 1990).

Franzén: **ZFC**+some infinity axiom may represent exactly the "human demonstrated mathematics", and we do not know whether that system is consistent (Franzén, 2005).

Penrose's reply to the first (Putnam's) argument is:

> The most usual kind of mistake that a mathematician might make is of no real concern to us here, being something that is correctable by that mathematician on further contemplation or when the error is pointed out by someone else. (2011, p. 351)

It is debatable whether this is indeed a satisfactory reply to Putnam. In any case, it is certainly irrelevant to Franzén's argument, and actually to Davis' one too. The inconsistencies in the systems suggested by the great logicians that Davis mentions were indeed pointed out to them by others, but it was not clear at all what their mistakes had been, and how to "correct" them. All of the principles they used seemed "certainly correct", and yet the whole system of each of them was inconsistent. It follows that there was something deeply inconsistent in their collections of beliefs, and it is not certain at all that this deep inconsistency disappeared after the obvious problems with their mistakes had been discovered. Therefore it is not inconceivable that some deep inconsistency exists in the mathematical "mind" of each of us.[10] In this connection, the following fact is rather telling: throughout the second half of the 19th century (if not already before), mathematicians were implicitly working within an inconsistent theory: naive set theory.[11]

Let us turn now to assumption 2 above. First, let us emphasize that it is indeed absolutely necessary for the argument of Lucas and Penrose to assume that our recognition of a certain formal system *F* as being equivalent to our "mind" (with respect to the (Π^0_1)-arithmetic sentences) should be mathematically certain. Otherwise, even under the assumption that we know with certainty the consistency of our mind, we would not be able to infer the consistency of *F*, or (equivalently) its Gödel's sentence, with any more mathematical certainty than *F* itself can. However, already Gödel admitted in (1951) that it is possible that the "mathematical human mind" is equivalent to a Turing machine which is unable to understand itself, and that to demonstrate that this is indeed the case (or at least that this is highly plausible), it suffices to bring forward a machine that empirically seems to be equivalent to our "mind". These observations of Gödel suffice to render the assumption of Lucas-Penrose under discussion as unwarranted. However, we would like to go one step further: to note that plausible candidates for *F* do exist. (This is a possibility that Lucas has obviously taken as just theoretical.) Actually, such candidates were already mentioned above. Thus according to Franzén and Shipman, *F* might be **ZFC** extended with some infinity axioms. But if we talk about the set of (Π^0_1)-arithmetic sentences that can be proved with certainty, then a much better candidate was already (partially) discussed in Section 5.1: it is **ZFC** itself.

[10] Note that that in Section 5.3 some incoherence, if not an inconsistency, is pointed out in the views of Penrose himself about the status of Euclidean geometry!

[11] Another interesting example is provided by the debate on the axiom of choice. Some of the great mathematicians that strongly objected to its use, like Borel and Lebesgue, did not notice that they had implicitly used it themselves in their work…

This explicit suggestion might immediately raise a particular case of the following standard objection:

> As long as we see mathematical theories, or algorithms, as fundamentally similar to what we know as mathematics, we tend to assume that all the theories that are encompassing our knowledge of the natural numbers must, in principle, be based on a series of transparent basic truths (axioms) and be developed due to the applications of known, correct logical rules. If so, every such theory, if presented to us, must be fully understood, or at least understandable. And this full understanding implies our knowledge of its consistency and presumably also soundness. Therefore, out-Gödeling is, indeed, possible. (Krajewski, 2020, p. 41)

Or in the words of Gödel himself, his second incompleteness theorem

> makes it impossible that someone should set up a certain well-defined system of axioms and rules and consistently make the following assertion about it: All of these axioms and rules I perceive (with mathematical certitude) to be correct, and moreover I believe that they contain all of mathematics. If someone makes such a statement, he contradicts himself. For if he perceives the axioms under consideration to be correct, he also perceives (with the same certainty) that they are consistent. Hence he has a mathematical insight not derivable from his axioms. (1951, p. 309)

It seems to follow that it makes no sense to fully trust the (Π^0_1)-arithmetic theorems of **ZFC**, but less than fully trust the consistency of **ZFC**. However, this conclusion is again based on a subtle confusion, the danger of which was again noted by Gödel himself. In a footnote to the last quote he observed about the person mentioned in it (the one who sets up a certain well-defined system of axioms and rules) that "If he only says 'I believe I shall be able to perceive one after the other to be true' he does not contradicts himself" (1951, p. 309).

What Gödel means here is that there is a difference between knowing with certainty the truth of each theorem of some system considered alone, (which means knowing with certainty an infinite numbers of claims), and between knowing the single claim that all of those sentences are true (a claim which is different from every such sentence). Thus we may be able to know with certainty any instance of Goldbach's conjecture, without ever knowing with certainty Goldbach's conjecture itself. Similarly, what I claim about **ZFC** is not that I sufficiently understand it to take its (Π^0_1)-arithmetic theorems as established with absolute certainty just because they are theorems of **ZFC**. I am only claiming the following:

- The fact that a certain arithmetics sentence ψ is a theorem of **ZFC** is a very good reason to believe its truth (for the reasons explained above, which are partially empirical). However, this theoremhood alone does not provide us absolute certainty in the truth of ψ.

- On empirical ground, I strongly believe that every (Π^0_1)-arithmetic sentence that will ever be proved with absolute certainty belongs to the set of theorems of **ZFC**.
- On empirical ground again, I see it as very plausible that the converse is true too: for every theorem ψ of **ZFC** there is some absolutely certain formal system F such that ψ is also a theorem of F. (F may e.g. be a system which we recognize as obtained from **PA** by the addition of some formalized reflection principles; see Feferman, 1962.)
- We do not know, and most probably we shall never know, the consistency of **ZFC** with absolute certainty.

I suspect that many people (including perhaps Gödel) would claim that although the situation I describe might in principle be possible, it is very unlikely to be the real one. I think that on the contrary, the facts as we know them at present support it. Nevertheless, I would like to end this section by pointing out an example in which a very similar state of affairs is accepted by most specialists to actually be the case. This is the case of predicative mathematics that was described above (and I personally take as identical to the "absolutely certain mathematics"). Without any connection to the debate on Lucas-Penrose theses, Feferman (1964) and Schütte (1965) independently characterized it by some (equivalent) formal systems that (so they claimed) prove exactly the arithmetic sentences that a real predicativist is able to prove with what s/he takes as absolute certainty. In the case of Feferman this was done in (1964) using a transfinite sequence of formal theories. Feferman maintained that a true predicativist can prove with certainty each theorem of each theory in this sequence, but he is not capable of seeing that he is able to do so, or the adequacy of the union of those systems as a whole. In other words: according to Feferman, he can exactly characterize what a predicativist (like me) can prove, although a real predicativist cannot do it (unless he abandons his principles). Feferman thinks therefore that he can know with full certainty a sentence which is equivalent to the consistency of my certain mathematics, while I myself cannot know it with certainty.[12] If he is right, then from Feferman's point of view (and almost every logician agrees) I (or at least my "mathematical mind") am equivalent to a Turing Machine. I do not feel insulted by this, but it is still difficult for me to accept that I am equivalent to a Turing Machine, while some other people (e.g. Lucas and Penrose) are not. Maybe this very human feeling is a sign that I am not exactly a Turing Machine after all…

[12] Although Feferman was very sympathetic with predicativism, and it is clear that it reflects his views better than any other known "ism", he has declared that he is not a real predicativist himself.

7. Conclusions

We have shown that the name "Lucas-Penrose thesis" encompasses several different theses. All these theses refer to extremely vague concepts, and so are either practically meaningless, or obviously false. The arguments for the various theses, in turn, are based on confusions with regard to the meaning(s) of these vague notions, and on unjustified hidden assumptions concerning them. All these observations are true also for all interesting versions of the much weaker (and by far more widely accepted) thesis known as "Gödel disjunction".

Now Penrose, e.g., has provided in (1994, and in other papers) "replies" to almost every argument made above. However, each of these "replies" is connected only to some of the theses he is trying to make (although he does not distinguish between them), and frequently they contradict each other. These and similar confusions, in turn, are frequently the result of the the inadequacy of natural languages for dealing with precise notions and propositions. My conclusion from this state of affairs is that an argument that cannot be fully formalized cannot be taken as a mathematical proof. What is more: if there is a debate about the soundness of an argument, then in order to resolve it one should first of all fully formalize it. One important outcome of such a full formalization is that it makes all the hidden assumptions explicit.

Another conclusion of this paper is the following dictum of Feferman: "It is hubris to think that by mathematics alone we can determine what the human mind can or cannot do in general" (2009, p. 213).

REFERENCES

Avron, A. (2020). Why Predicative Sets? In A. Blass, P. Cégielski, N. Dershowitz, M. Droste, B. Finkbeiner (Eds.), *Fields of Logic and Computation III, Eassys Dedicated to Yuri Gurevich on the Occasion of His 80th Birthday* (pp. 30–45). Springer.

Baaz, M., Papadimitriou, C. H., Putnam, H. W., Scott, D. S., & Harper, C. L. (Eds.). (2011). *Kurt Gödel and the Foundations of Mathematics: Horizons of Truth*. Cambridge: Cambridge University Press.

Boolos, G. (1995). Introductory Note to Kurt Gödel's "Some Basic Theorems on the Foundations of Mathematics and Their Implications". In S. Feferman et al. (Eds), *Collected Works, Volume III: Unpublished Essays and Lectures* (pp. 290–304). Oxford: Oxford University Press.

Charlesworth, A. (2016). A Theorem about Computationalism and "Absolute" Truth. *Minds and Machines*, *26*, 206–226.

Davis, M. (1990). Is Mathematical Insight Algorithmic? *Behavioral and Brain Sciences*, *13*, 659–660.

Ewald, W. (1996). *From Kant to Hilbert*. London: Clarendon Press.

Feferman, S. (1962). Transfinite Recursive Progressions of Axiomatic Theories. *Journal of Symbolic Logic*, *27*, 259–316.

Feferman, S. (1964). Systems of Predicative Analysis I. *Journal of Symbolic Logic*, *29*, 1–30.

Feferman, S. (1998). *In the Light of Logic*. Oxford: Oxford University Press.

Feferman, S. (2005). Predicativity. In S. Shapiro (Ed.), *The Oxford Handbook of the Philosophy of Mathematics and Logic* (pp. 590–624). Oxford: Oxford University Press.

Feferman, S. (2006). Are There Absolutely Unsolvable Problems? Gödel's Dichotomy. *Philosophia Mathematica*, *14*, 134–152.

Feferman, S. (2009). Gödel, Nagel, Minds, and Machines. *Journal of Philosophy*, *106*, 201–219.

Franzén, T. (2005). *Gödel's Theorem: An Incomplete Guide to Its Use and Abuse*. Wellesley: A.K. Peters.

Gödel, K. (1946). Remarks Before the Princeton Bicentennial Conference on Problems in Mathematics. In S. Feferman et al. (Eds.), *Collected Works, Volume II: Publications 1938–1974* (pp. 150–153). Oxford: Oxford University Press.

Gödel, K. (1951). Some Basic Theorems on the Foundations of Mathematics and their Implications. In S. Feferman et al. (Eds), *Collected Works, Volume III: Unpublished Essays and Lectures* (pp. 304–323). Oxford: Oxford University Press, 1951.

Gödel, K. (1990). *Collected Works, Volume II: Publications 1938–1974*. Oxford: Oxford University Press.

Gödel, K. (1995). *Collected Works, Volume III: Unpublished Essays and Lectures*. Oxford: Oxford University Press.

Koellner, P. (2016). Gödel's Disjunction. In L. Horsten & P. Welch (Eds.), *Gödel's Disjunction: The Scope and Limits of Mathematical Knowledge* (pp. 148–188). Oxford: Oxford university Press.

Koellner, P. (2018a). On the Question of Whether the Mind Can Be Mechanized, I: From Gödel to Penrose. *Journal of Philosophy*, *115*, 337–360.

Koellner, P. (2018b). On the Question of Whether the Mind Can Be Mechanized, II: Penrose's New Argument. *Journal of Philosophy*, *115*, 453–484.

Krajewski, S. (2020). On the Anti-Mechanist Arguments Based on Gödel Theorem. Studia Semiotyczne, 34(1), 9–56.

Horsten, L., & Welch, P. (Eds.). (2016a). *Gödel's Disjunction: The Scope and Limits of Mathematical Knowledge*. Oxford: Oxford university Press.

Horsten, L., & Welch, P. (2016b). Introduction. In L. Horsten & P. Welch (Eds.), *Gödel's Disjunction: The Scope and Limits of Mathematical Knowledge* (pp. 1–15). Oxford: Oxford University Press.

LaForte, G., Hayes, P. J., & Ford, K. M. (1998). Why Gödel's Theorem Cannot Refute Computationalism. *Artificial Intelligence*, *104*, 265–286.

Lucas, J. R. (1961). *Minds, Machines and Gödel. Philosophy*, *36*, 112–137.

Penrose, R. (1989). *The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics*. Oxford: Oxford University Press.

Penrose, R. (1994). *Shadows of the Mind: A Search for the Missing Science of Consciousness*. Oxford: Oxford University Press.

Penrose, R. (2011). Gödel, the Mind, and the Laws of Physics. In M. Baaz et al. (Eds.), *Kurt Gödel and the Foundations of Mathematics: Horizons of Truth* (pp. 339–358). Cambridge: Cambridge University Press.

Poincaré, H. (1906). Les Mathématiques et la Logique, II, III. *Revue de Métaphysique et Morale*, *14*, 17–34, 294–317.

Poincaré, H. (1909). La Logique de l'infini. *Revue de Métaphysique et Morale*, *17*, 461–482.

Richard, J. (1905). Les Principes des Mathematiques et les Problémes des Ensembles. *Revue general des sciences pures et appliqués*, *16*, 541–543.

Putnam, H. W. (2011). Gödel Theorem and Human Nature. In M. Baaz et al. (Eds.), *Kurt Gödel and the Foundations of Mathematics: Horizons of Truth* (pp. 325–338). Cambridge: Cambridge University Press.

Schütte, K. (1965). Predicative Well-Ordering. In J. Crossley and M. Dummett (Eds.), *Formal Systems and Recursive Functions* (pp. 279–302). Oxford: North-Holland.

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

Shapiro, S. (2016). Idealization, Mechanism, and Knowability. In L. Horsten & P. Welch (Eds.), *Gödel's Disjunction: The Scope and Limits of Mathematical Knowledge* (pp. 189–207). Oxford: Oxford university Press.

Wang, H. (1996). *A Logical Journey*. Cambridge: The MIT Press.

Weyl, H. (1918). *Das Kontinuum: Kritische Untersuchungen über die Grundlagen der Analysis*. Leipzig: Veit.

Weyl, H. (1987). *The Continuum: A Critical Examination of the Foundation of Analysis*. Kirksville, Missouri: Thomas Jefferson University Press.

Williamson, T. (2016). Absolute Provability and Safe Knowledge of Axioms. L. Horsten & P. Welch (Eds.), *Gödel's Disjunction: The Scope and Limits of Mathematical Knowledge* (pp. 243–253). Oxford: Oxford University Press.

Ye, F. (2011). *Strict Finitism and the Logic of Mathematical Applications*. New York: Springer.

STUDIA SEMIOTYCZNE (SEMIOTIC STUDIES), t. XXXIV, nr 1 (2020), s. 109–158
ISSN 0137-6608, e-ISSN 2544-073X
DOI: 10.26333/sts.xxxiv1.06
Article

JEFF BUECHNER [*]

USING KREISEL'S WAY OUT TO REFUTE LUCAS-PENROSE-PUTNAM ANTI-FUNCTIONALIST ARGUMENTS

SUMMARY: Georg Kreisel (1972) suggested various ways out of the Gödel incompleteness theorems. His remarks on ways out were somewhat parenthetical, and suggestive. He did not develop them in subsequent papers. One aim of this paper is not to develop those remarks, but to show how the basic idea that they express can be used to reason about the Lucas-Penrose-Putnam arguments that human minds are not (entirely) finitary computational machines. Another aim is to show how one of Putnam's two anti-functionalist arguments (that use the Gödel incompleteness theorems) avoids the logical error in the Lucas-Penrose arguments, extends those arguments, but succumbs to an absurdity. A third aim is to provide a categorization of the Lucas-Penrose-Putnam anti-functionalist arguments.

KEYWORDS: functionalism, Computational Liar, Gödel incompleteness theorems, finitary computational machine, mathematical certainty, finitary reasoning, epistemic refutation, metaphysical refutation, epistemic justification, recursively unsolvable, epistemic modality, finitary computational description.

1. Introduction

J. R. Lucas (1961) argued that for any finitary computational machine hypothesized to simulate full human mentality, there will be a Gödel sentence for that machine it cannot prove to be true, but which human beings can prove to be

[*] Rutgers University. The Saul Kripke Center, CUNY, The Graduate Center. E-mail: buechner@newark.rutgers.edu. ORCID: 0000-0002-6679-8209.

true. David Lewis (1969) responded that Lucas (and any other human being) can prove the Gödel sentence for that machine to be true if and only if they can also prove the theorems in Lucas arithmetic. But Lewis doubts a finitary human can do that, since Lucas arithmetic uses infinitary rules of inference—and so there might be infinitely many premises in a given proof. Lucas (1970), in turn, responded that Lewis failed to appreciate the dialectical character of Lucas' argument. Lewis (1979), in response, argued that even appreciating the dialectical character of the Lucas argument, Lucas cannot prove true the Gödel sentence of any finitary machine hypothesized to simulate full human mentality.

Roger Penrose (1989; 1994) improved upon Lucas' argument by proposing a neurobiological mechanism by which human beings might "see" the truth of the Gödel sentence of any finitary computational machine hypothesized to simulate full human mentality. Hilary Putnam argued (1995), famously, that Penrose commits a simple logical error. The finitary computational machine might have a program so long that no human being could physically survey it—and thus not be able to prove that it is consistent. If so, then even if full human mentality is not completely described by that finitary program, our failure to prove its consistency would not distinguish us from the finitary computational machine which (by the Gödel incompleteness theorems) fails to prove its own consistency. If so, the Gödel incompleteness theorems could not be used to arrive at a conclusion that functionalism as a theory of the human mind is a false theory, since it could not be demonstrated that there is an objective truth human minds can verify that no finitary computational machine can verify. The Penrose error is that even if human minds can "see" the truth of the Gödel sentence for the finitary computational machine that is hypothesized to describe human mentality, physically human beings are finite (in terms of time and space limitations). If the program of the finitary computational machine is so long that no human could survey it (such as read it) in their lifetime, then no human being could "see" that it is consistent (if it is). It is a logical error in Penrose's argument, since it is a possibility that, if true, undermines the argument by showing that the conclusion of the argument is false. The burden of proof is on Penrose's shoulders—to show that the possibility cannot be true. But this Penrose cannot do, since the ultimate finitary computational description of human mentality is yet to be written (if, in fact, there is one).

Putnam went on to construct an anti-functionalist argument using the Gödel incompleteness theorems (1988; 1994a; 1994b), applying it to both demonstrative and non-demonstrative reasoning. He does not apply the Gödel incompleteness theorems to a finitary computational program hypothesized to simulate full human mentality. Instead, he exploits the Kaplan-Montague paradox—the basic idea of which is the Computational Liar. The Computational Liar shows—if Putnam is right—that any attempt to formalize human reasoning must fail because any formal description of human reasoning can always be transcended by human reasoning. (Although Putnam does not make it, a distinction needs to be made between (i) prima facie, any formal system can be transcended by another formal system and (ii) any formal description of human reasoning can be transcended by human reasoning. It would be a mistake to reduce (ii) to (i)—that is not what Putnam claims.)

But his argument leads to a dilemma. If not all methods of inquiry are shown to be subject to the Gödel incompleteness theorems, one can take Kreisel's way out. But if all methods of inquiry are subject to the Gödel incompleteness theorems, there is an absurdity. I will provide (in section 7 of this paper) a categorization of the Lucas-Penrose-Putnam anti-functionalist arguments employing the Gödel incompleteness theorems.

What Putnam did not notice is that there is another way to show that human minds and any finitary computational machine hypothesized to simulate human minds are epistemically indistinguishable (even if they are de facto metaphysically distinguishable). What the Gödel incompleteness theorems show is that it is impossible to either prove the Gödel sentence of a formal system subject to the Gödel incompleteness theorems or to prove the consistency of that formal system using finitistic reasoning within that formal system (which delivers its theorems in the epistemic modality of mathematical certainty). Not even an infinitary mind can do that—an infinitary mind would use infinitary reasoning.

However, it is left open that either the Gödel sentence of a formal system subject to the Gödel incompleteness theorems or the consistency of that formal system can be proved with less than mathematical certainty or in some other epistemic modality. Both a human mind and a finitary computational machine might be able to do that. If so, both can prove the same thing, and no difference can be made between the two. This is the lesson from Kreisel's way out of the Gödel incompleteness theorems—and if taken, adds an interesting wrinkle to the Lucas-Penrose-Putnam anti-functionalist arguments. (Roger Penrose, in a preface to a reprinting of *The Emperor's New Mind* [Penrose, 1999], notes that one loophole to his argument is that "our capacity for [mathematical] understanding might be […] inaccurate, but only approximately correct". He says he will address this loophole to his argument in *Shadows of the Mind* [1994], but he does not.)

2. Kreisel's Way Out of the Gödel Incompleteness Theorems

Kreisel (1972) raises the question of whether there is non-mathematical evidence that can be used to establish the soundness of a formal system F (adequate for mathematical reasoning, and so subject to the Gödel incompleteness theorems). He observes that it does not logically follow from the fact that a formal system is subject to the second Gödel incompleteness theorems that there are absolutely no means available to prove its consistency. It only follows logically that its consistency cannot be mathematically demonstrated with mathematical certainty using finitistic reasoning. It is left open that its consistency can be proved by other means, viz., mathematically with less than mathematical certainty (typically by statistical reasoning) and non-mathematically, with less than mathematical certainty, by abstract philosophical reasoning (*a priori* reasoning that is not encodable into a formal system).

He believes that there are two different ways to realize the possibility of nonmathematical evidence to prove the soundness of F, both of which are left open by the Gödel incompleteness theorems. The first kind of nonmathematical evidence to prove the soundness of *F* is inductive evidence and the second is a metaphysical nonmathematical interpretation. Both kinds of evidence require substantial explanation—unfortunately, Kreisel's explanations are brief.

Nonmathematical inductive evidence is taken by Kreisel to be based on our experience with formal systems, such as our experience with Principia Mathematica. In one way of understanding what our experience of formal systems delivers, our confidence in the soundness of formal systems is acquired by various case studies of formal systems. Kreisel rejects this view—calling it a sham—for two distinct reasons. The first reason is that we have little or no experience of proving the soundness of a formal system by inductive methods. From this Kreisel thinks it follows that we have no good ideas about what are the appropriate statistical principles that would be used in evaluating the inductive evidence. Without statistical principles we have a data set, but no means by which to find in it the data which is necessary for establishing the soundness of some formal system. Whatever statistical principles we choose, one job which they must be able to do is to ascertain that the nonmathematical inductive evidence establishes that the entire formal system is sound, and not that only some subsystem of the formal system is sound.

The second reason Kreisel rejects the idea of nonmathematical inductive evidence for establishing the soundness of a formal system is that it is not done by using the experience we acquire from case studies of soundness proofs of formal systems. It is, instead, done by—at least in the case of Principia Mathematica—reflection on the intended meaning of the terms in the language of Principia Mathematica. However, what is interesting about Kreisel's point is that the act of reflecting upon what is the intended meaning of the terms in the language of a formal system may or may not be a computable procedure. There might not be a computational description of such acts. If there is no computational description of such acts, then there is some cognitive activity that humans can do which no machine can do. In which case, there would be a difference between humans and machines even if neither humans nor machines can prove the Gödel sentence of some formal system. Of course it would be a research project to show that acts of reflection upon the intended meanings of terms in some language (whether it is a formal language or not) have no computational description. (We shall see below that, using an ingenious Gödelian argument, Putnam attempts to close the door on both statistical methods and abstract philosophical methods for demonstrating CON(**PA**) by arguing that they are subject to the Gödel incompleteness theorems.)

The other way of proving the soundness of *F* is by an abstract but nonmathematical interpretation of *F*. Kreisel cites as an analogy the identification in intuitionistic mathematics of what is mathematical with what is intuitionistically acceptable. He notes that in intuitionism set-theoretic concepts are metaphysical and then claims that it might be possible to establish the soundness of some set-theoretic formal system using a metaphysical nonmathematical interpretation. Kreisel believes that this way of proving the soundness of *F* is more realistic than using inductive evidence to establish the soundness of *F*. I don't know what he means by "realistic" in this context. Perhaps he means that there is a wealth of mathematical and foundational work in intuitionism, and so we have a better understanding of what an abstract nonmathematical interpretation of *F* would look like than we do of statistical principles.

An interpretation is usually understood to be a map from syntactical objects (that is, symbols) to objects which need not be syntactical—perhaps mathematical objects. What, then, is a nonmathematical interpretation? Could it still be a map and yet be nonmathematical? And what does it mean to say it is metaphysical? Kreisel restricts the metaphysical nonmathematical interpretation to an abstract metaphysical nonmathematical interpretation. But if it is a map and it is abstract, it is not clear how it could not be mathematical.

Regardless of what Kreisel actually means by a metaphysical nonmathematical interpretation of *F*, using it to establish the soundness of *F* is different from proving the soundness of *F* within a classical formal system using finitary reasoning in the following respect: the proof of soundness of *F* within a classical formal system using finitary reasoning will be with mathematical certainty. (See below for a discussion of Church's view that the theorems of a given system of logic are proved with mathematical certainty.) On the other hand, the proof of the soundness of *F* using a metaphysical nonmathematical interpretation will perhaps not be with mathematical certainty. Kreisel's way out is the use of statistical proofs of consistency of **PA** with less than mathematical certainty or proofs in another epistemic modality such as (nonmathematical philosophical proofs). For more on the epistemic modality of a proof see 4.1 below.

3. Penrose on the Role of Trust in Mathematics

The key idea of Kreisel's way out is that one might be able to prove CON(**PA**) with less than mathematical certainty (using statistical methods) or in some other epistemic modality (such as metaphysical nonmathematical reasoning). Throughout the rest of this paper we will see how these possibilities enter into the Lucas-Penrose-Putnam anti-functionalism arguments. Recently Penrose has argued that trust plays an important role in mathematical proofs (2016). He claims that in order to trust a mathematical argument, we must trust that the rules of the formal system are sound. In cases where it cannot be established that the formal system is consistent because of the restriction imposed by the second Gödel incompleteness theorem, we need to trust that the formal system is consistent. If we do, then we can prove true the Gödel sentence and the consistency of that formal system by ascending to a stronger formal system—which we trust to be consistent.

We can view trust in the soundness of the rules of a formal system as an epistemic modality alternative to mathematical certainty delivered by proofs in a formal system. What Penrose fails to see is that if a finitary computational machine can meaningfully trust a formal system to be consistent, then there is no metaphysical difference between it and human minds. The move Penrose makes to show that human minds can determine the consistency of CON(**PA**) is one which defats his anti-functionalist argument, since it is open that finitary computational machines can do the same. The burden of proof is upon Penrose—to show that no finitary computational machine can exhibit the attitude of trust. (See Buechner, 2011, for an argument that finitary computational machines can engage in relations of trust with other finitary computational machines and with human beings.)

4. Two Uses of the Gödel Incompleteness Theorems in Refuting Functionalism

I introduce a distinction between two different uses of the Gödel incompleteness theorems in anti-functionalist arguments. This distinction has not been made in the literature—and it is important to make it because the conclusions of the arguments made under each use are significantly different. Perhaps the reader is puzzled: "Isn't there only one use of the Gödel theorems in refuting functionalism?" There are two different ways in which one can attempt to refute functionalism using the Gödel incompleteness theorems, and the conclusions about functionalism differ in each. Additionally, each method of refutation opens up different possibilities in the Lucas-Penrose-Putnam anti-functionalism arguments.

4.1. Metaphysical Uses of the Gödel Incompleteness Theorems in Refuting Functionalism

One way of using the Gödel incompleteness theorems in anti-functionalist arguments concludes that the human mind does not have the nature of a finitary computational machine, in which case, functionalism is false. This refutation establishes a metaphysical difference between human minds and finitary computational machines: human minds do not have the nature of such machines.

The metaphysical use of the Gödel incompleteness theorems in refuting functionalism is found in (Gödel, 1995; Lucas, 1961; Penrose, 1989): if it can be shown there is a mathematical truth that can be proved by a human mind, but that cannot be proved by a finitary computational machine (that, by hypothesis, finitely computationally models that human mind) then the human mind is not computationally modeled by that finitary computational machine. Whatever is the nature of the human mind, it does not have the nature of a finitary computational machine, since the human mind is different from the finitary computational machine in virtue of its causal powers, which enable it to prove a theorem that the latter cannot prove.

Another way of putting the same point: the human mind can prove that the program of the finitary computational device which purports to model it is correct, while the program cannot prove of itself that it is correct (assuming that there is no additional program embodied in the finitary computational device). So there is a cognitive power that the human mind possesses that is not possessed by the finitary computational machine. A human mind could justify the truth of the claim that the program that purports to describe it is correct, while the program itself cannot do that. But if the program, by hypothesis, describes all of the cognitive powers of the human mind, then it cannot be a complete finitary computational description of the human mind, since it lacks (at least) one cognitive power a human mind possesses.

This application of the Gödel incompleteness theorems shows functionalism is a false philosophical view by demonstrating that human minds are not identical with finitary computational machines. This non-identity claim is a metaphysical claim about the nature of the human mind: they do not have the nature of finitary computing machines. Functionalism is the view that human minds are identical with finitary computational machines (of some kind). The metaphysical argument (using the Gödel incompleteness theorems) demonstrates that human minds are not identical with finitary computing machines. Hence functionalism is false if the metaphysical argument is sound.

The Gödel incompleteness theorems (in the context of this metaphysical argument) provide a mathematical proof that the human mind is not identical to a finite computing machine and thus does not have the nature of a finite computing machine. (This claim can be generalized: the Gödel incompleteness theorems provide a mathematical proof that the human mind is not identical to any kind of finite computing machine and thus does not have the nature of any kind of finite computing machine. It can be generalized because the Gödel sentence unprovable in finitary computing machine1 can be proved in a stronger finitary computing machine2. However, a new Gödel sentence can be expressed in finitary computing machine2 that cannot be proved in it. This is true for all finitary computing machines.) So we have a mathematical proof of a negative metaphysical claim about the human mind: it is not any kind of finitary computing machine. We will call this use of the Gödel theorems "MGF" ("Metaphysical claims that are consequences of using the Gödel theorems to refute functionalism".)

It would be a mistake to claim that the Gödel incompleteness theorems specify an exact bound on the extent of the metaphysical difference between human minds and a given finitary computing machine. For instance, given a finitary computing machine that cannot prove its program is consistent, the extent to which the human mind differs from it is that the human mind can prove the program is consistent. This is not informative, since it says nothing positive about the cognitive functions necessary for human minds to prove that the program describing their mentality is consistent. It does say something negative, though. It says that no human mind can prove the program is consistent by simulating a finitary computing machine.

What is not usually addressed in metaphysical refutations of functionalism that use the Gödel incompleteness theorems is the epistemic modality of the provability relation in the formal system in which the reasoning occurs. A (sound) proof in a formal system (whether or not it is subject to the Gödel incompleteness theorems) proves a theorem with mathematical certainty. Our justification for believing the theorem is true is that it has been proved with mathematical certainty. So the Gödel theorems need to be qualified: the second incompleteness theorem says that no formal system subject to the Gödel incompleteness theorems can prove its own consistency with mathematical certainty. Here the epistemic modality—the way in which we come to know the truth of the claim made in the proof—is mathematical certainty. But there are other ways than mathematical certainty by which we can come to know the truth of a claim made in a proof. As the epistemic modality of a proof changes, so does the nature of the proof.

It is left open by the Gödel theorems that the formal system can prove its consistency with less than mathematical certainty or in some other epistemic modality. A statistical proof that a formal system (that is subject to the Gödel incompleteness theorems) is consistent has less than mathematical certainty. (Probabilistic proofs have this feature; see Wigderson, 2019.) A nonmathematical philosophical proof that such a formal system can prove its consistency would be a proof in another epistemic modality than that of a proof in logic or in mathematics. A proof using diagrams or pictures would be a proof in an epistemic modality other than mathematical certainty because the nature of a picture proof differs from the nature of a proof in a system of logistic. Intuitionistic reasoning in Brouwer's version of intuitionism might also be an example. Only a proof using a symbol system found in the formal languages of logic or in classical mathematics would have mathematical certainty. (Understanding in what epistemic modalities other than mathematical certainty there can be proofs of mathematical truths is an important and open research topic.)

If the only means of achieving mathematical certainty that *S* is true is to prove S in a formal system by finitistic reasoning within that formal system, then if S is either a Gödel sentence for that formal system or a consistency claim about that formal system, it follows that no human being (whether finitary in its cognitive powers or infinitary in its cognitive powers) can prove S is true with mathematical certainty using finitary reasoning within that formal system. So no human mind can prove the master program for a finitary computing machine simulating human mentality is correct with mathematical certainty by engaging in finitistic reasoning described by that master program. If so, human minds are indistinguishable from the finitary computing machine. On the other hand, there is no prohibition on the human mind proving the correctness of the master program with either less than mathematical certainty or in some other epistemic modality. But neither is the finitary computing machine prohibited from this, either. (This is so, unless proof with less than mathematical certainty or in another epistemic modality is subject to the Gödel incompleteness theorems. In that case, it is ruled out for the finitary computing machine to do that. But then it is also ruled out for human beings to do so as well.) If human minds can perform infinitary reasoning, and can prove the correctness of the master program using infinitary reasoning, this would distinguish human minds from finitary computing machines (which, by definition, cannot perform infinitary reasoning). But since it is an open question whether human minds can perform infinitary reasoning, this line of argument cannot establish its conclusion.

If the MGF argument is sound, then we know, with mathematical certainty, that we are not finitary computing machines. What is the provenance of the qualifier "mathematical certainty"? The Gödel theorems show that any formal system subject to the them cannot prove its Gödel sentence nor its consistency sentence with mathematical certainty using finitistic formalizable reasoning within that formal system. Why mathematical certainty? Why not logical certainty? Because there are different systems of logic—such as relevance logic—what is provable with logical certainty in one kind of logic might not be provable in some other kind of logic. Since the finitary reasoning in classical first-order logic can be described mathematically, the theorems of that logic are said to be proved with mathematical certainty.

Where does the claim that proofs in a formal system of logic carry mathematical certainty come from? Alonzo Church (1956) uses the phrase "mathematical certainty" in his discussion of proofs in mathematics that are translated into first-order logic. For Church, the only way to achieve mathematical certainty is a proof system where the axioms are effectively specified and in where, for any line in the proof, there is an effective procedure by which one can tell that it is an authentic line in the proof. This finitary reasoning in first-order logic can be described mathematically. An auditor of a proof

> [M]ay fairly demand a proof, in any given case, that the sequence of formulae put forward is a proof; and until this supplementary proof is provided, he may refuse to be convinced that the alleged theorem is proved. This supplementary proof ought to be regarded […] as part of the whole proof of the theorem, and the primitive basis of the logistic system ought to be so modified as to provide this, or its equivalent. (Church, 1956, p. 53)

The only logistic systems for which Church's requirement is satisfied are those in which the axioms and the rules of inference are effectively specified—these are finitary proof systems in which there are only finitely many lines in a proof and the pedigree of each line in the proof can be effectively ascertained. Infinitary logistic systems are different, for rules of inference are not effectively specified. A mind that has infinitary capacities can effectively specify them, but the notion of "effectiveness" then belongs to alpha-recursion theory, a theory of effectivity for infinite minds. Church obviously assumed human minds are finitary in his discussion.

So if the MGF argument is sound, then we know, with mathematical certainty, that human minds are not identical with any kind of finitary computing machine. This is an extraordinarily strong claim. Compare it with the following claim: we know, with mathematical certainty, that B follows from A and $A \rightarrow B$. This claim is trivial. On the other hand, one does not know with mathematical certainty that one is (now) looking at a tree. The claim an MGF argument makes is strong, then, in the sense that the information it establishes about the nature of the human mind has important value. (I do not suggest, in using the phrase "extraordinarily strong", that the claim is thereby unlikely to be true.)

But the strength of the claim should make us suspicious of it. The assumption that underlies the metaphysical claim is that human minds can prove the correctness of the finitary computing machine's master program (for simulating human mentality). But we have seen that this assumption needs to be qualified: human minds can prove, with mathematical certainty using finitistic reasoning, the correctness of the computing machine's master program. This, though, is highly unlikely to be true. If a human mind has infinitary cognitive capacities, it might do so (for instance, by employing Turing's infinitary procedure; see Turing, 1939). But do we have infinitary cognitive capacities? Some philosophers and cognitive scientists believe we do not have infinitary cognitive capacities. Others believe that we do. So a stalemate is reached in the absence of evidence concluding one or the other position.

If the assumption underlying the MGF argument is changed by changing the qualification to "with less than mathematical certainty or in some other epistemic modality", then the MGF argument cannot establish its conclusion, since it is also available for a finitary computing machine to prove the correctness of its own master program with less than mathematical certainty or in some other epistemic modality. Thus the metaphysical claim is bankrupt and the refutation of functionalism using the Gödel incompleteness theorems is drained of its force. This is a significant philosophical result overlooked in the anti-functionalism debate. If it is true that human minds are not completely describable by a finitary computational machine and that human minds are able to verify the consistency of Peano arithmetic, i.e., CON(**PA**), how is it done? It cannot be done by employing a recursively axiomatized finite proof system to do it, since for any such proof system (strong enough to capture arithmetic), the Gödel incompleteness theorems apply. On the other hand, if we use a recursively axiomatized finite proof system which is too weak to be subject to the Gödel incompleteness theorems, then this will not distinguish us from any finitary computational machines, since finitary computational machines are also capable of proving theorems in such weak proof systems.

In such a finitary proof system, there is nothing human minds can prove which a finite computational machine (of the appropriate kind) cannot prove. How, then do we differ from the finite machine? We know from Gentzen's proof of CON(**PA**) by transfinite induction, that infinitely long derivations can secure CON(**PA**). We also know that within formalized systems of Peano arithmetic, proofs of transfinite induction for any ordinal up to, but not including the infinite ordinal epsilon$_0$, are available. However, we need transfinite induction along a well-ordered path of length epsilon$_0$ to prove CON(**PA**). The issue, then, is this:

if human minds know the truth of CON(**PA**) with mathematical certainty, is the only mathematical method by which we do it the use of infinitely long derivations? There cannot be a finitary method of reasoning that proves CON(**PA**) with mathematical certainty within the formal system for **PA**. One can find stronger formal systems in which CON(**PA**) can be proved by finitistic reasoning, but only if CON(stronger formal system) can be verified. If it is verified, then we do it this way only if we have infinitary cognitive capacities, and that is at present an open question.

4.2. Epistemic Uses of Gödel's Incompleteness Theorems in Refuting Functionalism

MGF arguments show the nature of the human mind differs from the nature of physical finitary computing machines. MGF arguments are philosophically satisfying, since they rule out one metaphysical possibility about the nature of the human mind—that our minds have the nature of finitary computational machines. Even though they do not have the resources to describe the true nature of the human mind, their importance lies in showing what the human mind is not. But MGF arguments are not the only use of the Gödel theorems in the functionalism debate. Even if we assume that human minds are finitary computing machines, we can still enlist the Gödel incompleteness theorems to make philosophically important claims about the human mind. Call these uses of the Gödel theorems "EGF" ("Epistemic claims that are consequences of using the Gödel incompleteness theorems to refute functionalism"). There are two different kinds of EGF arguments.

4.2.1. The first kind of EGF argument.

Assume that human minds are finitary computational in nature. (However, the argument is the same if human minds cannot be fully described by finitary computational machines.) Suppose human cognition is finitely computationally described by computer program *P*. If we assume human beings can prove truths of Peano arithmetic, *P* is subject to the Gödel incompleteness theorems (since *P* must be equipped with enough syntax to arithmetize metamathematics, which is necessary for the Gödel theorems to take root). CON(*P*) expresses the consistency (or correctness) of *P*. Since it is equivalent to *P*'s Gödel sentence, it follows that *P* can't prove it is consistent. Assuming we are correctly described by *P*, human beings cannot verify the consistency of *P*.

Since the project of cognitive science is to find *P*, then that project can never be epistemically justified (since it cannot be established that *P* is consistent). Any science of the human mind that views the human mind as a finitary computing machine will not be able to epistemically justify its claims, because we cannot verify that the correct program of the finitary computing machine is consistent. Human beings will not be able to prove, with mathematical certainty, *P* is consistent. Human beings cannot prove the consistency of *P* in the epistemic modality of mathematical certainty. To do so, our reasoning about *P* would have accord with that of a finitary computing machine, to which the notion of "proof with mathematical certainty" applies. This is a radical form of philosophical skepticism: we have a mathematical proof (of which we are mathematically certain) that we cannot know, with mathematical certainty, the correct computational theory of how our minds work.

EGF arguments do more than provide a new form of philosophical skepticism. They also address the competence/performance distinction essential for the viability of cognitive science. A critical distinction is made in cognitive science between how the human mind actually works and how it ought to work—between a performance level description and a competence level description of the human mind. Without such a distinction, the very idea of a psychological law is jeopardized. EGF arguments show three basic assumptions essential for cognitive science to be viable cannot consistently obtain: (i) that the human mind can be represented (at a level of computational description) by a computational device, (ii) that its cognitive capacities can be viewed as finitely computable functions and (iii) that there is a competence description of the human cognitive mind. The Gödel incompleteness theorems show the first two assumptions are incompatible with the third. If we take the first two to be part of Marr's (2010) implementation level and the third to be Marr's theory of the function (the what, i.e., the function, which is computed), Gödel's theorems reveal an incompatibility in Marr's foundational program for cognitive science. (For details, see Buechner, 2010.)

4.2.2. The second kind of EGF argument.

Assume that human minds are not finitary computational in nature (but that we do not know this fact). If so, any finitary computational machine conjectured to describe human mentality fails to do so—it either fails to describe all of human mentality or else it falsely describes parts of human mentality. Suppose it is conjectured human mentality is correctly described by computer program *P*, which is subject to the Gödel incompleteness theorems. Suppose, additionally, the length of *P* is infeasibly long for a human being to survey. In which case, no human being will be able to establish that *P* is consistent.

Since no human being will be able to verify that *P* is consistent (which is an epistemic claim), we cannot use the mathematical theory of computation or cognitive science to show that there is a metaphysical difference between human mentality and a finitary computational machine. Although this kind of EGF argument does not refute functionalism, it reveals a shortcoming in it—that we cannot use it to establish metaphysical claims about the human mind. Additionally, since cognitive science and functionalism might be false theories (if *P* is inconsistent), any psychological claims made within cognitive science and any philosophical claims made within functionalism might be false, and we could never fully justify those claims no matter how much evidence we had supporting them.

4.3. Correct and Incorrect Readings of the Gödel Theorems

In arguments that use the Gödel theorems to attempt to refute functionalism and in critical discussions of those arguments, an obvious point has been overlooked. What the Gödel incompleteness theorems show is that there is no mathematically certain finitistic mathematical proof of the Gödel sentence and the consistency sentence of any formal system susceptible to the Gödel theorems. We cannot fintistically prove, with mathematical certainty, the Gödel sentence and the consistency sentence of Gödelizable formal systems. What is overlooked is the epistemic modality of mathematical certainty that qualifies the proof relation. Perhaps it is overlooked since the method of proof within a system of logic is what delivers mathematical (or logical) certainty.

The standard reading is that we cannot prove CON(**PA**), period. By failing to qualify "prove", it appears the claim is that there is no proof of any kind of CON(**PA**). This is an incorrect reading of the Gödel incompleteness theorems. The correct reading is that we cannot prove CON(**PA**) with mathematical certainty by finitistic reasoning in a formal system for **PA**. (John von Neumann, in his tribute to Gödel, notes that "for no such system can its freedom from inner contradiction be demonstrated with the means of the system itself" [1969, p. x]. This is a correct reading of the Gödel incompleteness theorems.)

It does not follow, however, that we cannot prove CON(**PA**) with less than mathematical certainty or prove it in some other epistemic modality than mathematical certainty (as Kreisel rightly noted). (The claims of statistical proofs are with less than mathematical certainty. Epistemic modalities other than mathematical certainty might include pictorial proofs and nonmathematical philosophical reasoning.) The same remarks hold if we transpose the discussion of the Gödel incompleteness theorems to the context of what we know about CON(**PA**). If we substitute "know the truth of" for "prove", the same point applies. We cannot know the truth of CON(**PA**) with mathematical certainty. It is left open by the Gödel theorems that we can know the truth of CON(**PA**) with less than mathematical certainty and that we can know the truth of CON(**PA**) in some epistemic modality other than mathematical certainty.

If we accept a mathematical epistemology in which we can know mathematical propositions with less than mathematical certainty or in some other epistemic modality than mathematical certainty, new possibilities become available for the functionalism debate. For instance, if there are formal systems (in which the Gödel incompleteness theorems hold) in which CON(**PA**) is proved with less than mathematical certainty and the epistemic modality in which it is proved satisfies a reasonable notion of epistemic justification, then the limitations of the Gödel incompleteness theorems might be dramatically circumvented. Substitute "the correctness of its own computer program" for "CON(**PA**)" in the preceding sentence. If an anti-functionalist enlists the Gödel theorems to refute functionalism, she must show that the notion of justification under which a finite machine can prove the correctness of its own computer program with less than mathematical certainty is normatively bankrupt. Suppose that human beings are finitary computational machines. Define the goal of cognitive science to be discovery of the master computer program for the human mind. Assume the cognitive activities cognitive scientists engage in when they attempt to discover the master computer program are themselves described in that program. Suppose that in the future a cognitive scientist claims to have found the master computer program. Do we require that her belief that this is the correct master computer program must be mathematically certain in order to count as being epistemically justified? Whether that requirement does or does not appear to be too strong, it is clear that it is a question that must be addressed wherever the Gödel theorems are enlisted in the functionalism debate.

Even within mathematics there is evidence that this demand is negotiable. Mathematical proofs not formalized within a system of logic do not satisfy the stringent demands of mathematical certainty. Only proofs that are formalized in a formal system whose axioms, rules of inference and application of rules of inference are recursively specified can satisfy those stringent demands. Proofs in, for instance, algebraic topology do not meet them, though mathematicians do not feel that they need to translate those proofs into a formal system before they can be said to know (with adequate justification) the truths of algebraic topology.

The consequence is that no finitary being can prove CON(**PA**) finitistically with mathematical certainty. The reason this is so is obvious. If mathematical certainty is secured only in virtue of a finitistic proof within a system of logic, no finite being can prove CON(**PA**) with mathematical certainty unless they construct a finitistic proof of it within a system of logic. But the Gödel theorems forbid this. (A being with infinitary powers can construct a proof of CON(**PA**) with mathematical certainty only if constructions in a system of logic requiring infinitary operations confer mathematical certainty upon the theorems proved within that system. Church did not consider this matter in his discussion of mathematical certainty.)

When anti-functionalists, such as Penrose, claim that human beings can know CON(**PA**) they must qualify their claim. We cannot know CON(**PA**) with mathematical certainty. But if we can know it with less than mathematical certainty or in some epistemic modality than mathematical certainty, it is possible that a finitary computational machine can acquire that knowledge as well. If so, the Gödel incompleteness theorems cannot drive a wedge between what a human being can know and what a finitary computational machine can know.

5. Putnam's First Version of His Argument That Not All Methods of Inquiry Can Be Formalized

An early argument Putnam (1988) uses against the view that methods of inquiry can be formalized by a finitary computational machine is his Gödelian argument that that there can be no prescriptive competence description of human reasoning (including the reasoning in mathematical proofs). Suppose that there is a description *P* of human prescriptive mathematical competence. There will be many functions that are provably recursive according to *P*. List the index of each partial recursive function that *P* can prove to be total recursive. There will be infinitely many functions on this list—since a mathematician can (in principle) prove infinitely many functions are general recursive. This list of functions can be diagonalized, and the diagonal function will be total, since there are infinitely many functions on the list.

However, if it could be proved that *P* is a sound proof procedure, it could also be proved that the diagonal function is a total recursive function. Unfortunately, such a proof would also show that *P* is inconsistent. Why is that? Suppose that the proof is on the list—in which case, the diagonal function would be on the list. But by the definition of a diagonal function, if it is the j^{th} member on the list, then diagonal function (j) = diagonal function (j + 1). It follows that any formalization of human mathematical proof ability cannot both (i) be sound and (ii) can be proven to be sound using human mathematical proof abilities.

Putnam's conclusion needs to be emended: no formalization of human mathematical proof ability can both be sound and be such that it is part of human mathematical proof ability to finitarily prove that soundness, with mathematical certainty and from within *P*. We cannot prove with mathematical certainty and finitistic reasoning that *P* is correct. It follows that we cannot prove with mathematical certainty and finitistic reasoning that the competence theory for human mathematical proof ability is correct.

It is impossible for us—whether we are or are not subject to the Gödel incompleteness theorems—to finitarily prove with mathematical certainty from within *P* that the competence level description is true of us. If we were able to finitarily prove it is true of us, with mathematical certainty and from within P, we would have proven that the formal theory encapsulated by the competence description is consistent. But this is prohibited by Gödel's second incompleteness theorem. Notice we would have to ascend to a stronger computational system to finitarily prove, with mathematical certainty, the consistency of our competence description. If so, then the competence description that we finitarily prove to be correct, with mathematical certainty, in the stronger system is not our competence description. Since we ascended to a new computational system, the competence description of the weaker computational system is no longer true of us.

Suppose that human minds are not subject to the Gödel incompleteness theorems. The Gödel incompleteness theorems rule out the possibility that a finitary human mind can finitarily prove, with mathematical certainty, that a finitary computer program that simulates it is correct. What this means is that whether human minds are or are not subject to the Gödel incompleteness theorems, the human mind cannot finitarily prove with mathematical certainty that a program that simulates it is correct. Thus whether human minds are or are not subject to the Gödel incompleteness theorems, they cannot justify claims in cognitive science about its computational structure. EGF arguments do not need to show that there is something a human mind can do that any finitary computing machine cannot do in order to make philosophically interesting claims about the mind. In this case, the claim concerns the limits of cognitive science in providing a rigorous, scientific study of the human mind.

EGF argument (such as the one Putnam makes above) must (as we argued earlier) make a very strong assumption: that justifications of claims in cognitive science are mathematically certain. This follows from the use of the Gödel incompleteness theorems. We know, with mathematical certainty, that we cannot, with mathematical certainty, finitarily prove the correctness of the program, *P*, that describes our competence. If *P* is the master program for human cognition, we can't mathematically prove it is correct with mathematical certainty. Do any other scientific disciplines impose such stringent epistemic requirements upon the claims they make? I think it is too high a price to ask of cognitive science, and one that is incompatible with the epistemic demands other scientific disciplines impose upon their own claims. This is an important issue that deserves further attention.

Notice that statistical methods and proof methods in an epistemic modality other than that of mathematical certainty (we will call them 'weak methods') will be included in *P*. There's no absurdity or inconsistency in this inclusion, since they do not finitarily prove the correctness of *P* with mathematical certainty. Rather, they prove it with less than mathematical certainty or in some other epistemic modality. The central issue for EGF arguments is what we should take as the standard of epistemic justification of *P*. If we take the standard of epistemic justification to be mathematical certainty, then they refute computational functionalism. If the standard is less than mathematical certainty or some other epistemic modality, they lose all their potency in refuting functionalism.

This version of Putnam's anti-functionalist argument using the Gödel incompleteness theorems—that there can be no prescriptive competence description of human mathematical reasoning—succeeds only if the epistemic modality of the proof relation is that of mathematical certainty achieved by finitistic reasoning. Where that is not the case, the argument fails.

6. An Exposition of Putnam's Second Gödelian Argument Against Functionalism

Whether there is or is not a finitary computational description of total human mentality is an open question. However, if we cannot (now) know the ultimate finitary computational description of total human mentality—should there be one—then we cannot (now) know whether its program is (or is not) infeasibly long. This presents an irresolvable difficulty for any MGF or EGF arguments—such as the Lucas-Penrose arguments. To assume the program is feasibly long—and one which can be shown consistent by human minds—is a logical error. Putnam diagnosed this error in Penrose's argument. As we saw earlier, since is it possible the program is infeasibly long, it is therefore possible that even if human minds do not have a complete finitary computational description, they cannot be distinguished from finitary computational machines because they will not be able to prove the consistency of an infeasibly long program. (Even if we do have infinitary minds, our physical bodies in some of their aspects are finitarily restricted—and so we would not be able to read all of the lines in a program which is infeasibly long.) To neglect this possibility is a logical error. Yet Putnam makes a Gödelian argument against functionalism without making either logical error—he does not assume the program is feasibly long and he does not have to consider the possibility that it is infeasibly long. How is it done?

One way out of this difficulty for EGF and MGF arguments is to show that all epistemically justified methods that prove CON(*P*) with less than mathematical certainty or in some other epistemic modality (the weak methods) are subject to the Gödel incompleteness theorems. Putnam claims that all weak methods are subject to the Gödel incompleteness theorems. This argument appears in *Reflexive Reflections* (Putnam, 1994b). The argument employs Gödel's second incompleteness theorem. In what follows, I use the acronym "PGA" ("Putnam's use of the second Gödel incompleteness theorem in his argument that all weak methods are subject to the Gödel incompleteness theorems").

PGA claims that our prescriptive inductive competence is subject to the Gödel incompleteness theorems. Putnam cites his earlier work on Carnapian inductive logics and on computational learning theory, only to assert that it does not matter whether this work is taken into account in PGA, since PGA will assume there is some finitary computational description of our prescriptive inductive competence and that one does not need to know what that description looks to make the PGA argument. "P" denotes a finitary computational description of our inductive (or non-demonstrative) and demonstrative prescriptive competence.

Putnam uses an idea in the Montague-Kaplan *Paradox of the Knower* (Feferman, 1960) that is an application of self-reference. It is The Computational Liar (CL):

(CL) There is no evidence on which acceptance of the sentence CL is justified (Putnam, 1994b)

CL is arithmetizable, and its arithmetization is a sentence of arithmetic to which the Gödel diagonal lemma applies. The diagonal lemma tells us that for any predicate that is definable in the language of Peano arithmetic, there is some sentence that is true if and only if its Gödel number is false of that predicate. The diagonal lemma allows us to couple *P* with CL.

> It follows from Gödel's work that there is a sentence of mathematics which is true if and only if *P* does not accept that very sentence on any evidence, where *P* is any procedure itself definable in mathematics—not necessarily a recursive procedure. (Putnam, 1994b)

In an important caveat to CL, Putnam says that "[…] if the inductive logic *P* uses the notion of degree of confirmation rather than the notion of acceptance, then one replaces 'is justified' by 'has instance confirmation greater than .5', […]" (1994b, p. 426, note 5). This is significant, since the notions of a justified belief and of acceptance of a justified belief play critical roles in non-quantitative models of inductive reasoning, while "has instance confirmation greater than .5" and "degree of confirmation" play critical roles in both quantitative and logical models of inductive reasoning. This caveat gives us reason to think that Putnam takes *P* to be a computational description of any kind of inductive reasoning and not just logical models of inductive reasoning, such as those found in computational learning theory.

If there is evidence which justifies the acceptance of CL, it easily follows that CL is false, and it is a sentence of pure mathematics. Since *P* formalizes our prescriptive competence in demonstrative and non-demonstrative reasoning, our (fully justified) reasoning tells us to accept a mathematically false proposition.

The negation of CL is that there is evidence on which the acceptance of CL is justified. If there is evidence on which the acceptance of the negation of CL is justified, then we know from what was just established above that CL is a mathematically false sentence. (Putnam notes that it is an omega inconsistency.) It follows that should *P* converge on CL—that is gives an answer to CL—to which we are justified (by *P*), then that evidence for the answer licenses us to accept a mathematical falsehood. So it has been established that CL cannot be shown true or shown false using *P*, which is a computational description of our prescriptive competence in demonstrative and non-demonstrative (inductive) reasoning. (Gödel assumed that the formal system in which he worked is omega-consistent in order to show that proof of the negation of the Gödel sentence leads to contradiction, in this case, an omega-inconsistency. Omega-consistency is weaker than consistency. If a formal system is omega-consistent, it follows that it is consistent. Putnam makes the same assumption.)

Given that anyone is justified in believing that if *P* converges on CL, it licenses one to believe a sentence that is mathematically false, Putnam formulates a criterion of adequacy (CA) for accepting any formalization of human prescriptive demonstrative and non-demonstrative competence

(CA) The acceptance of a formal procedure *P* as a formalization of (part or all) of prescriptive inductive (demonstrative and non-demonstrative) competence is only justified if one is justified in believing that *P* does not converge on *P*'s own Gödel sentence (i.e., CL) as argument.

From CL and CA, it follows that no human being can demonstrate that *P* is prescriptive whenever our minds work in the exact way that *P* says they should work. When we believe CA and also believe that *P* is both complete and also correct in describing our prescriptive demonstrative and non-demonstrative competence, it easily follows that we will believe that *P* does not converge on CL. However, that is to believe CL. But notice that this belief is justified, and that (by assumption) all justification of beliefs can be formalized in *P*. Since we are committed to believing CL, we are in a contradiction. That is Putnam's ingenious PGA.

Notice that Putnam has not made any claims that there is something human minds can do that no finitary computing machine can do, nor has he assumed that *P* is feasibly long. (That is why Putnam does not commit the logical error that Penrose commits). He has, though, shown that *P* could not be justified within cognitive science without licensing us to believe a contradiction. One consequence of PGA is that any formal theory proposed in cognitive science of how we do inductive reasoning cannot be justified without also licensing us to believe a contradiction. (This is a disturbing and important result that has not caught the attention of cognitive scientists working on the problem of formally characterizing inductive reasoning.)

6.1. PGA and the Kaplan-Montague Paradox

Is it really the case that the key terms in CL can be arithmetized? If they cannot be arithmetized, then PGA fails. I contrast Putnam's Computational Liar with the version that Kaplan and Montague (1960) constructed in order to show the Gödel incompleteness theorems extend to the modal predicates "knowledge" and "necessity". Kaplan and Montague needed to find for the knowledge predicate suitable analogues of the Hilbert-Bernays derivability conditions for the provability predicate. Montague employed a weak epistemic system consisting of the four schemata:

(i.) $K\alpha \rightarrow \alpha$

(ii.) $K\alpha$, if α is an axiom of first-order logic

(iii.) $K(\alpha \rightarrow \mu) \rightarrow (K\alpha \rightarrow K\mu)$

(iv.) $K(K\alpha \rightarrow \alpha)$

Montague (1963) appreciated Tarski's insight (1983), in the latter's proof of the indefinability of truth in first-order logic, that two prima facie consistent theories cannot always be combined into a consistent theory. In Tarski's indefinability work, Robinson arithmetic relativized to ß cannot be combined with Tarski's schema for the language of Robinson arithmetic relativized to ß and extended with a truth predicate T. Montague saw that this insight can be generalized: two prima facie true theories, one a theory of its own syntax and the other a theory that has principles capturing the logic of concepts such as knowledge, belief or necessity, cannot be combined into a consistent theory. The tool necessary for the proof is the Gödel diagonal lemma:

Suppose T is an extension of Robinson arithmetic relativized to ß. Let α be a formula whose only free variable is v_0. Then there is a sentence ζ such that:

$$\vdash T\ \zeta \text{ if and only if } \alpha(\zeta/v_o)\text{, where,}$$
$$\text{if } n \text{ is the Gödel number of } \zeta,\ \zeta \text{ is the } n^{\text{th}} \text{ numeral.}$$

The key to the Montague-Kaplan proof is the fact that knowledge is a property of "proposition-like" objects recursively built from atomic constituents. Given enough arithmetic, it is easy to associate with each "proposition-like" object a Gödel number. Then, structural properties and relations between "proposition-like" objects can be arithmetically simulated by explicitly defined arithmetical predicates of the Gödel numbers of the "proposition-like" objects.

Recall Putnam's Computational Liar:

CL There is no evidence on which acceptance of the sentence CL is justified.

We need to arithmetize the properties and relations in CL in order to use Gödel's diagonal lemma. Can "evidence", "acceptance", and "justified" be arithmetized? It is not obvious that they can. Consider the ramified type theory in Russell and Whitehead's *Principia Mathematica*. No one has succeeded in showing it is subject to the Gödel incompleteness theorems, for there is no general theory of the intensional provability relation. It will do no good to simply assert that consistency cannot be proved within any sufficiently strong system because Gödel's second incompleteness theorem tells us this. Richmond Thomason (1980; 1989) has pointed out in this connection that "it has never been possible to state the [second incompleteness] theorem at this level of generality with a degree of precision that will support a mathematical proof" (1989, p. 54).

Intensional provability relations link arithmetical theories to a given set of propositions when the arithmetical theory is able to prove each of the propositions in the set. That there cannot be a general theory of the kind Thomason specifies follows from an interesting result on the peculiarities of the intensional proof relation. It is a result of Feferman (1960) that Gödel's arithmetical formalization of the proposition that Peano arithmetic is consistent can be proved, under substitution of different linguistic expressions for the same classes of numbers in that arithmetical formalization.

PGA requires that "evidence", "acceptability", and "justified" can be arithmetized. We can formalize the evidence relation and the property of acceptance within computable learning theory, but this raises the question of whether that formalization captures all of the uses of these terms in inductive reasoning and if the terms can be arithmetized. What of the property of being justified? How would we axiomatize its basic features in the way that Kaplan and Montague axiomatized the basic features of knowledge? What happens to PGA if the notion of being justified is omitted? Without it, we cannot say that *P* tells us that we are prescriptively justified in believing an arithmetically false sentence. Thus we will not be able to show that an absurdity results if *P* converges upon either CL or the negation of CL. In which case, we cannot even express the condition of adequacy that is necessary for obtaining the contradiction.

Objection: It is true that omitting the notion of "justifies" in PGA blocks deriving the contradiction. But that is not a problem for the anti-functionalist end to which PGA is applied. You succumb to a dilemma if you argue there is no obvious arithmetization of "is justified". The first horn is that if there is an arithmetization of "is justified", then the contradiction is secured. For the second horn, suppose it cannot be arithmetized. If so, then it cannot be part of cognitive science. Thus, either way, cognitive science is in jeopardy. On the first horn, cognitive science cannot prove that it is correct and on the second horn, inductive reasoning can't be computationally described. On either horn, the anti-functionalist wins.

Response: The first horn of the dilemma is that if "is justified" is arithmetizable, then PGA is secured. Below we argue that even if PGA is sound, it cannot be used to secure the claim that human minds are not finitary computing machines or the claim that cognitive science cannot be justified. The second horn is easily dismissed, though. That "X" is not arithmetizable does not logically imply "X" is not formalizable. Why think any property or relation whatsoever, even though formalizable, can be arithmetized? Certainly, Gödel numbers can be assigned to formalized sentences and to formalized properties. But it does not follow from that fact that any formalized property is arithmetizable. The example of Principia ramified type theory, discussed above, illustrates the point. The burden of proof is upon Putnam, to show that the epistemic property of being justified, under a suitable formalization, can be arithmetized. (Artemov-Fitting logics of justification are not a method of reasoning to achieve justification, but a method for reasoning about justifications. An open question is whether a Montague-Kaplan type paradox could be constructed using their justification predicate.)

6.2 Strengthened PGA Leads to an Absurdity

One problem with PGA is that if not all inductive methods or, more broadly, methods of inquiry into the world, are subject to the Gödel incompleteness theorems, then it is possible that in using methods that are subject to the Gödel incompleteness theorems, we can employ weak inductive methods that are not subject to the Gödel incompleteness theorems to prove CON(method subject to the Gödel incompleteness theorems) or the Gödel sentence (of a method subject to the Gödel incompleteness theorems) in another epistemic modality or with mathematical certainty less than the degree of mathematical certainty of the proof procedure of the formal system in which the methods are formalized. Both human minds (that have or do not have a finitary computational description) and finitary computing machines that are subject to the Gödel incompleteness theorems can use weak methods that are not subject to the Gödel incompleteness theorems. Any EGF or MGF argument that ignores this possibility commits a logical error no less serious than the logical error Penrose commits in his anti-functionalist argument. On the other hand, if the above possibility is taken seriously, then EGF and MGF arguments can fail. What can be done? One suggestion is to show all methods of inquiry into the world are subject to the Gödel incompleteness theorems.

Suppose we strengthen PGA in the following way: all methods of inquiry into the world are subject to the Gödel incompleteness theorems. (Putnam appears to say this is how he wants his argument to be interpreted; see Putnam, 1988.) Such methods include all inductive methods, all demonstrative methods and all methods to which Putnam calls attention in (1988): rational interpretation, reasonable reasoning and general intelligence. Although he makes the strengthened PGA argument in (1994a), he alludes to it in:

> This is analogous to saying the true nature of rationality—or at least of human rationality—is given by some "functional organization", or computational description […]. But if the description is a formalization of our powers to reason rationally *in toto*—a description of all our means of reasoning—then inability to know something by the "methods formalized by the description" is inability to know that something in principle. (Putnam, 1988)

Strengthened PGA claims all inductive methods, all notions of epistemic justification, all methods of inquiry into the nature of the world are subject to the Gödel incompleteness theorems. The truth of (x) CON(method of inquiry$_x$) is essential to the soundness of PGA. If we can't prove (x) CON(method of inquiry$_x$), then we cannot show that strengthened PGA is sound. Why is that? If we can't prove (x) CON(method of inquiry$_x$), method of inquiry$_x$ might be inconsistent, in which case anything is provable. If so, we can't prove that the epistemic notions of "acceptance" and "justifies" are subject to the Gödel incompleteness theorems. Even if we can prove CON(method of inquiry$_i$) using method of inquiry$_j$ (a stronger extension of method of inquiry$_i$), if CON(method of inquiry$_j$) can't be proved, then it's possible that both CON(method of inquiry$_i$) and NOT-CON(method of inquiry$_i$) can be proved within method of inquiry$_j$. If each method of inquiry is subject to the Gödel incompleteness theorems, then no method of inquiry can be proved consistent. If no method of inquiry can be proved consistent, it is possible no method of inquiry is consistent.

I will now argue that strengthened PGA engenders an absurdity. Suppose that all methods of inquiry (such as statistical methods and methods that deliver proofs in another epistemic modality) are subject to the Gödel incompleteness theorems. That supposition would have as a consequence that all of our reasoning (in whatever method of inquiry that reasoning occurs) about the Gödel incompleteness theorems is subject to the Gödel incompleteness theorems. In which case, that reasoning might not be correct, and so that reasoning could not be epistemically justified. Why is that? For any method of reasoning, its' consistency cannot be proved. Thus it is left open that any method of reasoning might be inconsistent. Consider the following: for any chain of reasoning that establishes proposition *p*, it is possible there is another chain of reasoning that establishes not-*p*. This is so because it is possible that all methods of inquiry are inconsistent. If so, one could validly reason to p and one could validly reason to not-*p*, for any inconsistent method of inquiry. Thus, for all *p*, *p* cannot be epistemically justified, since for each *p*, one might validly infer not-*p* and *p*. This is an absurdity. Take this absurdity to be a reductio of the argument that all forms of reasoning are subject to the Gödel incompleteness theorems.

Given this absurdity, the most natural explanation of it is that one assumed that all methods of inquiry are subject to the Gödel incompleteness theorems. Give up that assumption, and the absurdity is removed. But giving up that assumption means there must exist some methods of inquiry that are not subject to the Gödel incompleteness theorems. If so, it is possible that any such method can prove CON(*P*) or CON(method of inquiry subject to the Gödel incompleteness theorems) with less than mathematical certainty or in some other epistemic modality. And if that is the case, then any finitary computational machine could also make such inferences. No cognitive difference would be registered between human minds and any finitary computational machine. (There might be significant cognitive differences between human minds and finitary computational machines which can compute functions that human minds cannot compute, owing to resource limitations, such as the length of time allowed for computing values of the function.)

7. A Fundamental Logical Problem for EGF and MGF

We now introduce a logical difficulty that arises in MGF and EGF arguments, how anti-functionalists might respond to it and whether Putnam can satisfactorily respond to it. We remark that a difficulty noticed by George Boolos (1986) will not be considered here. Boolos argued the Gödel disjunction (Gödel, 1995) is not derivable from the Gödel incompleteness theorems without first clarifying what it means for a human mind to be equivalent to a finite computing machine. What does it mean to assert that the human mind is equivalent to a Turing machine? We do not consider it here, because Nathan Salmon (2001) has convincingly argued the Gödel disjunction can be used to make philosophically interesting claims about the limitations of the human mind even if we do not have a precise description of what it is for human minds to be equivalent to Turing machines.

7.1. The Logical Problem Confronting EGF and MGF Arguments Is Recursively Unsolvable

The possibilities Kreisel (1972) notes for finitistically proving CON(**PA**) with less than mathematical certainty or in some other epistemic modality must be taken seriously by anti-functionalists who offer EGF or MGF arguments. Failure to take them into account is a logical error in EGF or MGF arguments. Why is that? Where a human agent can finitistically prove CON(*P*) with less than mathematical certainty or in some other epistemic modality—and that is how such human agents prove CON(**PA**) and the Gödel sentence for **PA**, so also might a finitary computational machine hypothesized to provide a computational description of human mentality. If so, neither MGF nor EGF arguments can distinguish human mentality from finitary computational machines. Failure to consider this possibility is a logical error in Lucas-Penrose-Putnam arguments. However, taking this possibility into account is a recursively unsolvable problem. The anti-functionalist is then faced with a dilemma: either the anti-functionalist fails to take into account Kreisel's way out, in which case they commit a logical error in their argument or else they do take it into account, in which case they must solve a recursively unsolvable problem.

The anti-functionalist might voice the following objection to the claim that they commit a logical error by failing to take into account Kreisel's way out: "The functionalist must find a specific method of inquiry or program that proves CON(*P*) with less than mathematical certainty or in another epistemic modality. The anti-functionalist is not required to find such a method. No logical error is committed by failing to consider the possibility of such a way out". This objection can be easily dismissed. The anti-functionalist makes the claim that a human mind not fully characterized by any finitary computational machine can determine the truth of CON(*P*). But it is possible that there is a method of inquiry or a program that can determine CON(*P*) with less than mathematical certainty or in some other epistemic modality. It is up to the anti-functionalist to dismiss that possibility. To dismiss it, the anti-functionalist must prove a negative existential claim: there is no such method of inquiry or program. It will be shown below that dismissing this possibility is a recursively unsolvable task.

Recall that Putnam's objection to Penrose's argument is that the program *P* might be so large that it cannot be humanly surveyed, and so no human could establish CON(*P*). Putnam only needs to cite the possibility that the program *P* is so large that no human could survey it. Since it is a possibility which, if true, would undermine Penrose's argument, Penrose must respond to it. It is not a legitimate argumentative move for Penrose to reply that Putnam must provide an actual *P* which cannot be humanly surveyed. The burden of proof is on Penrose—to show that the actual *P* can be humanly surveyed. Of course, *P* has yet to be written, since we do not now have a complete finitary computational description of human mentality (should there be one), so Penrose cannot counter Putnam. That is why Putnam's critique of Penrose's argument is so devastating.

Anti-functionalists who wish to avoid that logical error by taking these possibilities into account confront a computationally daunting task. Call that task "DISJUNCTION". It is the following: The anti-functionalist must show that either: (i) each method or program for mathematically or non-mathematically finitistically proving, with less than mathematical certainty or in some other epistemic modality, the consistency of *P* (the ultimate computer program that completely describes the human cognitive mind) is subject to the Gödel incompleteness theorems or (ii) if that cannot be done, because such a method or program is not subject to the Gödel incompleteness theorems, show that the proofs delivered by those methods or programs are not epistemically justified.

From DISJUNCTION there is a dilemma for anti-functionalists using EGF or MGF arguments:

First horn: The anti-functionalist must show, for each possible method or program capable of finitistically demonstrating the consistency of *P* with less than mathematical certainty or in some other epistemic modality, that it is either subject to the Gödel incompleteness theorems or that, where it is not subject to the Gödel incompleteness theorems, it is epistemologically inadequate.

Second horn: If the anti-functionalist does not enumerate all of these possibilities, a logical error is committed in their EGF or MGF argument.

DISJUNCTION has logical complexity Π(1,2). Suppose an anti-functionalist offers an MGF argument. In virtue of DISJUNCTION, they must be able to perform an infinitary computational task. If they have infinitely many resources, they will be able to complete the task. If not, then they will not. But if they do not complete the task, then they commit a logical error in MGF. Thus the anti-functionalist who uses an MGF argument must either have the capacity to make infinitary computations or else commits a logical error. But it is not known whether human beings do or do not have infinitary computational capacities.

The anti-functionalist must show that human beings can prove CON(*P*), but the machine for which *P* is its program cannot prove CON(*P*). Neither the human nor the machine can finitistically prove CON(*P*) with mathematical certainty in the program for *P*. So the anti-functionalist must finitistically prove CON(*P*) with less than mathematical certainty or in some other epistemic modality that is not available to the machine. To show these methods are not available to the machine, she must (according to DISJUNCTION) be able to make infinitary computations to canvass all of the possibilities for doing just that or else commit a logical error. But, once again, it is not known whether human beings do or do not have infinitary computational capacities.

7.2. DISJUNCTION is Π(1,2) in the analytic hierarchy

(a) The first disjunct in DISJUNCTION.

How many methods of reasoning are there for finitistically proving CON(*P*) with less than mathematical certainty or in some other epistemic modality? Since formal systems such as **PM**, **FOL**, and sentential logic prove their truths with mathematical certainty, and since the Gödel theorems tell us that we cannot finitistically establish CON(*P*) with mathematical certainty, those formal systems cannot be used. But probabilistic formal systems can deliver their truths with less than mathematical certainty.

For instance, assume we use a statistical method based on a Carnapian measure function to finitistically prove CON(*P*) with less than mathematical certainty. Then there are infinitely many possible methods that can be used, since Carnapian inductive logics employ a caution parameter that has infinitely many values and which differentiates different logics (Carnap, 1952; 1962). (This establishes an existence proof that there are infinitely many inductive methods. For recent work on new probabilistic proof methods in randomness and computation, see Wigderson, 2019) How many different systems of formal inductive reasoning are there? How many probabilistic logics are there? How many hybrid modal probabilistic logics? Thus far we have the following computational problem: (i) Look at each method for finitistically proving CON(*P*) with less than mathematical certainty or in some other epistemic modality. (ii) Show it is subject to the Gödel incompleteness theorems.

What of proving some proposition in an epistemic modality other than that of mathematical certainty? For instance, philosophical nonmathematical reasoning that cannot be translated into first-order logic might be an example. One problem, though, is that Hilbert's thesis that any argument can be translated into first-order logic makes it difficult to claim that there is reasoning in a natural language that cannot be captured in first-order logic.

There are infinitely many applicable methods of reasoning with less than mathematical certainty or in another epistemic modality. Each of them must be enumerated and checked for being subject to the Gödel incompleteness theorems. And there is an additional regress-like wrinkle. It is the following. Suppose a program P^* proves CON(*P*) with less than mathematical certainty or in some other epistemic modality. The anti-functionalist needs to verify that P^* is subject to the Gödel incompleteness theorems. (If not, then neither an MGF nor an EGF argument can be deployed.)

The wrinkle is that even if P^* is subject to the Gödel incompleteness theorems, there might be a program P^{**} that can be used to mathematically and finitistically prove CON(P^*) with less than mathematical certainty or in some other epistemic modality. Suppose that P^{**} is shown to be subject to the Gödel incompleteness theorems. If so, there is a possibility there is a program P^{***} that can be used to mathematically prove CON(P^{**}) with less than mathematical certainty or in some other epistemic modality. So we have the possibility of an infinite regress for each program or method for proving CON(P^*) and its star relatives that we have shown to be subject to the Gödel incompleteness theorems.

The procedure then, is the following. Look at each method$_{1,i}$ for proving CON(P) with less than mathematical certainty or in some other epistemic modality. Show it is subject to the Gödel incompleteness theorems. If it is, look at each method$_{2,j}$ for proving CON("method$_{1,i}$") with less than mathematical certainty or in some other epistemic modality. Show it is subject to the Gödel incompleteness theorems. If it is, look at each method$_{3,k}$ for proving CON("method$_{2,j}$") with less than mathematical certainty or in some other epistemic modality. Show it is subject to the Gödel incompleteness theorems. Continue in this way ad infinitum.

Let's consider an objection the anti-functionalist might raise to the specter of the infinite regress. She tells us that there will be no infinite regress, because of her dialectical situation in EGF or MGF arguments. Whenever computational functionalists propose a method *M*, all she has to do is to show *M* is subject to the Gödel incompleteness theorems. She plays a waiting game. She waits for the computational functionalist to propose a method, and only then does she need to show that the proposed method is subject to the Gödel incompleteness theorems (Lewis, 1969; 1979; Lucas, 1961; 1970).

This objection fails, for two reasons. The first is that methods of proof that prove a theorem with less than mathematical certainty or in some other epistemic modality are methods of proof that will be used to prove the consistency of the methods for proving CON(P) that are susceptible to the Gödel incompleteness theorems. So we are still considering a specific machine *M* and not any other machine, *M'*. The anti-functionalist does not, contra J. R. Lucas, play a wait and see game with the computational functionalist.

Second, all MGF and EGF arguments are responsible to certain epistemic standards: if there are any relevant possibilities that undermine the arguments, they must be examined. If it is possible there is a method or program *P* not subject to the Gödel incompleteness theorems that finitistically proves CON(M) with less than mathematical certainty or in some other epistemic modality, then that undermining possibility must be discharged.

The anti-functionalist implicitly makes a negative existence claim in EGF and MGF arguments: there is no method or program subject to the Gödel incompleteness theorems by which CON(P) can be finitistically shown correct with less than mathematical certainty or in some other epistemic modality. Since there are infinitely many possibilities for finitistically proving CON(P) with less than mathematical certainty or in some other epistemic modality, each of them must be taken into account. If not then the negative existence claim fails.

7.3. How Program Length Contributes to the Complexity of DISJUNCTION

Suppose that *P* is so long it can't be surveyed by any human agent, whether they are finitistically computationally describable or not. If that is the case, we will not know if there are any programs or methods that can be used to prove CON(P) with less than mathematical certainty or in some other epistemic modality. But there might be ways of compressing the length of P so that we can then determine if there are methods that can be used to prove CON(P). One way of doing this is to reduce P to some program P^* that is humanly surveyable. (One then looks at methods for proving CON(P^*) with less than mathematical certainty or in some other epistemic modality.) There are three ways in which this can be done. One method is by a relative interpretation of P in P^*, another is by a translation of P into P^* and the third is a reduction of P to P^*. There are logical differences between interpretations, translations and reductions, which are the subject of reductive proof theory. What is common to all three is that the map from P into P^* is recursive and preserves negation. The latter condition ensures that logical consistency is preserved under the map.

The maps between P and P^* preserve consistency, provided P^* is consistent. Since the assumption is that P is consistent, we need to find a short and consistent P^*. Suppose P^* is not feasibly short. It is possible there is a P^{**} that is consistent and feasibly short to which P^* can be reduced or translated or into which it can be interpreted. At each level of reduction for which there is a consistent and infeasibly long P^{n*}, it is possible that in a reduction to the next level, by either a translation, reduction or interpretation, there is a consistent and feasibly short $P^{(n+1)*}$.

To avoid in EGF and MGF arguments the logical error committed by Penrose, we have to consider the possibility that P is infeasibly long and then to consider how it might be compressed. The possibility of an infinite chain of reductions of length omega is a prospect that cannot be *a priori* ruled out. (The chain length could be omega, since a reduction might not decrease the length of P^{n*}.) There are also other methods that can compress P. For instance, P could be translated into another programming language in which compression devices called MACROS are available or other higher-order programming constructs that facilitate program compression. There are infinitely many different programming systems, so there are that many possibilities that might need examination in the search for a feasibly short P. There are also speed-up theorems in the theory of computability that tell us there's no recursive bound on the speed-up of some programs (over the initial program for which there is speed-up).

The anti-functionalist can object to the preceding infinite regress generated by program compression considerations in the same way she objected to the first infinite regress above: “The computationalist must first present to me a feasibly short P. Once that is done, we can then see if there are methods or programs that finitistically prove CON(P) with less than mathematical certainty or in some other epistemic modality”. Once again, the anti-functionalist misconceives of her epistemic situation in the anti-functionalism dialectic. If it is possible that there is a feasibly short P, then she must examine the possibilities under which it can be obtained. Many of these possibilities (such as relative interpretability) might be dead-ends, might generate infinite regresses or might create trade-off problems.

7.4. The First Disjunct of DISJUNCTION is Π(1,2) in the Analytic Hierarchy

We first noted that there might be infinitely many distinct methods for finitistically proving CON(*P*) with less than mathematical certainty or in some other epistemic modality. For each such method, the anti-functionalist must show either that it is subject to the Gödel incompleteness theorems or that it is not epistemically justified. We then noted that for each method *M* that proves CON(*P*) and is shown subject to the Gödel incompleteness theorems, there might be a method M^* that proves CON(method *M*) with less than mathematical certainty or in some other epistemic modality. If so, the anti-functionalist must show method M^* is subject to the Gödel incompleteness theorems. In general, for each *M* that is shown subject to the Gödel incompleteness theorems, there might be an M^* that proves its correctness for which it must be shown it is subject to the Gödel incompleteness theorems. After that, we saw that if *P* (or any of the methods or any of the M^*'s) is infeasibly long, we need to see if we can compress it to obtain a feasibly short *P* (or short M^*, etc.) Each of these feasibly short *M*'s must then be shown to be subject to the Gödel incompleteness theorems.

There are infinitely many methods of reasoning that might finitistically prove CON(*P*) with less than mathematical certainty or in some other epistemic modality. For each method M_i subject to the Gödel incompleteness theorems, it is possible there is a method or program that finitistically proves CON(M_i) with less than mathematical certainty or in some other epistemic modality. Let M_j be the method that finitistically proves CON(M_i), where $i \neq j$. If M_i is subject to the Gödel incompleteness theorems, then there might be an M_k ($i \neq j \neq k$) that finitistically proves CON(M_i) and which must then be shown by the anti-functionalist to be subject to the Gödel incompleteness theorems. For each of the infinitely many M_i's, there are infinitely many M_i^{n*}'s. Finally, for every M_i and M_i^{n*}, it is possible it is infeasibly long and thus we need to look for a compression of it into a feasibly short program. But for each M_i and M_i^{n*}, there might be an infinite sequence of compression reductions R_i.

Each method or procedure can be considered to be a function from the natural numbers to natural numbers. Determining that a method or procedure is or is not subject to the Gödel incompleteness theorems is a recursive predicate. The predicate is applied to each method or procedure, of which there are infinitely many. So there is a quantifier over the set of methods and procedures—it is a function quantifier. For all such methods or procedures, it is possible there exists a method or procedure not subject to the Gödel incompleteness theorems which verifies its consistency with less than mathematical certainty or in some other epistemic modality. (M_x) ($\exists M_y$) (M_y is not subject to the Gödel incompleteness theorems AND M_y proves CON(M_x) with less than mathematical certainty or in some other epistemic modality). In the analytic hierarchy, this sentence has logical complexity Π(1,2).

(b) The second disjunct in DISJUNCTION

Recall the second disjunct in DISJUNCTION: If a method or program for proving CON(*P*) with less than mathematical certainty or in some other epistemic modality is not subject to the Gödel incompleteness theorems, then show that the proofs delivered by that method or program are not epistemically justified. The anti-functionalist must show that for each method or program examined by the procedure described in the first disjunct of DISJUNCTION that is not subject to the Gödel incompleteness theorems, it is not epistemically justified. This must be done to save any EGF or any MGF argument. Suppose the anti-functionalist argument is an EGF argument. The claim is: *P* cannot be proved correct because it is subject to the Gödel incompleteness theorems (and thus cognitive science cannot be justified). But there might be other ways to prove CON(*P*) with less than mathematical certainty or in another epistemic modality. If those ways are subject to the Gödel incompleteness theorems, the claim remains intact. If any of those ways are not subject to the Gödel incompleteness theorems, they prima facie refute the claim. The only way to save the claim is to show that the methods or programs not subject to the Gödel incompleteness theorems are not epistemically justified. That is, proofs delivered by those methods or programs are not epistemically warranted.

Since any method or any procedure might not be subject to the Gödel incompleteness theorems, then every subset of the infinite methods tree might need to be tested for epistemic adequacy—that it is epistemically justified. Of course, no point in the infinite methods tree might need to be tested, if every point represents a method or program that is subject to the Gödel incompleteness theorems.

How we can show that a method or program is not epistemically justified? If what is proved by a method has a 50% chance of being true, we can conclude the method is not justified. However, what do we say when the probability of being true is greater than ½? What is the cut-off point? What if we do not have sufficient statistics for showing the likelihood of what a method proves? What epistemological theory do we employ in assessing epistemic justification of a method? Even if we are guided by statistical methods used in the sciences, those methods still make philosophical presuppositions about the nature of probabilities.

Suppose that a method uses nonmathematical philosophical reasoning (Kreisel, 1972) that contains no quantitative information necessary for obtaining probabilities. How do we assess these methods for epistemic justification? Is the epistemic justification of a quantitative method different in kind from the epistemic justification of a non-quantitative method? What does it mean to say we search the space of epistemologies for various construals of epistemic justification (Audi, 1988; Lehrer, 1990)? Given that EGF and MGF arguments are philosophical arguments claiming to refute a philosophical position in the philosophy of mind, any elucidation of the notion "epistemic justification of *P* (for any *P*)" must be philosophically respectable. If the philosophical construal of "epistemic justification of *P* (for any *P*)" is not philosophically respectable, the anti-functionalist will not be able to satisfy the second disjunct of DISJUNCTION.

These issues concerning epistemic justification are critical problems for the anti-functionalist. Without establishing that methods or procedures not subject to the Gödel incompleteness theorems are not epistemically justified, EGF and MGF arguments fail. The anti-functionalist must be prepared to decide what counts as epistemic justification of the correctness of *P* (for any *P*), and so what counts as the epistemic justification of cognitive science. Being able to assess the epistemic justification of methods that prove CON(*P*) with less than mathematical certainty or in another epistemic modality is a necessary condition for the success of EGF and MGF arguments. An important philosophical project, then, is elucidation of the notion "epistemic justification of proofs of CON(*P*) with less than mathematical certainty or in another epistemic modality".

7.5. Chains and Tangled Chains in the Methods Tree Exhibiting Defeater Relations

Suppose that a method or procedure is not subject to the Gödel incompleteness theorems and that it is not epistemically justified. Does it follow it can be dismissed by the anti-functionalist? No, for this method might epistemically justify CON(M^*), where method M^* is not subject to the Gödel incompleteness theorems and epistemically justifies CON(*P*). This may happen if we allow relative interpretations, translations and reductions between *P*, the method and M^*. But it can happen even if these relations do not occur. There might be chains in the methods tree, of arbitrary length, in which a method that does not epistemically justify *P* epistemically justifies a method which epistemically justifies *P*. Such chains can be of arbitrary length. Each of these chains must be examined by the anti-functionalist. It is well-known is epistemology that justification of a proposition can be defeated and can be restored after defeat, given the appropriate conditions (Pollock, 1999). The same can happen with methods for proving CON(*P*).

For example, suppose we have a chain in the methods tree of length 1,000 in which the 1,000th element in the chain is not subject to the Gödel incompleteness theorems. It is a method that does proves CON(*P*) with less than mathematical certainty or in some other epistemic modality, but is not epistemically justified when considered in isolation from all of the other methods in the chain. However, the 529th method in the chain epistemically justifies the 530th method in the chain, which, in turn, epistemically justifies the 531st method in the chain. This continues, until the 1000th element in the chain is epistemically justified.

Even if the *n*th method in a chain is not epistemically justified by the *n*-1st method in that chain (where the two methods are consider in isolation from all other methods), it does not follow the anti-functionalist can dismiss it, since there might be chains, of arbitrary length starting with the *n*-*k*th method, between the *n*-1st and *n*-*k*th methods, which transmit epistemic justification in such a way that the *n*-1st method is epistemically justified, and in consequence of this, is able to epistemically justify the n^{th} method. Additionally, one method in a chain might defeat epistemic justification of another method in the chain. If a chain of methods is finitely long, the power set of that chain consists of all subsets of methods which might need to be considered by the anti-functionalist. If a chain is infinitely long (because there are infinitely many methods or programs), then all possible chains that can be built with those methods or programs will have the power set of that infinitely long chain, and need to be considered by the anti-functionalist.

An additional complication in building such chains is the existence of methods or programs that defeat epistemic justification of CON(P) or of CON(M_i). Moreover, those methods or programs might not be formalized or even formalizable—suppose they are instances of what Kreisel means by "nonmathematical philosophical reasoning". Justification can be achieved by many different forms of reasoning. If your aim is to show that some proposition is not justified, then you must consider all of the ways in which it could be justified.

Suppose that M_k defeats justification of M_i, and M_i can prove correct CON(P) with less than mathematical certainty or in another epistemic modality. However, there might be a method or program M_{k-i} that defeats justification of M_k, thus restoring M_i so that it can prove correct CON(P). Call this a tangled chain of methods or programs. Notice this problem is similar to the logical problem facing defeater epistemologies (Pollock, 1999). There might be chains of defeaters, of arbitrary length, in which the 999th member of the chain defeats the 347th member of the chain, while the 876th member of the chain defeats the 999th. Simply enumerating and individually assessing each element in the chain is not enough. Each element in the chain must be evaluated for justificatory relations with every other sequence of elements in the chain.

Although formalizable methods or procedures can be considered to be functions over the natural numbers, I am less confident about methods or procedures for, say, nonmathematical philosophical reasoning. Perhaps they can be formalized and considered to be functions over the natural numbers. But the relation of one method justifying another might not be recursive, and might not even be formalizable. So it might be that no logical complexity measure can be assigned to the second disjunct of DISJUNCTION.

We have the following results:

(i) It is possible there are epistemically justified methods or programs which prove, with less than mathematical certainty or in some other epistemic modality, CON(P). EGF arguments must show there are no methods which can do that. If not, the conclusion of the EGF argument—that cognitive science cannot be demonstrated to be a correct theory—fails. EGF arguments assume human minds have a finitary computational description. Showing there are no epistemically justified methods or programs which can prove CON(P) with less than mathematical certainty or in some other epistemic modality is recursively unsolvable. Finitary human minds that have a finitary computational description cannot complete this task. If human minds have a metarecursive computational structure, they might be able to complete the task. But we do not know if human minds have a metarecursive computational structure.

(ii) To save the MGF conclusion that there is a cognitive task human minds can do that finitary computing machines can't, it must be shown either (a) that human minds can prove CON(*P*) with mathematical certainty or (b) that there is no epistemically justified method or program by which CON(*P*) can be proved, with less than mathematical certainty or in some other epistemic modality. Since only an infinitary mind can prove CON(*P*) with mathematical certainty (and only if mathematical certainty can be defined for an infinitistic system of reasoning), (a) has no empirical basis in cognitive science. There is no empirical evidence that human minds can perform infinitary tasks, such as constructing infinite proof trees. EGF arguments must establish (b), and we saw they cannot do so, because it is a recursively unsolvable task. It is a mystery how a human mind, even one that has no finitary computational description, could complete the task (unless it has a metarecursive computational structure, but we do not know whether this is so.)

8. A Categorization of Anti-Functionalist Arguments Using the Gödel Incompleteness Theorems Into Sixteen Cases

There are sixteen cases that are determined by partitioning anti-functionalist arguments into (i) epistemic and metaphysical uses of the Gödel incompleteness theorems—that is, EGF and MGF arguments, (ii) Penrose error cases (infeasibly long programs), and (iii) showing some, but not all weak inductive methods, are subject to the Gödel incompleteness theorems (PGA) and showing that all methods of inquiry into the world (i.e., all inductive methods) are subject to the Gödel incompleteness theorems (strengthened PGA).

There are eight cases when PGA or strengthened PGA succeeds. There are an additional eight cases when PGA or strengthened PGA fails. (We contend they both fail.) What is surprising is that even if PGA or strengthened PGA succeeds, the anti-functionalist acquires virtually no advantage over the computational functionalist in anti-functionalism arguments. It's important to note that in all MGF cases it is not assumed that human minds are finitary, nor is it assumed that they are infinitary. If human minds are infinitary and have a metarecursive structure, should we consider them to have a computational description analogous to finite minds with a computational structure? If human minds are infinitary and do not have a metarecursive structure, we should not consider them to have a computational description. But it is unknown whether human minds are or not infinitary. Similarly, although some cognitive scientists and philosophers believe human minds are finitary and can be described computationally, it is not known whether they are finitary.

In the first kind of EGF argument, it is assumed human minds are finite. Not so for the second kind of EGF argument (see Section 4.2.2 above). However, the second kind of EGF argument shows that metaphysical claims established by

MGF arguments are epistemically justified. Thus MGF arguments need to be categorized—the second kind of EGF argument does not. The phrase "EGF arguments" below refers to the first kind of EGF argument.

8.1. The First Eight Cases: PGA and Strengthened PGA Succeed

A successful PGA shows that some, though not all, weak methods of inquiry are subject to the Gödel incompleteness theorems. The first four cases cover a successful PGA. There are two cases for an EGF refutation of functionalism and two cases for an MGF refutation of functionalism. The two cases for each are when the computational description *P* is feasibly short and when it is infeasibly long.

Case (i): Recall that EGF arguments assume human minds are fully characterized by a finitary computational description. Suppose *P* is feasibly short. Since not all weak inductive methods have been shown to be subject to the Gödel incompleteness theorems, there may be weak methods that prove CON(*P*) with less than mathematical certainty or in another epistemic modality. If so, an EGF argument fails, since it is the point of an EGF argument to show that human minds cannot justify the finitary computational description *P* of themselves. That is, there isn't a proof of CON(*P*) that is epistemically justified. But a weak method might provide such a proof.

Case (ii): Suppose an EGF argument and that *P* is infeasibly long. Since not all weak methods have been shown to be subject to the Gödel incompleteness theorems, use weak methods to perform a statistical analysis to recover the full size of *P* from the fragments available. Then use weak methods to establish CON(*P*), with less than mathematical certainty. The EGF argument fails, for the same reasons in case (i).

Case (iii): Assume an MGF argument. Recall that MGF arguments show human minds do not have a finitary computational description, and argue that human minds are metaphysically different from finitary computing machines, since there are cognitive activities we can perform, that finitary computing machines cannot perform. Suppose that *P* is feasibly short. Even if human minds do not have a finitary computational description, we cannot use weak inductive methods or programs subject to the Gödel incompleteness theorems to establish CON(*P*), in the epistemic modality of the proof procedures of the weak methods. We can only use weak methods or programs not subject to the Gödel incompleteness theorems to establish CON(*P*) with less than mathematical certainty or in another epistemic modality. However, finitary computing machines can do the same thing, so we can't establish a metaphysical difference between them and human minds. The MGF argument fails.

Case (iv): Assume an MGF argument and that *P* is infeasibly long. Even if human minds do not have a finitary computational description, we cannot use weak inductive methods subject to the Gödel incompleteness theorems to do a statistical analysis of the fragments of *P* and recover *P* from that analysis and then prove CON(*P*). We can only use weak methods or programs not subject to the Gödel incompleteness theorems to do this. However, so can finitary computing machines. Once again, there is no metaphysical difference which we can establish between them and finitary human minds. The MGF argument fails.

Now we look at the four cases in which strengthened PGA succeeds. Recall that strengthened PGA shows that all methods of inquiry into the structure of the world (i.e., all inductive methods) are subject to the Gödel incompleteness theorems. The four cases are analogous to the four cases for PGA.

Case (v): Assume an EGF argument and that *P* is feasibly short. If so, then there are no weak methods or programs that can be used to show CON(*P*). In that case, the EGF argument succeeds, since we have shown that a human mind with a computational description *P* cannot justify *P*.

Case (vi): Assume an EGF argument and that *P* is infeasibly long. Since there are no weak methods or programs available for a statistical analysis of fragments of *P* to recover *P*, nor for showing CON(*P*), it follows that the EGF refutation succeeds. We have shown that a human mind with a finitary computational description *P* cannot justify *P*.

Case (vii): Assume an MGF argument and that *P* is feasibly short. There are no weak methods that can be used to show CON(*P*). In which case, even human minds with no finitary computational description will not be able to justify *P*. However, finitary computing machines cannot do this either. In which case, there is no discernible metaphysical difference (concerning computability) between human minds with no finitary computational description and finitary computing machines. Hence, the MGF argument fails.

Case (viii): Assume an MGF argument and that *P* is infeasibly long. There are no weak methods or programs that can be used to perform a statistical analysis on a fragment of *P* and recover *P*, nor to show CON(*P*). In which case, even human minds with no finitary computational description will not be able to justify *P*. However, finitary computing machines cannot do this either. In which case, there is no discernible metaphysical difference (concerning computability) between human minds with no finitary computational description and finitary computing machines. Hence, the MGF argument fails.

These analyses reveal an interesting truth. It is that all MGF arguments fail, even though either PGA or strengthened PGA succeeds. On the other hand, though EGF arguments fail even where PGA succeeds, EGF arguments succeed where strengthened PGA succeeds. Thus there is a critical philosophical difference between MGF and EGF arguments.

Note that if it is demonstrated that human minds are able to construct infinite proof trees and do not have a metarecursive structure that allows for a computational description analogous to a computational description of finitary minds, then all MGF arguments will succeed wherever PGA and strengthened PGA succeed. Using the Gödel theorems to refute functionalism by an MGF argument can only succeed if it is a fact (and known to us) that human minds can construct infinite proof trees, but have no metarecursive structure that allows for a computational description analogous to a computational description of finitary minds. If that cannot be demonstrated, then even though PGA or strengthened PGA succeeds, no MGF argument can succeed.

8.2. The Second Set of Cases: PGA and Strengthened PGA Fail

We now look at the same kinds of cases, under the assumption that PGA and strengthened PGA fail (in the way in which I have argued they fail). Cases ix–xii will happen when PGA fails. That is, PGA fails to show that some weak inductive methods are subject to the Gödel incompleteness theorems.

Case (ix): Suppose an EGF argument and that *P* is feasibly short. Since it has not been shown that any weak methods are subject to the Gödel incompleteness theorems, all weak methods are available for proving CON(*P*), with less than mathematical certainty. So a human mind that has a finitary computational description can prove *P* is correct (i.e., justify *P*). Since there are more weak methods available for proving CON(*P*) with less than mathematical certainty, and in other epistemic modalities, than there are when PGA succeeds, EGF arguments fail more often when PGA fails than they do when PGA succeeds.

Case (x): Suppose an EGF argument and *P* is infeasibly long. Since it has not been shown that any weak inductive methods are subject to the Gödel incompleteness theorems, all weak inductive methods are available for statistically recovering P and proving CON(*P*). So a human mind that has a finitary computational description can epistemically justify *P*. Since there are more weak methods available for recovery of *P* and proof of CON(*P*), and in other epistemic modalities, than there are when PGA succeeds, EGF arguments fail more often when PGA fails than they do when PGA succeeds.

Case (xi): Suppose an MGF argument and that *P* is feasibly short. Although all weak inductive methods are available for proving CON(*P*) with less than mathematical certainty, all of these methods are also available to finitary computing machines. In which case, there is no means of discerning a metaphysical difference (concerning computability) between human minds with no finitary computational description and finitary computing machines. MGF arguments fail when strengthened PGA fails, but no worse (or no better) than they failed when PGA succeeded.

Case (xii): Suppose an MGF argument and *P* is infeasibly long. Although all weak inductive methods are available for statistically recovering *P* and for proving CON(*P*) with less than mathematical certainty, all of these methods are available to the finitary computing machine. In which case, there are no means of discerning a metaphysical difference (concerning computability) between human minds with no finitary computational description and finitary computing machines. MGF arguments fail when strengthened PGA fails, but no worse (or no better) than they did when PGA succeeded.

Now we examine the four cases when strengthened PGA fails, because of the absurdity to which it succumbs. Recall the absurdity: *P* encompasses all of the epistemically adequate weak methods *M* of inquiry into the world that could prove CON(*P*) with less than mathematical certainty or in another epistemic modality. Suppose that all methods of inquiry are subject to the Gödel incompleteness theorems. For each method *M*, we cannot prove that it is consistent. So it is possible that each method *M* is inconsistent. For any chain of reasoning that establishes proposition *p*, it is possible there is another chain of reasoning that establishes not-*p*. One could validly reason to p and one could validly reason to not-*p*, for any inconsistent method of inquiry. Thus, for all *p*, *p* cannot be epistemically justified, since for each *p*, one might validly infer not-*p* and validly infer *p*. This is an absurdity. Take this absurdity to be a reductio of the argument that all forms of reasoning are subject to the Gödel incompleteness theorems.

Case (xiii): Suppose an EGF argument and *P* is feasibly short. The reasoning is exactly the same as it is for case (ix). All weak methods are available for proving CON(*P*) with less than mathematical certainty or in another epistemic modality. So a human mind that has a finitary computational description can epistemically justify *P*. Since there are more weak methods available for proving CON(*P*) with less than mathematical certainty, and in other epistemic modalities, than there are when PGA succeeds, EGF arguments fail more often when strengthened PGA fails than they do when strengthened PGA succeeds.

Case (xiv): Suppose an EGF argument and *P* is infeasibly long. The reasoning is exactly the same as it is for case (x). All weak methods are available for statistically recovering and proving CON(*P*) with less than mathematical certainty. So a human mind that has a finitary computational description can epistemically justify *P*. Since there are more weak methods available for recovery of *P* and proof of CON(*P*), and in other epistemic modalities, than there are when PGA succeeds, EGF arguments fail more often when strengthened PGA fails than they do when strengthened PGA succeeds.

Case (xv): Suppose a MGF argument and *P* is feasibly short. The reasoning is exactly the same as it is for case (xi). All weak methods are available for proving CON(*P*) with less than mathematical certainty or in some other epistemic modality to finitary computing machines and human minds with no finitary computational description. In which case, there is no means of discerning a metaphysical difference (concerning computability) between human minds with no computational description and finitary computing machines. MGF arguments fail when strengthened PGA fails, but no worse (or no better) than they did when strengthened PGA succeeded.

Case (xvi): Suppose a MGF argument and *P* is infeasibly long. The reasoning is exactly the same as it is for case (xii). All weak methods are available for statistically recovering *P* and for proving CON(*P*), with less than mathematical certainty or in another epistemic modality, to finitary computing machines and human minds with no finitary computational description. In which case, there is no means of discerning a metaphysical difference (concerning computability) between human minds with no finitary computational description and finitary computing machines. MGF arguments fail when strengthened PGA fails, but no worse (or no better) than they did when strengthened PGA succeeded.

That concludes the categorization of cases under PGA and strengthened PGA, where they succeed and where they fail. Do we have any reason to believe that *P* will be infeasibly long? Now, we have no such reason. We do not know what ultimate cognitive science will look like, so we do not know, now, whether in ultimate cognitive science the ultimate program *P* will be infeasibly long. We do not have a theory of feasible computability that will tell us whether programs that have outputs of certain kinds are feasibly short. We do not know if human minds can be completely described computationally. We do not know if there is an ultimate cognitive science.

9. Twelve Objections to the Absurdity Engendered by Strengthened PGA

There are several anti-functionalist objections to the absurdity that threatens to destroy PGA and strengthened PGA and thus threatens to destroy EGF and MGF arguments. I enumerate and respond to them below.

Objection 1: Even if *P* is infeasibly long, human minds can epistemically justify *P*, though no finite computing machine (which *P* formally characterizes) can. Since all epistemically adequate weak methods of inquiry into the world—including any that confer empirical justification upon CON(*P*)—are, by PGA, subject to the Gödel incompleteness theorems, no finite computing machine formally characterized by *P* can employ those methods to prove, with less than mathematical certainty or in another epistemic modality, CON(*P*). However, human minds can do that, since statistical methods fall under the weak methods subsumed by *P* and statistical methods are employed where human minds face resource limitations or do not have all the facts. The burden of proof is on the shoulders of the functionalist, to show that for programs greater than length L no statistical method subsumed under *P* can empirically justify CON(*P*).

Response: It is true that no finitary computing machine formally characterized by *P* can use the statistical methods subsumable under *P*, provided that strengthened PGA succeeds. But human minds, even under the assumption they have no finitary computational description, are similarly forbidden. If all formalized statistical methods are shown by strengthened PGA to be subject to the Gödel incompleteness theorems, then no finitary human mind can use them to recover *P* and then prove CON(*P*).

Objection 2: Finitary human mind can empirically justify CON(*P*) by reducing its consistency problem to a consistency problem for a formal system that does not subsume any of the weak methods of inquiry into the world that are subsumed by *P*. We then use weak methods to prove, with less than mathematical certainty, CON(REDUCING FORMAL SYSTEM) and use the reduction to conclude CON(*P*).

Response: If *P* subsumes all methods of inquiry into the world, then any formal system that does not subsume them is probably not a formal system to which *P* can be reduced. Suppose that, for the sake of argument, it is. Reductive proof theory requires there is a recursive function that maps every proof in the reduced system to a proof in the reducing system. Moreover, this mapping must itself be provable in a formal system that is, in general, included in the reducing system. When these conditions are satisfied, the reducing system will be a conservative extension of the reduced system. There is nothing in the reduced system that cannot be proved in the reducing system and, more importantly, there is nothing in the language of the reduced system that can be proved in the reducing system, though not proved in the reduced system. In other words, for any proof in PGA that any epistemically adequate weak method in *P* is subject to the Gödel incompleteness theorems, there will be a corresponding proof in the reducing system that whatever is the analogue of the weak method in *P* is subject to the Gödel incompleteness theorems.

Objection 3: If *P* is infeasibly long, it fails as an explanatory theory in cognitive science. Any finitary computational description we can't follow is one that can't be explanatory for us. Thus, under the assumption human beings have no finitary computational description that characterizes their complete mentality, an infeasibly long *P* secures for anti-functionalists the conclusion that cognitive science is not justified. If a scientific theory has no explanatory value, it loses epistemic justification.

Response: This objection does not advance the anti-functionalist even one square forward in the functionalism debate. If it turns out that *P* is infeasibly long, then human beings might never discover it. What we do discover will be an approximation to *P* which we do find explanatory and that is not infeasibly long. The objection which the anti-functionalist just voiced is really a skeptical objection, and it is one which could be voiced in any scientific discipline whatsoever. The anti-physicalists can say that the ultimate theory of physics is super-long and thus has no explanatory value. The same response to the anti-functionalist holds here as well. Yes—it is a worry, but no—it is not a worry that gives the anti-functionalist any advantage, for it is a general skeptical worry.

Objection 4: Let's try to refine the preceding objection. Genuine warranted assertibility and epistemic justification have no finitary computational description. These methods, because they are not formalizable, are not subject to the Gödel incompleteness theorems. Finitary human minds—under the assumption they have no finitary computational description—can use these methods to produce a proof of CON(*P*). So there is something a finitary human mind can do that no finitary computing machine can do.

Response: This is a confused objection. How can methods resisting formalization be used to prove the correctness of a formal system? Strengthened PGA shows that all epistemically adequate weak methods are subject to the Gödel incompleteness theorems. Thus it shows that all epistemically adequate weak methods have no complete finitary computational description. But if strengthened PGA fails, it is left open that there are formalizable epistemically adequate weak methods that can prove, with less than mathematical certainty or in another epistemic modality, CON(*P*). If strengthened PGA fails, then the anti-functionalist must compute the solution to a recursively unsolvable problem, in order to show that there are no epistemically adequate weak methods that are not subject to the Gödel incompleteness theorems. The point is that the only way we have of showing that there is no complete finitary computational description of X is by using a Gödelian argument. Strengthened PGA is such a Gödelian argument, but it fails.

Objection 5: The absurdity is a travesty of mathematical reasoning. If you are right, then you have shown that the Gödel theorems in their original context—proving the incompleteness of Peano arithmetic and the unprovability of CON(**PA**)—fail to work. One can run your absurdity argument on the provability predicate and easily reach the absurd conclusion that there is no unprovable sentence in Peano arithmetic. You would have shown that Gödel is wrong. Since that is too absurd to consider, we must conclude that you are wrong!

Response: That is an important objection However, you did not think very clearly about the matter at hand. The provability predicate is not defined by Peano arithmetic. We have independent reasons for believing in its cogency and we could construct it even if Peano arithmetic did not exist. What we are able to do in Peano arithmetic is to arithmetize it and then employ the diagonal lemma to secure the incompleteness theorems.

The situation is quite different when it comes to program *P*—the computational description of our methods of inquiry into the world. Recall that in PGA the analogue of the notion of "proof" for Peano arithmetic is the notion of "justifies". However, *P* defines the notion of justification. If there were no *P*, there would be no notion of justification. If it turns out that the notion of justification cannot itself be justified—and that is exactly what PGA attempts to show—then we have no coherent notion of justification. If there are truths about justification we are forbidden from justifying, the notion is incoherent. In which case, we can't appeal to the Montague-Kaplan-Thomason axioms for axiomatizing "justifies" so that it can meaningfully satisfy the Gödel diagonal lemma, since we have no reason to think that these axioms applied to "justifies" are true. On the other hand, we do have independent reasons for thinking that the Hilbert derivability conditions for the provability predicate are true, independently of the question of the consistency of Peano arithmetic.

Objection 6: You cannot be serious that human minds with no finitary computational description have no epistemic advantages over finitary computing machines. Can't a human mind with no finitary computational description survey an infeasibly long *P*? If not, then what could possibly be the difference between the human minds and finitary computing machines? Are you proposing that they are identical?

Response: No, we are not. But just because a human mind has no finitary computational description does not entail it is able to construct infinite proof trees or that it has the computational resources to survey an infeasibly long *P*. Human minds that have no finitary computational description might not have any epistemic advantages over finitary computing machines. Even infinitary agents cannot prove the consistency of Peano arithmetic using a finitary and effective proof, since finitary and effective proofs of it are prohibited by Gödel's incompleteness theorems. If all weak methods for proving CON(**PA**) are subject to the Gödel incompleteness theorems, then an agent with an infinitary mind can only employ an infinitary method to prove CON(**PA**). In which case, the anti-functionalist must demonstrate that human minds are infinitary or give up the view that there is an epistemic difference between human minds that have no finitary computational description and finitary computing machines governed by *P*.

If human minds, under the assumption they have no finitary computational description, prove CON(*P*) with less than mathematical certainty or in another epistemic modality, by weak methods not subject to the Gödel incompleteness theorems, they are not distinguishable from finitary computing machines that can similarly employ those weak methods to prove CON(*P*). If those weak methods are subject to the Gödel incompleteness theorems, then neither the human mind that has no finitary computational description nor the finitary computing machine can prove CON(*P*) in the characteristic epistemic modality of the proof procedures of the formal systems that formalize the weak methods.

The anti-functionalist wants to prove that all weak methods which could, under some standard of epistemic adequacy, prove CON(*P*), with less than mathematical certainty or in some other epistemic modality, are subject to the Gödel incompleteness theorems. Yet this task is just what engenders the absurdity. If all weak methods which could, under some standard of epistemic adequacy, prove CON(*P*) with less than mathematical certainty or in some other epistemic modality, are subject to the Gödel incompleteness theorems, then they cannot be used to prove CON(*P*), even by minds that have no finitary computational description. This is so, because whatever the epistemic modality of the proof of CON(*P*) no agent, no matter what its computational structure (whether finitary or metarecursive), can prove CON(*P*) in that epistemic modality.

Objection 7: An epistemic use of the Gödel theorems does, in fact, render a metaphysical conclusion. It shows that the cognitive structure of the human mind is subject to the Gödel incompleteness theorems. That, in turn, shows that we cannot be metaphysically distinguished from finitary computing machines.

Response: However, that is a moot conclusion, since the anti-functionalist who employs an EGF argument proceeds from the assumption that the human mind has a finitary computational description. That is, she proceeds from the adoption of the metaphysical picture of the human kind as a finitary computing machine. The Gödel theorems tell us about the limitations faced by such finitary computational descriptions, but the basic metaphysics is already in place. EGF arguments don’t conclude to a metaphysical conclusion, as is done in MGF arguments.

Objection 8: That the anti-functionalist falls into an absurdity in escaping from the simple logical error of Penrose is a clever observation, but it is false. We do not say that an absurdity arises out of the fact that Peano arithmetic is subject to the Gödel incompleteness theorems. A formal system strong enough to carry out (minimally) Robinson arithmetic is one for which we cannot, with mathematical certainty, employing a finitistic and effective proof procedure, prove its consistency. However, that we cannot is not license for us to infer that we can reasonably doubt that Peano arithmetic is subject to the Gödel incompleteness theorems. That is absurd. It is too easy a move. Certainly, we would have encountered someone in mathematics making it long ago. But no one did, because it is nothing short of being numbingly stupid.

Response: You are quite right about Peano arithmetic. No absurdity—of the kind we have specified—arises, and it would be numbingly stupid to claim one does. It is the assumption that all forms of reasoning are subject to the Gödel incompleteness theorems which produces the absurdity. The absurdity shows the assumption is false. Gödel did not prove that all formal systems are incomplete—only those that are string enough for Peano arithmetic.

Additionally, the epistemic situations with respect to Peano arithmetic and with respect to *P* are quite different. There is probably not a single mathematician who genuinely doubts the consistency of Peano arithmetic. There are infinitary proofs of it—Gentzen discovered one in the mid-thirties and Ackermann polished it five years later. We have good reason to believe that the Gentzen proof works. We have, then, no reason to believe that the Gödel incompleteness theorems fail to hold of a formal system that encompasses Peano arithmetic. There is no absurdity, even though we cannot prove, with mathematical certainty, using a finitistic and effective proof procedure, the consistency of Peano arithmetic. From that we do not conclude that Peano arithmetic might be inconsistent.

The epistemic situation is much different with respect to *P*, which is a finitary computational description based on a cognitive theory, an ultimate one at that. We do not have the same intuitions about its consistency that we have about the consistency of Peano arithmetic, for we do not even have the cognitive theory that underlies *P*. It is a suppositional device to carry out the anti-functionalist argument. Nor, for the same reasons, do we have an infinitary proof of CON(*P*). If *P* encompasses all finitary methods of inquiry into the world, and we show that all of these methods are subject to the Gödel incompleteness theorems, then we have no methods of inquiry left with which to carry out the consistency proof of *P*, other than infinitary ones. We cannot, however, say that we have good reason to believe that *P* is consistent, since we have no idea what it will look like and, even if we did, it is still based on a cognitive theory which has to be tested. If we cannot test it, because all our procedures for testing it are subject to the Gödel incompleteness theorems, we are in an epistemic situation of maximal ignorance. We have no good reason to believe it is consistent and no good reason to believe it is inconsistent. In that epistemic situation, we cannot accept the result that all epistemic methods of inquiry are subject to the Gödel incompleteness theorems. The absurdity cannot be dismissed by comparing it with the disanalogous epistemic situation in Peano arithmetic. It is, then, a genuine epistemic problem for the anti-functionalist.

Objection 9: You mistakenly think that since PGA and strengthened PGA incur an absurdity, it is left open for finitary human minds and finitary computing machines to use any weak methods of empirical inquiry into the structure of the world. The absurdity does not entitle the agent to use all weak methods. Given there is an absurdity, how would you determine the weak methods which escape being subject to the Gödel incompleteness theorems because of the absurdity? You cannot stipulate there are weak methods that can be used by a human agent. Just as a paradoxical sentence (such as the Liar sentence) can't be assumed true, agents can't conclude from the absurdity of strengthened PGA that there are weak methods that are not subject to the Gödel incompleteness theorems and that are thereby legitimate to use.

Response: That is a perceptive point, but it is misguided. The analogy with Liar sentences is not acceptable. Once we show a Liar sentence is paradoxical, we cannot assume it is true, nor can we assume that it is false. In some truth theories, we withhold assignment of a truth-value to it, in which case it has a null functional status in our discourses.

On the other hand, the assumption in strengthened PGA that led to the absurdity is that all methods of inquiry into the world are subject to the Gödel incompleteness theorems. The absurdity shows that assumption is false—not all methods of inquiry into the world are subject to the Gödel incompleteness theorems. So it is left open that there are weak methods which are not subject to the Gödel incompleteness theorems.

Objection 10: Any intuitions about the consistency of *P* must be seen as evidence for the claim that we have infinitary capacities. We would not have those intuitions unless there is some infinitary reasoning process, below the threshold of conscious perception, which accounts for them. The best explanation of why we have these intuitions is that there is some infinitary reasoning mechanism in us which causes us to have those intuitions. Thus, even though there is an absurdity for the anti-functionalist who wants to show all weak methods are subject to the Gödel incompleteness theorems, the intuitions we would (since *P* does not exist—it is merely a hypothetical construct) have about the correctness of *P* are reliable indicators of our infinitary capacities. The absurdity is no hindrance to the anti-functionalist, since human minds are infinitary and we do not even need PGA.

Response: If we do have intuitions that *P* is consistent, and we set a probability level for the reliability of those intuitions higher than the reliability we would—in probabilistic terms—rate the weak methods for showing *P* is consistent, with less than mathematical certainty or in some other epistemic modality, and we know that there are no other weak methods available and that only infinitary methods can prove the correctness of *P* with mathematical certainty, what can we reasonably conclude about the nature of our cognitive capacities? We can't reasonably conclude that we have infinitary cognitive capacities. It would be the case that the best explanation of our intuitions is that an infinitary reasoning mechanism causes us to have them if we had no alternative explanations of them. But we have alternative explanations of how we could have such intuitions, and these explanations do not posit infinitary reasoning processes. For instance, we have experiences with cognitive theories of inductive reasoning, and we see an analogy between them and *P*. If they are known to be consistent, we conclude that it is highly likely *P* is consistent as well. We might, also, be simply mistaken. Our probabilistic intuitions are notoriously shaky, a fact well-known to cognitive psychologists. In that case, the best explanation for our intuitions is that we have made errors in probabilistic reasoning. If we had independent evidence the human mind performs infinitary operations, then the explanation of our intuitions about the correctness of P in terms of infinitary operations would be superior to the two alternatives we have just cited. But, in the absence of that evidence, the two alternatives are not inferior to it, since they are sensitive to established work in cognitive psychology, while there is no established work that shows we have infinitary reasoning powers.

Objection 11: We can use the Gödel incompleteness theorems to show that there are capacities which human minds have that finitary computing machines do not have. Let the formal system characterizing the capacities of a finitary computing machine be P. Suppose P is subject to the Gödel incompleteness theorems. Then the finitary computing machine can't prove CON(P) and can't prove its own Gödel sentence. However, a human mind can prove CON(P) and the Gödel sentence in P by ascending to a more powerful formal system, P^*, that contains P. The finitary computing machine characterized by P, however, cannot ascend to P^*.

Response: That point is well-known in the functionalism debate. Perhaps ascent to P^* may prove futile, since P^* may be so long that finitary human minds cannot survey it and thus cannot prove that it is consistent. That is the Penrose error.

However, even if we discount the Penrose error, there is still a problem. Recall that what the second Gödel incompleteness theorem rules out is the possibility of finitistically proving, with mathematical certainty, and within the system P, CON(P). If one ascends to P^*, then CON(P) can be proved finitistically with mathematical certainty, period. However, this is true only if one can finitistically prove, with mathematical certainty, that P^* is consistent. But now the Gödel theorems take root in P^*. It is impossible to finitistically prove CON(P^*) with mathematical certainty, within P^*. That means that the ascent to P^* is futile unless P^* can be proved consistent. But that cannot be done within P^*. It can only be done by ascending to a stronger system P^{**} that contains both P and P^*. Within P^{**}, one can finitistically prove CON(P) and CON(P^*) with mathematical certainty, but only if P^{**} is consistent.

Notice the epistemic pattern which emerges. For any n less than omega, one can finitistically prove with mathematical certainty CON(P^n) in the formal system P^{n+1} only if one can finitistically prove, with mathematical certainty, CON(P^{n+1}). However, for any n less than omega, it is impossible to finitistically prove CON(P^n) with mathematical certainty within P^n.

The anti-functionalist will have to ascend infinitely high to the infinitary formal system Pomega, in order to finitistically prove, with mathematical certainty, CON(P). That is just to say that the anti-mechanist will have to possess the cognitive capacity to construct an infinite proof tree in order to finitistically prove, with mathematical certainty, CON(*P*). Indeed, this is true for any P^n, where *n* is less than omega.

It easily follows from these considerations that the anti-functionalist has no advantage over the functionalist in showing that there are cognitive capacities which finitary human minds possess, but which a finitary computing machine lacks. If human minds possess an infinitary cognitive capacity, there is something we possess that finitary computing machines lack. But there is no conclusive evidence that we possess an infinitary cognitive capacity. It is open to us to prove CON(*P*) with less than mathematical certainty or in another epistemic modality, but it is open to finitary computing machines to do the same as well. If the methods for proving CON(*P*) with less than mathematical certainty or in some other epistemic modality are subject to the Gödel incompleteness theorems, then the very same considerations expressed above will apply to this case also. In which case, the anti-functionalist has no advantage over the functionalist in demonstrating there is a cognitive capacity which human minds possesses that finitary computing machines lack.

Objection 12: The anti-functionalist using an MGF argument has an avenue of escape. Although there cannot be a finitistic proof within *P* that establishes, with mathematical certainty, CON(*P*), it is possible for a human mind (not susceptible to the Gödel incompleteness theorems) to prove CON(*P*), with mathematical certainty, by using mathematical reasoning that is not subject to the Gödel incompleteness theorems.

Response: That is a good objection, but it might not work. If the mathematical reasoning in question is captured by a formal system that is not subject to the Gödel incompleteness theorems, it might be too weak to finitistically prove CON(*P*) with mathematical certainty. Perhaps CON(*P*) could be finitistically proved with mathematical certainty in the ramified type theory of *Principia Mathematica*. But since there is no adequate theory of its intensional proof predicate (which is why it is not subject to the Gödel incompleteness theorems), it is not known whether such a proof will have mathematical certainty.

On the other hand, if there is a system of mathematical reasoning which is not subject to the Gödel incompleteness theorems only because it cannot be formalized (justified perhaps on philosophical grounds), such as Brouwer's view of intuitionism, it is not known whether such reasoning can establish its conclusions with mathematical certainty and it is not known whether such reasoning is (or is not) finitary.

There are systems of mathematical reasoning that are captured only by infinitary formal systems (such as the system in Turing's completeness theorem), that are not subject to the Gödel incompleteness theorems. But there is no conclusive evidence human agents can engage in infinitary reasoning, where proper infinitary reasoning implies the ability of the reasoner to construct infinitary proof trees. This will not help the anti-functionalist who uses an MGF argument.

The moral, then, is that the anti-functionalist can dream of a system of finitary mathematical reasoning which can finitistically prove CON(P) with mathematical certainty, and which is not subject to the Gödel incompleteness theorems But we have no reason to believe such a system of mathematical reasoning exists, nor that it is logically possible.

10. The Epistemology of Mathematical Certainty: A New Project for Philosophy of Mind

Proving that an arbitrary mathematical sentence is true is beyond the pale of a mechanical proof procedure, since the set of mathematical truths is not recursive, not recursively enumerable, and not definable in arithmetic. This is another reason why mechanical proof procedures that verify a proof of a theorem in mathematics must be mechanical. If we attempt to show that each line in a proof preserves truth by showing that each line in the proof is true in and of itself and without examining how it was obtained, there is no guarantee we will be able to complete the job of verifying the proof of the theorem (even if we have the time and resources). On the other hand, if the proof verification procedure is mechanical, then we do not check that each line of the proof is true. Rather, we check that it has the requisite syntactical form. The relation "p is a proof of α" is recursive, where "α" is a sentence in some language and "p" is a proof of that sentence. It follows that all of the theorems in that language are recursively enumerable. There is a fundamental dichotomy between proof and truth arising from these considerations. Mathematical truth is not recursively enumerable, while mathematical provability is recursively enumerable. One way of describing the Gödelian incompleteness phenomena is that they witness this dichotomy.

If we relax the standards of mathematical proof, we might not have assurance that intersubjective agreement can be reached as to whether a derivation is a legitimate proof of its conclusion. In which case, we cannot be assured we will be mathematically certain of the truth of the theorem derived. It is the epistemological requirement in mathematics that a proof establish with mathematical certainty the truth of its conclusion that allows the anti-functionalist to capitalize on the Gödel incompleteness theorems in EGF and MGF arguments. Relaxing this requirement in mathematics is relevant to the philosophy of mind. We must ask: what is the epistemic goodness of weak mathematical methods—those which do not confer mathematical certainty on what they establish?

An area in philosophy of mathematics that connects with philosophy of mind is mathematical intuitionism. Can intuitionistic reasoning as originally envisaged by Brouwer deliver mathematical certainty? Is it infinitistic? If so, does it have a metarecursive computational structure? Work needs to be done to explore Kreisel's musing:

> There is the old and familiar idea, or: idealization, which regards a thought and, in particular, a proof of a general proposition as an infinite object. [I]nfinite ob-

> jects are better representations of proofs than the words we use to communicate proofs… (1967, p. 203)

and Brouwer:

> These mental mathematical proofs that in general contain infinitely many terms must not be confused with their linguistic accompaniments, which are finite and necessarily inadequate, hence do not belong to mathematics. (1967, p. 460, note 8)

A virtue of Lucas-Penrose-Putnam anti-functionalist arguments is that they connect mathematical logic with the philosophy of mind and might cast light on issues in the foundations of mathematics.

Note 1: An anonymous reviewer of this paper made several important remarks: that Kreisel (1965; 1967) and Gödel (in his *Dialectica* paper; see Gödel, 1990) perhaps hold the view that human minds are capable of infinitary mental proofs, that Gödel (1995) perhaps believes mathematics is empirical (and so statistical methods would be an appropriate means of proving theorems), and that there is an interesting problem in Kripke's Schema (formalizing Brouwer's creating subject)—namely, the assumptions in his argument using the schema are incompatible with infinitary mental proofs. Van Atten (2018) provides an excellent discussion of this matter. If human minds are capable of infinitary mental proofs, the question of whether such mental acts have a metarecursive computational structure is raised and with it, whether such a computational structure can be accommodated within functionalism. I thank the anonymous reviewer for these remarks and other useful suggestions.

Note 2: This paper revises and expands two earlier versions (Buechner, 2007; 2010). The most prominent changes are the nature of the problem that I contend arises for Putnam's use of the Gödel incompleteness theorems to refute functionalism and the nature of the problem that arises for functionalists whose burden of proof is to show there are no ways (that avoid the incompleteness theorems) of establishing the consistency of first-order arithmetic with less than mathematical certainty or in some other epistemic modality than that of mathematical certainty. The most significant overlap is in the categorization of the Lucas-Penrose-Putnam anti-functionalist arguments. Although there are changes of emphasis in that categorization in this paper, I still believe it is a significant contribution to the role of the Gödel incompleteness theorems in the functionalism debate.

REFERENCES

Audi, R. (1988). *Belief, Justification, and Knowledge*. Belmont, CA: Wadsworth Publishing Company.

Boolos, G. (2001). Introductory Note to 'Some Basic Theorems on the Foundations of Mathematics and Their Implications'. In: S. Feferman et al. (Eds.), *Collected Works, Volume 3: Unpublished Essays and Lectures* (pp. 290–307). New York: Oxford University Press.

Brouwer, L. E. J. (1967). On the Domains of Definition of Functions. In J. van Heijenoort (Ed.), *From Frege to Gödel* (pp. 446–463). Cambridge, Mass.: Harvard University Press.

Buechner, J. (2007). *Gödel, Putnam, and Functionalism: A New Reading of 'Representation and Reality'*. Cambridge: MIT Press

Buechner, J. (2010). Are the Gödel Incompleteness Theorems Limitative Results for the Neurosciences? *J Biol Phys*, *36*, 23–44.

Buechner, J., Tavani, H. (2011). Trust and Multi-Agent Systems: Applying the "Diffuse, Default Model" of Trust to Experiments Involving Artificial Agents. *Ethics and Information Technology*, *13*, 39–51.

Carnap, R. (1952). *The Continuum of Inductive Methods*. Chicago: University of Chicago Press.

Carnap, R. (1962). *Logical Foundations of Probability* (2nd ed.). Chicago: University of Chicago Press.

Church, A. (1956). *Introduction to Mathematical Logic* (vol. 1). Princeton: Princeton University Press.

Feferman, S. (1960). Axiomatization of Metamathematics in a General Setting. *Fundamenta Mathematicae*, *49*, 35–92.

Gödel, K. (1990). On a Hitherto Unutilized Extension of the Finitary Standpoint. In: S. Feferman et al. (Eds.), *Collected Works, Volume II: Publications 1938–1974* (pp. 241–252). New York: Oxford University Press.

Gödel, K. (1995). Some Basic Theorems on the Foundations of Mathematics and Their Implications. In: S. Feferman et al. (Eds.), *Collected Works, Volume III: Unpublished Essays and Lectures* (pp. 304–323). New York: Oxford University Press.

Kaplan, D., Montague, R. (1960). A Paradox Regained? *Notre Dame Journal of Formal Logic*, *1*, 79–90.

Kreisel, G., Sacks, G. (1965). Metarecursive Sets. *Journal of Symbolic Logic*, *30*, 318–338.

Kreisel, G. (1965). Mathematical Logic. In: T. L. Saaty (Ed.), *Lectures on Modern Mathematics* (vol. 3, pp. 95–195). New York: Wiley.

Kreisel, G. (1967). Mathematical Logic: What Has It Done for the Philosophy of Mathematics? In: R. Schoenmann (Ed.), *Bertrand Russell: Philosopher of the Century* (pp. 201–272). London: George Allen & Unwin.

Kreisel, G. (1970). Church's Thesis: A Kind of Reducibility Axiom for Constructive Mathematics. In: A. Kino, J. Myhill, R. Vesley, (Eds.), *Intuitionism and Proof Theory* (pp. 121–150). Amsterdam: North-Holland.

Kreisel, G. (1972). Which Number Theoretic Problems Can Be Solved in Recursive Progressions on Π(1,1) Paths Through O? *Journal of Symbolic Logic*, *37*, 311–334.

Lehrer, K. (1990). *Theory of Knowledge*. Boulder, Colorado: Westview Publishing Company.

Lewis, D. (1969). Lucas Against Mechanism. *Philosophy*, *44*, 231–233.

Lewis, D. (1979). Lucas Against Mechanism II. *Canadian Journal of Philosophy*, *4*, 373–376.

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

Lucas, J. (1970). Mechanism: A Rejoinder. *Philosophy*, *45*, 149–151.

Marr, D. (2010). *Vision: A Computational Investigation Into the Human Representation and Processing of Visual Information*. Cambridge, MA: MIT Press.

Montague, R. (1963). Syntactical Treatments of Modality, with Corollaries on Reflection Principles and Finite Axiomatizabiity. *Acta Philosophica Fennica*, *16*, 153–167.

Penrose, R. (1999). *The Emperor's New Mind, with a New Preface by the Author*. New York: Oxford University Press.

Penrose, R. (1994). *Shadows of the Mind: A Search for the Missing Science of Consciousness*. New York: Oxford University Press.

Penrose, R. (2016). On Attempting to Model the Mathematical Mind. In: S. B. Cooper (Ed.), *The Once and Future Turing* (pp. 361–378). New York: Cambridge University Press.

Pollock, J. Cruz, J. (1999). *Contemporary Theories of Knowledge* (2nd edition). Lanham, Maryland: Rowman and Littlefield.

Putnam, H. (1988). *Representation and Reality*. Cambridge, MA: MIT Press.

Putnam, H. (1994a). *Words and Life*. Cambridge, MA: Harvard University Press.

Putnam, H. (1994b). Reflexive Reflections. In: H. Putnam, *Words and Life* (pp. 416–427). Cambridge, MA: Harvard University Press.

Putnam, H. (1995). Review of Roger Penrose's *Shadows of the Mind*. *Bulletin of the American Mathematical Society*, *32*, 370–373.

Salmon, N. (2001). The Limits of Human Mathematics. *Philosophical Perspectives*, *15*, 93–117.

Tarski, A. (1983). The Concept of Truth in Formalized Languages. In: A. Tarski, *Logic, Semantics, Metamathematics* (2nd edition, pp. 152–278). Indianapolis, Indiana: Hackett Publishing Company.

Thomason, R. (1980). A Note on Syntactical Treatments of Modality. *Synthese*, *44*, 391–395.

Thomason, R. (1989). Motivating Ramified Type Theory. In G. Chierchia, B. H. Partee, R. Turner (Eds.), *Properties, Types and Meaning, Volume 1: Foundational Issues* (pp. 47–62). Norwell, MA: Kluwer.

Turing, A. (1939). Systems of Logic Based on Ordinals. *Proceedings of the London Mathematical Society*, *45*, 161–228.

van Atten, M. (2018). The Creating Subject, the Brouwer-Kripke Schema, and Infinite Proofs. *Indagationes Mathematicae*, 29, 1565–1636.

von Neumann, J. (1969). Tribute to Dr. Gödel. In: J. J. Bulloff, T. C. Holyoke, S. W. Hahn (Eds.), *Foundations of Mathematics: Symposium Papers Commemorating the Sixtieth Birthday of Kurt Gödel* (pp. ix–x). Berlin: Springer-Verlag.

Wigderson, A. (2019). *Mathematics + Computation*. Princeton, NJ: Princeton University Press.

STUDIA SEMIOTYCZNE (SEMIOTIC STUDIES), t. XXXIV, nr 1 (2020), s. 159–182
ISSN 0137-6608, e-ISSN 2544-073X
DOI: 10.26333/sts.xxxiv1.07
Article

YONG CHENG [*]

GÖDEL'S INCOMPLETENESS THEOREM AND THE ANTI-MECHANIST ARGUMENT: REVISITED[1]

SUMMARY: This is a paper for a special issue of *Semiotic Studies* devoted to Stanislaw Krajewski's paper (2020). This paper gives some supplementary notes to Krajewski's (2020) on the Anti-Mechanist Arguments based on Gödel's incompleteness theorem. In Section 3, we give some additional explanations to Section 4–6 in Krajewski's (2020) and classify some misunderstandings of Gödel's incompleteness theorem related to Anti-Mechanist Arguments. In Section 4 and 5, we give a more detailed discussion of Gödel's Disjunctive Thesis, Gödel's Undemonstrability of Consistency Thesis and the definability of natural numbers as in Section 7–8 in Krajewski's (2020), describing how recent advances bear on these issues.

KEYWORDS: Gödel's incompleteness theorem, the Anti-Mechanist Argument, Gödel's Disjunctive Thesis, intensionality.

[*] Wuhan University, School of Philosophy. E-mail: world-cyr@hotmail.com. ORCID: 0000-0003-2408-3886.

[1] This paper is the research result of the Humanities and Social Sciences of Ministry of Education Planning Fund project "Research on Gödel's Incompleteness Theorem" (project no: 17YJA72040001). I would like to thank the fund support by the Humanities and Social Sciences of Ministry of Education Planning Fund in China. I thank the lamp for my feet, and the light on my path. I would also like to thank Peter Koellner for sending me his drafts of (2018a; 2018b) prior to publication, and the referee for detailed helpful comments.

1. Introduction

Gödel's incompleteness theorem is one of the most remarkable and profound discoveries in the 20th century, an important milestone in the history of modern logic. Gödel's incompleteness theorem has wide and profound influence on the development of logic, philosophy, mathematics, computer science and other fields, substantially shaping mathematical logic as well as foundations and philosophy of mathematics from 1931 onward. The impact of Gödel's incompleteness theorem is not confined to the community of mathematicians and logicians, and it has been very popular and widely used outside mathematics.

Gödel's incompleteness theorem raises a number of philosophical questions concerning the nature of mind and machine, the difference between human intelligence and machine intelligence, and the limit of machine intelligence. It is well known that Turing proposed a convincing analysis of the vague and informal notion of "computable" in terms of the precise mathematical notion of "computable by a Turing machine". So we can replace the vague notion of computation with the mathematically precise notion of a Turing machine. In this paper, following Koellner in (2018), we stipulate that the notion "the mind cannot be mechanized" means that the mathematical outputs of the idealized human mind outstrip the mathematical outputs of any Turing machine.[2] A popular interpretation of Gödel's first incompleteness theorem (G1) is that G1 implies that the mind cannot be mechanized. The Mechanistic Thesis claims that the mind can be mechanized. In this paper, we will not examine the broad question of whether the mind can be mechanized, which has been extensively discussed in the literature (e.g. Penrose, 1989; Chalmers, 1995; Lucas, 1996; Lindström, 2006; Feferman, 2009; Shapiro, 1998; 2003; Koellner, 2016; 2018; 2018; Krajewski, 2020). Instead we will only examine the question of whether G1 implies that the mind cannot be mechanized.

This is a paper for a special issue of *Semiotic Studies* devoted to Krajewski's paper (2020). We first give a summary of Krajewski's work in (2020). In (2020), Krajewski gave a detailed analysis of the alleged proof of the nonmechanical, or non-computational, character of the human mind based on Gödel's incompleteness theorem. Following Gödel himself and other leading logicians, Krajewski refuted the Anti-Mechanist Arguments (the Lucas Argument and the Penrose Argument), and claimed that they are not implied by Gödel's incompleteness theorem alone. Moreover, Krajewski (2020) demonstrated the inconsistency of Lucas's arithmetic and the semantic inadequacy of Penrose's arithmetic. Krajewski (2020) also discussed two consequences of Gödel's incompleteness theorem directly related to Anti-Mechanist Arguments: our consistency is not provable (Gödel's Undemonstrability of Consistency Thesis), and we cannot define the

[2] In this paper, we will not consider the performance of actual human minds, with their limitations and defects; but only consider the idealized human mind and look at what it can do in principle (Koellner, 2018a, p. 338).

natural numbers. The discussion in Krajewski's paper is mainly from the philosophical perspective. However, the discussion in this paper is mainly from the logical perspective based on some recent advances on the study of Gödel's incompleteness theorem and Gödel's Disjunctive Thesis. Basically, we agree with Krajewski's analysis of the Anti-Mechanist Arguments and his conclusion that Gödel's incompleteness theorem alone does not imply that the Anti-Mechanist Arguments hold. However, some discussions in (2020) are vague. Moreover, in the recent work on Gödel's Disjunction Thesis one finds precise versions which can actually be proved. The motivation of this paper is to give some supplementary notes to Krajewski's recent paper (2020) on the Anti-Mechanist Arguments based on Gödel's incompleteness theorem.

This paper is structured as follows. In Section 2, we review some notions and facts we will use in this paper. In Section 3, we give some supplementary notes to Section 5–6 in Krajewski's (2020) and classify some misunderstandings of Gödel's incompleteness theorem related to Anti-Mechanist Arguments. In Section 4, we give a more detailed discussion of Gödel's Disjunctive Thesis as in Section 7 in Krajewski's (2020) based on recent advances of the study on Gödel's Disjunctive Thesis in the literature. In Section 5, we give a more precise discussion of Gödel's Undemonstrability of Consistency Thesis and the definability of natural numbers as in Section 8 in Krajewski's paper.

2. Preliminaries

In this section, we review some basic notions and facts used in this paper. Our notations are standard. For textbooks on Gödel's incompleteness theorem, we refer to (Enderton, 2001; Murawski, 1999; Lindström, 1997; Smith, 2007; Boolos, 1993). There are some good survey papers on Gödel's incompleteness theorem in the literature (Smoryński, 1977; Beklemishev, 2010; Kotlarski, 2004; Visser, 2016; Cheng, in press).

In this paper, we focus on first order theory based on countable language, and always assume the arithmetization of the base theory with a recursive set of non-logical constants. For a given theory T, we use $L(T)$ to denote the language of T. For more details about arithmetization, we refer to (Murawski, 1999). Under the arithmetization, any formula or finite sequence of formulas can be coded by a natural number (called the Gödel number of the syntactic item). In this paper, $\ulcorner\varphi\urcorner$ denotes the numeral representing the Gödel number of φ.

We say a set of sentences Σ is *recursive* if the set of Gödel numbers of sentences in Σ is recursive.[3] A theory T is *decidable* if the set of sentences provable in T is recursive; otherwise it is *undecidable*. A theory T is *recursively axiomatizable* if it has a recursive set of axioms, i.e. the set of Gödel numbers of axioms of T is recursive. A theory T is *finitely axiomatizable* if it has a finite set of axioms. A theory T is *essentially undecidable* iff any recursively axiomatizable consistent

[3] For ease of exposition, we will pass back and forth between the two.

extension of T in the same language is undecidable. We say a sentence φ is *independent* of T if $T \nvdash \varphi$ and $T \nvdash \neg\varphi$. A theory T is *incomplete* if there is a sentence φ in $L(T)$ which is independent of T; otherwise, T is *complete* (i.e., for any sentence φ in $L(T)$, either $T \vdash \varphi$ or $T \vdash \neg\varphi$). Informally, an interpretation of a theory T in a theory S is a mapping from formulas of T to formulas of S that maps all axioms of T to sentences provable in S. If T is interpretable in S, then all sentences provable (refutable) in T are mapped, by the interpretation function, to sentences provable (refutable) in S. Interpretability can be accepted as a measure of strength of different theories. For the precise definition of interpretation, we refer to (Visser, 2011) for more details.

Theorem 2.1 (Tarski, Mostowski, Robinson, 1953, Theorem 7, p. 22)**.** *Let T_1 and T_2 be two consistent theories such that T_2 is interpretable in T_1. If T_2 is essentially undecidable, then T_1 is also essentially undecidable.*

Robinson Arithmetic **Q** was introduced in (1953) by Tarski, Mostowski and Robinson as a base axiomatic theory for investigating incompleteness and undecidability.

Definition 2.2. Robinson Arithmetic **Q** is defined in the language $\{\mathbf{0}, \mathbf{S}, +, \cdot\}$ with the following axioms:

$\mathbf{Q}_1$: $\forall x \forall y(\mathbf{S}x = \mathbf{S}y \rightarrow x = y)$;

$\mathbf{Q}_2$: $\forall x(\mathbf{S}x \neq \mathbf{0})$;

$\mathbf{Q}_3$: $\forall x(x \neq \mathbf{0} \rightarrow \exists y(x = \mathbf{S}y))$;

$\mathbf{Q}_4$: $\forall x \forall y(x + \mathbf{0} = x)$;

$\mathbf{Q}_5$: $\forall x \forall y(x + \mathbf{S}y = \mathbf{S}(x + y))$;

$\mathbf{Q}_6$: $\forall x(x \cdot \mathbf{0} = \mathbf{0})$;

$\mathbf{Q}_7$: $\forall x \forall y(x \cdot \mathbf{S}y = x \cdot y + x)$.

The theory **PA** consists of axioms $\mathbf{Q}_1$–$\mathbf{Q}_2$, $\mathbf{Q}_4$–$\mathbf{Q}_7$ in Definition 2.2 and the following axiom scheme of induction:

$$(\varphi(\mathbf{0}) \wedge \forall x(\varphi(x) \rightarrow \varphi(\mathbf{S}x))) \rightarrow \forall x\varphi(x),$$

where φ is a formula with at least one free variable x.

Let $\mathfrak{N} = \langle \mathbb{N}, +, \times \rangle$ denote the standard model of **PA**. We say $\varphi \in L(\mathbf{PA})$ is a true sentence of arithmetic if $\mathfrak{N} \models \varphi$. We define that $Th(\mathbb{N}, +, \cdot)$ is the set of sentence φ in $L(\mathbf{PA})$ such that $\mathfrak{N} \models \varphi$. Similarly, we have the definition of $Th(\mathbb{Z}, +, \cdot)$, $Th(\mathbb{Q}, +, \cdot)$ and $Th(\mathbb{R}, +, \cdot)$.

We introduce a hierarchy of $L(\mathbf{PA})$-formulas called the "arithmetical hierarchy" (Murawski, 1999; Hájek, Pudlák, 1993). Bounded formulas (Σ^0_0, or Π^0_0, or

Δ_0^0 formula) are built from atomic formulas using only propositional connectives and bounded quantifiers (in the form $\forall x \leq y$ or $\exists x \leq y$). A formula is Σ_{n+1}^0 if it has the form $\exists x\varphi$ where φ is Π_n^0. A formula is Π_{n+1}^0 if it has the form $\forall x\varphi$ where φ is Σ_n^0. Thus, a Σ_n^0-formula has a block of n alternating quantifiers, the first one being existential, and this block is followed by a bounded formula. Similarly for Π_n^0-formulas. A formula is Δ_n^0 if it is equivalent to both a Σ_n^0 formula and a Π_n^0 formula.

A theory T is said to be ω-consistent if there is no formula $\varphi(x)$ such that $T \vdash \exists x\varphi(x)$ and for any $n \in \omega$, $T \vdash \neg\varphi(\bar{n})$. A theory T is 1-consistent if there is no such formula $\varphi(x)$ which is Δ_1^0. A theory T is sound iff for any formula φ, if $T \vdash \varphi$, then $\mathfrak{N} \vDash \varphi$; a theory T is Σ_1^0-sound iff for any Σ_1^0 formula φ, if $T \vdash \varphi$, then $\mathfrak{N} \vDash \varphi$.

In the following, unless stated otherwise, let T be a recursively axiomatizable consistent extension of **PA**. There is a formal arithmetical formula $\mathbf{Proof}_T(x,y)$ (called Gödel's proof predicate) which represents the recursive relation $Proof_T(x,y)$ saying that y is the Gödel number of a proof in T of the formula with Gödel number x. Define $\mathbf{Prov}_T(x) \triangleq \exists y\mathbf{Proof}_T(x,y)$. Since we will discuss general provability predicates based on proof predicates, now we give a general definition of proof predicate which is a generalization of properties of Gödel's proof predicate $\mathbf{Proof}_T(x,y)$.

Definition 2.3. We say a formula $\mathbf{Prf}_T(x,y)$ is a proof predicate of T if it satisfies the following conditions:[4]

(1) $\mathbf{Prf}_T(x,y)$ is $\Delta_1^0(\mathbf{PA})$;[5]

(2) $\mathbf{PA} \vdash \forall x(\mathbf{Prov}_T(x) \leftrightarrow \exists y\mathbf{Prf}_T(x,y))$;

(3) for any $n \in \omega$ and formula φ, $\mathbb{N} \vDash \mathbf{Proof}_T(\ulcorner\varphi\urcorner, n) \leftrightarrow \mathbf{Prf}_T(\ulcorner\varphi\urcorner, n)$;

(4) $\mathbf{PA} \vdash \forall x\forall x'\forall y(\mathbf{Prf}_T(x,y)\wedge\mathbf{Prf}_T(x',y) \rightarrow x = x')$.

We define the provability predicate $\mathbf{Pr}_T(x)$ from a proof predicate $\mathbf{Prf}_T(x,y)$ by $\exists y\mathbf{Prf}_T(x,y)$, and the consistency statement $\mathbf{Con}(T)$ from a provability predicate $\mathbf{Pr}_T(x)$ by $\neg\mathbf{Pr}_T(\ulcorner\mathbf{0} \neq \mathbf{0}\urcorner)$.

D1: If $T \vdash \varphi$, then $T \vdash \mathbf{Pr}_T(\ulcorner\varphi\urcorner)$;

D2: If $T \vdash \mathbf{Pr}_T(\ulcorner\varphi \rightarrow \phi\urcorner) \rightarrow (\mathbf{Pr}_T(\ulcorner\varphi\urcorner) \rightarrow \mathbf{Pr}_T(\ulcorner\phi\urcorner))$;

D3: $T \vdash \mathbf{Pr}_T(\ulcorner\varphi\urcorner) \rightarrow \mathbf{Pr}_T(\ulcorner\mathbf{Pr}_T(\ulcorner\varphi\urcorner)\urcorner)$.

[4] We can say that each proof predicate represents the relation "y is the code of a proof in T of a formula with Gödel number x".

[5] We say a formula φ is $\Delta_1^0(\mathbf{PA})$ if there exists a Σ_1^0 formula α such that $\mathbf{PA} \vdash \varphi \leftrightarrow \alpha$, and there exists a Π_1^0 formula β such that $\mathbf{PA} \vdash \varphi \leftrightarrow \beta$.

D1–**D3** is called the Hilbert-Bernays-Löb derivability condition. Note that **D1** holds for any provability predicate $\mathbf{Pr}_T(x)$. We say that provability predicate $\mathbf{Pr}_T(x)$ is standard if it satisfies **D2** and **D3**. In this paper, unless stated otherwise, we assume that $\mathbf{Con}(T)$ is the canonical arithmetic sentence expressing the consistency of T and $\mathbf{Con}(T)$ is formulated via a standard provability predicate.

The reflection principle for T, denoted by $\mathbf{Rfn}_T$, is the schema $\mathbf{Pr}_T(\ulcorner\varphi\urcorner) \to \varphi$ for every sentence φ in $L(T)$. The reflection principle for T restricted to a class of sentences Γ will be denoted by Γ-$\mathbf{Rfn}_T$.

Let $\alpha(x)$ be a formula in $L(T)$. We can similarly define the provability predicate and consistency statement w.r.t. formula $\alpha(x)$ as follows. Define the formula $\mathbf{Prf}_\alpha(x,y)$ saying "y is the Gödel number of a proof of the formula with Gödel number x from the set of all sentences satisfying $\alpha(x)$". Define the provability predicate $\mathbf{Pr}_\alpha(x)$ of $\alpha(x)$ as $\exists y \mathbf{Prf}_\alpha(x,y)$ and the consistency statement $\mathbf{Con}_\alpha(T)$ as $\neg\mathbf{Pr}_\alpha(\ulcorner\mathbf{0} \neq \mathbf{0}\urcorner)$. We say that formula $\alpha(x)$ is a numeration of T if for any n, $T \vdash \alpha(\overline{n})$ iff n is the Gödel number of some sentence in T.

3. Some notes on Gödel-Based Anti-Mechanist Arguments

There has been a massive amount of literature on the Anti-Mechanist Arguments due primarily to Lucas and Penrose (see Lucas, 1961; Penrose; 1989) which claim that G1 shows that the human mind cannot be mechanized. The Anti-Mechanist Argument began with Nagel and Newman in (2001) and continued with Lucas's publication in (1961). Nagel and Newman's argument was criticized by Putnam in (1960) and earlier by Gödel (Feferman, 2009), while Lucas's argument was much more widely criticized in the literature. See Feferman (2009) for a historical account and Benacerraf (1967) for an influential criticism of Lucas. Penrose proposed a new argument for the Anti-Mechanist Argument in (1994; 2011). Penrose's new argument is the most sophisticated and promising Anti-Mechanist Argument which has been extensively discussed and carefully analyzed in the literature (Chalmers, 1995; Feferman, 1995; Lindström, 2001; 2006; Shapiro, 1998; 2003; Gaifman, 2000; Koellner, 2016; 2018a; 2018b, etc.)

Most philosophers and logicians believe that variants of the arguments of Lucas and Penrose are not fully convincing. However, they do not agree so well on what is wrong with arguments of Lucas and Penrose. One strength of Krajewski's paper (2020) is that it provides a detailed review of the history of Anti-Mechanist Arguments based on Gödel's incompleteness theorem (Krajewski, 2020, Section 3) and an analysis of these Gödel-Based Anti-Mechanist Arguments (e.g. Lucas's argument in Section 4 and Penrose's argument in Section 6 in [Krajewski, 2020]). In this section, based on Krajewski's work, we give some supplementary notes of Krajewski's Sections 5–6.

For us, the Gödel-Based Anti-Mechanist Argument comes from some misinterpretations of Gödel's incompleteness theorem. To understand the source of these misinterpretations or illusions, we should first have correct interpretations of Gödel's incompleteness theorem. In the following, we first review some important facts about Gödel's incompleteness theorem which are helpful to clarify some misinterpretations of Gödel's incompleteness theorem.

Gödel proved his incompleteness theorem in (1931) for a certain formal system **P** related to Russell-Whitehead's Principia Mathematica and based on the simple theory of types over the natural number series and the Dedekind-Peano axioms (Beklemishev, 2010, p. 3). Gödel's original first incompleteness theorem (1931, Theorem VI) says that for formal theory T formulated in the language of **P** and obtained by adding a primitive recursive set of axioms to the system **P**, if T is ω-consistent, then T is incomplete. The following theorem is a modern reformulation of Gödel's first incompleteness theorem.

Theorem 3.1 (Gödel's first incompleteness theorem (G1))**.** *If T is a recursively axiomatized extension of* **PA**, *then there exists a Gödel sentence* **G** *such that:*

(1) if T is consistent, then $T \nvdash \mathbf{G}$;

(2) if T is ω-consistent, then $T \nvdash \neg\mathbf{G}$.

Thus if T is ω-consistent, then **G** is independent of T and hence T is incomplete. If T is consistent, Gödel sentence **G** is a true Π^0_1 sentence of arithmetic. Gödel's proof of G1 is constructive: one can effectively find a true Π^0_1 sentence **G** of arithmetic such that **G** is independent of T assuming T is ω-consistent. Gödel calls this the "incompletability or inexhaustability of mathematics". Note that only assuming that T is consistent, we can show that **G** is a true sentence of arithmetic unprovable in T. But it is not enough to show that $T \nvdash \neg\mathbf{G}$ only assuming that T is consistent. To show that $T \nvdash \neg\mathbf{G}$, we need a stronger condition such as "T is 1-consistent" or "T is Σ^0_1-sound".

Let T be a recursively axiomatized extension of **PA**. After Gödel, Rosser constructed Rosser sentence **R** (a Π^0_1 sentence) and showed that if T is consistent, then **R** is independent of T. Rosser improved Gödel's G1 in the sense that Rosser proved that T is incomplete only assuming that "T is consistent" which is weaker than "T is 1-consistent".

In this paper, let $\langle M_n : n \in \omega\rangle$ be the list of Turing machines and $Th(M_n)$ be the set of sentences produced by the Turing machine M_n. Let $C = \{n : Th(M_n)$ is a consistent theory$\}$ and $S = \{n : Th(M_n)$ is a sound theory$\}$.

The following proposition on inconsistency and unsoundness is from (Krajewski, 2020).

Proposition 3.2.

(1) If F is a partial recursive function such that $C \subseteq dom(F)$ and $F(n) \notin Th(M_n)$ for any $n \in C$, then $\{F(n) : n \in dom(F)\}$ is inconsistent.

(2) If F is a partial recursive function such that $S \subseteq dom(F)$ and $F(n) \notin Th(M_n)$ for any $n \in S$, then $\{F(n) : n \in dom(F)\}$ is inconsistent.

A natural question is: whether there exists such a function F with these properties. However, the effective version of Gödel's first incompleteness theorem (EG1) tells us that there exists a partial recursive function F such that for any $n \in \omega$, if $Th(M_n)$ is consistent, then $F(n)$ is defined and $F(n)$ is the Gödel number of a true arithmetic sentence which is not provable in $Th(M_n)$. Thus there exists such a function F with the properties as stated in Proposition 3.2.

One popular interpretation of EG1 is: for any Turing machine M_n, $F(n)$ picks up the true sentence of arithmetic not produced by M_n. However, this is a misinterpretation of EG1 which in fact says that for such a partial recursive function F, if $Th(M_n)$ is consistent, then $F(n)$ is the Gödel number of a true sentence of arithmetic which is not provable in $Th(M_n)$. A natural question is: whether there exists an effective procedure such that we can decide whether $Th(M_n)$ is consistent. The answer is negative since C is a complete Π^0_1 set as Koellner points out in (2018a).

Krajewski (2020) claimed that C and S are not recursive. However, as Krajewski (2020) commented, Proposition 3.2 on inconsistency and unsoundness does not require that for $n \in dom(F)$, $F(n)$ is the code of a true arithmetic sentence. But we do not see that C or S is not recursive from Proposition 3.2. However, if we add the condition that for $n \in dom(F)$, $F(n) \in \mathbf{Truth} \setminus Th(M_n)$, then we can show that C and S are not recursive. Let us take C for example and show that C is not recursive.

Proposition 3.3. *C is not recursive.*[6]

Proof. Suppose C is recursive. Let $A = \{F(n) : n \in C\}$. Then A is recursive enumerable. Suppose $A = Th(M_m)$ for some m. Note that $A \subseteq \mathbf{Truth}$, and so A is consistent. By the definition of C, $m \in C$ and hence $F(m) \in A$. But, on the other hand, $F(m) \notin Th(M_m) = A$ which leads to a contradiction. □

Since C is undecidable, it is impossible to effectively distinguish the case that $Th(M_n)$ is consistent and the case that $Th(M_n)$ is not consistent.

In fact, Theorem 3.2 can be generalized in the following form:

Theorem 3.4. *Let P be any property about first order theory (i.e. consistency, soundness, 1-consistency, etc). Let $C = \{n : Th(M_n)$ has property $P\}$. Suppose F is a partial recursive function satisfying the following conditions:*

(1) $C \subseteq dom(F)$,

[6] In fact, C is a complete Π^0_1 set as Koellner points out in (2018a).

(2) for each $n \in C$, $F(n) \notin Th(M_n)$.

Then, $\{F(n) : n \in dom(F)\}$ *does not have property* P.

Proof. Let $A = \{F(n) : n \in dom(F)\}$. Suppose A has property P. Since F is partial recursive, A is recursively enumerable. Suppose $A = Th(M_k)$ for some k. Since A has property P, we have $k \in C$. Thus, $F(k) \notin Th(M_k) = A$ which contradicts that $F(k) \in A$. □

Gödel announced the second incompleteness theorem (G2) in an abstract published in October 1930: no consistency proof of systems such as Principia, Zermelo-Fraenkel set theory, or the systems investigated by Ackermann and von Neumann is possible by methods which can be formulated in these systems (see Zach, 2007, p. 431). For a theory T, recall that $\mathbf{Con}(T)$ is the canonical arithmetic sentence expressing the consistency of T under Gödel's recursive arithmetization of T. The following is a modern reformulation of G2:

Theorem 3.5. *Let T be a recursively axiomatized extension of* **PA**. *If T is consistent, then* $T \nvdash \mathbf{Con}(T)$.

From G2, we cannot get that $\mathbf{Con}(T)$ is independent of T only assuming that T is consistent. It is provable in T that if T is consistent, then $T \vdash \mathbf{Con}(T) \leftrightarrow \mathbf{G}$ and thus $T \nvdash \mathbf{Con}(T)$. However, it is not provable in T that if T is consistent, then $T + \mathbf{Con}(T)$ is also consistent.[7] So it is not enough to show that $T \nvdash \neg\mathbf{Con}(T)$ only assuming that T is consistent. But we could prove that $\mathbf{Con}(T)$ is independent of T by assuming that T is 1-consistent which is stronger than the condition "T is consistent".[8] Let 1-$\mathbf{Con}(T)$ be the sentence in $L(\mathbf{PA})$ expressing that T is 1-consistent. Fact 3.6 is a summary of these results.

Fact 3.6. Let T be a recursively axiomatized consistent extension of **PA**.

(1) $T \vdash \mathbf{Con}(T) \to \mathbf{Con}(T + \neg\mathbf{Con}(T))$;

(2) $T \nvdash \mathbf{Con}(T) \to \mathbf{Con}(T + \mathbf{Con}(T))$;

(3) $T \vdash \mathbf{Con}(T) \to \mathbf{Con}(T + \mathbf{R})$;[9]

(4) $T \vdash 1\text{-}\mathbf{Con}(T) \to \mathbf{Con}(T + \mathbf{Con}(T))$.

An illusion of the application of Gödel's incompleteness theorem is that we can add consistencies (or Out-Gödeling) forever: from $\mathbf{Con}(T)$, we have

[7] See (Boolos, 1993, Theorem 4, p. 97) for a modal proof in **GL** of this fact using the arithmetic completeness theorem for **GL**.

[8] It is an easy fact that if T is 1-consistent and S is not a theorem of T, then $\mathbf{Pr}_T(\ulcorner S \urcorner)$ is not a theorem of T.

[9] Recall that **R** is the Rosser sentence.

Con(T + **Con**(T)), then **Con**(T + **Con**(T + **Con**(T))) and so on. However, by Fact 3.6, this does not hold. For the iteration of adding the consistency statement (or Out-Gödeling), we need a stronger condition: T is 1-consistent. The following fact shows the difference between **Con**(T) and 1-**Con**(T).

Fact 3.7 (Smoryński, 1977)**.** Let T be a recursively axiomatized consistent extension of **PA**. Then $T \vdash \mathbf{Con}(T) \leftrightarrow \Pi^0_1\text{-}\mathbf{Rfn}_T$ and $T \vdash 1\text{-}\mathbf{Con}(T) \leftrightarrow \Sigma^0_1\text{-}\mathbf{Rfn}_T$.

As a corollary of Fact 3.7, $1\text{-}\mathbf{Con}(T) \vdash 1\text{-}\mathbf{Con}(T + \mathbf{Con}(T))$ (see Proposition 3 in Pudlák, 1999). Thus, if we assume 1-**Con**(T), then we can prove **Con**(T), **Con**(T + **Con**(T)), **Con**(T + **Con**(T + **Con**(T))) and we can continue forever (note that the assumption 1-**Con**(T) is stronger than all these statements).

In summary, the differences between Rosser sentence and Gödel sentence, as well as between **Con**(T) and 1-**Con**(T) are very important. However, these differences are often overlooked in informal philosophical discussions of Gödel's incompleteness theorem.

4. Gödel's Disjunctive Thesis

The focus of Krajewski's paper (2020) is not about Gödel's Disjunctive Thesis even if he gives a very brief discussion of Gödel's Disjunctive Thesis related to the Anti-Mechanist Arguments in Section 7. In this section, we give a more detailed discussion of Gödel's Disjunctive Thesis and its relevance to the Mechanistic Thesis based on recent advances on the study of Gödel's Disjunctive Thesis. This section is a summary of Koellner's papers (2018a) and (2018b), and we follow Koellner's presentation very closely.

Gödel did not argue that his incompleteness theorem implies that the mind cannot be mechanized. Instead, Gödel argued that his incompleteness theorem implies a weaker conclusion: Gödel's Disjunctive Thesis (GD).

The first disjunct: The mind cannot be mechanized.

The second disjunct: There are absolutely undecidable statements.[10]

Gödel's Disjunctive Thesis (GD)**:** Either the first disjunct or the second disjunct holds.[11]

[10] In the sense that there are mathematical truths that cannot be proved by the idealized human mind.

[11] The original version of GD was introduced by Gödel in (1951; see p. 310): "So the following disjunctive conclusion is inevitable: either mathematics is incompletable in this sense, that its evident axioms can never be comprised in a finite rule, that is to say, the human mind (even within the realm of pure mathematics) infinitely surpasses the powers of any finite machine, or else there exist absolutely unsolvable diophantine problems of the type specified (where the case that both terms of the disjunction are true is not excluded, so that there are, strictly speaking, three alternatives)".

Gödel's Disjunctive thesis (GD) concerns the limit of mathematical knowledge and the possibility of the existence of mathematical truths that are inaccessible to the idealized human mind. The first disjunct expresses an aspect of the power of the idealized human mind, while the second disjunct expresses an aspect of its limitations.[12]

What about Gödel's view toward the first disjunct and the second disjunct? For Gödel, the first disjunct is true and the second disjunct is false; that is the mind cannot be mechanized and human mind is sufficiently powerful to capture all mathematical truths. Gödel's incompleteness theorem shows certain weaknesses and limitations of one given Turing machine. For Gödel, mathematical proof is an essentially creative activity and his incompleteness theorem indicates the creative power of human reason. Gödel believes that the distinctiveness of the human mind when compared to a Turing machine is evident in its ability to come up with new axioms and develop new mathematical theories. Gödel shared Hilbert's belief expressed in 1926 in the words: "in mathematics there is no ignoramuses, we should know and we must know" (see Reid, 1996, p. 192). Based on his rationalistic optimism, Gödel believed that we are arithmetically omniscient and the second disjunct is false.[13] However, Gödel admits that he cannot give a convincing argument for either the first disjunct or the second disjunct. Gödel thinks that the most he can claim to have established is his Disjunctive Thesis. For Gödel, GD is a "mathematically established fact" of great philosophical interest which follows from his incompleteness theorem, and it is "entirely independent from the standpoint taken toward the foundation of mathematics" (Gödel, 1951, p. 310).[14] In the following, we give a concise overview of the current progress on Gödel's disjunctive thesis based on Koellner's work in (2016; 2018a; 2018b).

Let $\mathbf{K}$ be the set of sentences in $L(\mathbf{PA})$ that the idealized human mind can know. Let $\mathbf{Truth}$ be the set of sentences in $L(\mathbf{PA})$ which are true in the standard model of arithmetic and $\mathbf{Prov}$ be the set of sentences in $L(\mathbf{PA})$ which are provable in $\mathbf{PA}$. Gödel refers to $\mathbf{Truth}$ as objective mathematics and $\mathbf{K}$ as subjective mathematics. Recall that a theory T in $L(\mathbf{PA})$ is sound if $T \subseteq \mathbf{Truth}$. In this paper, we assume that $\mathbf{K}$ is sound. However, from G1, we have $\mathbf{Prov} \subsetneq \mathbf{Truth}$ since Gödel's sentence is a true sentence of arithmetic not provable in $\mathbf{PA}$.[15]

[12] We refer to (Horsten, Welch, 2016), a recent comprehensive research volume about GD, for more discussions of the status of GD.

[13] For more discussions of the status of the second disjunct, we refer to (Horsten, Welch, 2016).

[14] In the literature there is a consensus that Gödel's argument for GD is definitive, but until now we have no compelling evidence for or against any of the two disjuncts (Horsten, Welch, 2016).

[15] Let us take Fermat's last theorem for another example. People have shown that Fermat's last theorem is a true sentence of arithmetic but, as far as I know, it is still an open problem whether Fermat's last theorem is provable in $\mathbf{PA}$. So Fermat's last theorem belongs to $\mathbf{K}$ but it is open whether it belongs to $\mathbf{Prov}$.

Note that GD concerns the concepts of relative provability, absolute provability, and truth. Before we present the analysis of GD, let us first examine two key notions about provability: relative provability and absolute provability. The notion of relative provability is well understood and we have a precise definition of relative provability in a formal system. But the notion of absolute provability is much more ambiguous and we have no unambiguous formal definition of absolute provability as far as we know. The notion of absolute provability is intended to be intensionally different from the notion of relative provability in that absolute provability is not conceptually connected to a formal system. In contrast to the notion of relative provability, there is little agreement on what principles of the notion "absolute provability" should be adopted. In this paper, we identify the notion of "relatively provable with respect to a given formal system F" with the notion of "producible by a Turing machine M" (where M is the Turing machine corresponding to F)[16] and we identify the notion of "absolute provability" with the notion of "what the idealized human mind can know".[17] Under this assumption, $\mathbf{K}$ is just the set of sentences that are absolutely provable.

In this paper, we assume without loss of generality that $\mathbf{Q} \subseteq Th(M_n)$ such that both G1 and G2 apply to $Th(M_n)$. For a natural number n, we say that a statement φ is *relatively undecidable* w.r.t. theory $Th(M_n)$ for some n if $\varphi \notin Th(M_n)$ and $\neg\varphi \notin Th(M_n)$. We say that a statement φ is *absolutely undecidable* if $\varphi \notin \mathbf{K}$ and $\neg\varphi \notin \mathbf{K}$. Let us first examine what the incompleteness theorem tells us about the relationship between $Th(M_n)$, $\mathbf{K}$ and $\mathbf{Truth}$.

Note that G1 tells us that for any sufficiently strong consistent theory F containing $\mathbf{Q}$, there are statements which are relatively undecidable with respect to F. But as Gödel argued, these statements are not absolutely undecidable; instead one can always pass to higher systems in which the sentence in question is provable (see Gödel, 1995, p. 35). For example, from G2, $\mathbf{Con(PA)}$ is not provable in $\mathbf{PA}$; but $\mathbf{Con(PA)}$ is provable in second order arithmetic ($\mathbf{Z_2}$). Since G2 applies to $\mathbf{Z_2}$, the Π^1_0-truth $\mathbf{Con(Z_2)}$ is not provable in $\mathbf{Z_2}$. But $\mathbf{Con(Z_2)}$ is provable in $\mathbf{Z_3}$ (third order arithmetic) which captures the Π^1_0-truth that was missed by $\mathbf{Z_2}$. This pattern continues up through the orders of arithmetic and up through the hierarchy of set-theoretic systems; at each stage a missing Π^1_0-truth is captured at the next stage (see Koellner, 2018a, p. 347).

Now let us examine the question of whether the incompleteness theorem shows that GD holds. From the literature, we have found a natural framework $\mathbf{EA_T}$ in which we can show that if the concepts of relative provability, absolute provability and truth satisfy some principles, then one can give a rigorous proof of GD, vindicating Gödel's claim that GD is a mathematically established fact (see Koellner, 2018a, p. 355).

[16] Note that sentences relatively provable with respect to a given formal system F can be enumerated by a Turing machine.

[17] Williamson (2016) makes the similar definition that a mathematical hypothesis is absolutely decidable if and only if either it or its negation can in principle be known by a normal mathematical process; otherwise it is absolutely undecidable.

Now we introduce two systems of epistemic arithmetic: **EA** and **EA$_{\mathbf{T}}$**. For the presentation of **EA** and **EA$_{\mathbf{T}}$**, we closely follow Koellner's discussion in (2016; 2018a). The first is designed to deal with $Th(M_e)$ and **K**, and the second is designed to deal with $Th(M_e)$, **K** and **Truth**. For **EA$_{\mathbf{T}}$**, we only require a typed truth predicate.[18] The basic system **EA** of epistemic arithmetic has axioms of arithmetic and axioms of absolute provability, and the extended system **EA$_{\mathbf{T}}$** has additional axioms of typed truth.[19] In **EA** and **EA$_{\mathbf{T}}$**, **K** is treated as an operator rather than a predicate. From results in Gödel (1986), Myhill (1960), Montague (1963), Thomason (1980), and others, if one formulates a theory of absolute provability with **K** as a predicate then inconsistency may come (see Koellner, 2016). The basic axioms of absolute provability are:[20]

K1: Universal closures of formulas of the form $\mathbf{K}\varphi$ where φ is a first-order validity.

K2: Universal closures of formulas of the form $(\mathbf{K}(\varphi \to \psi) \wedge \mathbf{K}\varphi) \to \mathbf{K}\psi$.

K3: Universal closures of formulas of the form $\mathbf{K}\varphi \to \varphi$.

K4: Universal closures of formulas of the form $\mathbf{K}\varphi \to \mathbf{KK}\varphi$.[21]

The language L(**EA**) is L(**PA**) expanded to include an operator **K** that takes formulas of L(**EA**) as arguments. The axioms of arithmetic are simply those of **PA**, only now the induction scheme is taken to cover all formulas in L(**EA**). For a collection Γ of formulas in L(**EA**), let $\mathbf{K}\Gamma$ denote the collection of formulas $\mathbf{K}\varphi$ where $\varphi \in \Gamma$. The system **EA** is the theory axiomatized by $\Sigma \cup \mathbf{K}\Sigma$, where Σ consists of the axioms of **PA** in the language L(**EA**) and the basic axioms of absolute provability. The language L(**EA$_{\mathbf{T}}$**) of **EA$_{\mathbf{T}}$** is the language L(**EA**) augmented with a unary predicate T. The system **EA$_{\mathbf{T}}$** is the theory axiomatized by $\Sigma \cup \mathbf{K}\Sigma$, where Σ consists of the axioms of **PA** in the language L(**EA$_{\mathbf{T}}$**), the basic

[18] A typed truth predicate is one that applies only to statements that do not themselves involve the truth predicate. In contrast, a type-free truth predicate is one which also applies to statements that themselves involve the truth predicate. The principles governing typed truth predicates are perfectly straightforward and uncontroversial, while the principles governing type-free truth predicates are much more delicate (Koellner, 2018a).

[19] These systems were first introduced by Myhill (1960), Reinhardt (1985a; 1985b; 1986) and Shapiro (1985), and then investigated by many others (e.g. Horsten, 1998; Leitgeb, 2009; Carlson, 2000; Koellner, 2016; 2018a and others).

[20] The basic conditions we will impose on knowability are: (1) if the idealized human mind knows φ and $\varphi \to \psi$ then the idealized human mind knows ψ; (2) if the idealized human mind knows φ then φ is true; (3) if the idealized human mind knows φ then the idealized human mind knows that the idealized human mind knows φ.

[21] **K1**-known as logical omniscience-says that **K** holds of all first-order logical validities; **K2** says that **K** is closed under modus ponens, and so distributes across logical derivations; **K3** says that **K** is correct; and **K4** says that **K** is absolutely self-reflective (Koellner, 2018a).

axioms of absolute provability (in the language $L(\mathbf{EA_T})$), and the Tarskian axioms of truth for the language $L(\mathbf{EA})$.

From the incompleteness theorem, Gödel made the following two claims about the relationship between $Th(M_e)$, $\mathbf{K}$ and $\mathbf{Truth}$.

Claim One: For any $e \in \mathbb{N}$, $\mathbf{K}(Th(M_e) \subseteq \mathbf{Truth}) \to Th(M_e) \subsetneqq \mathbf{K}$.[22]

Claim Two: Either $\neg\exists e(Th(M_e) = \mathbf{K})$ or $\exists\varphi(\varphi \in \mathbf{Truth} \wedge \varphi \notin \mathbf{K} \wedge \neg\varphi \notin \mathbf{K})$.[23]

Gödel's Claim One is formalizable and provable in $\mathbf{EA_T}$. In fact, something stronger is provable in $\mathbf{EA}$ as the following theorem shows:

Theorem 4.1 (Reinhardt, 1985a)**.** *Assume that S includes* $\mathbf{EA}$. *Suppose $F(x)$ is a formula with one free variable.*

(1) If for each sentence φ, $S \vdash \mathbf{K}(F(\ulcorner\varphi\urcorner) \to \varphi)$. Then there is a sentence ϕ such that $S \vdash \mathbf{K}\phi \wedge \mathbf{K}\neg F(\ulcorner\phi\urcorner)$.

(2) If for each sentence φ, $S \vdash \mathbf{K}(\mathbf{K}\varphi \to F(\ulcorner\varphi\urcorner))$. Then $S \vdash \mathbf{K}\neg\mathbf{K}(\mathbf{Con}(F))$.

From the following theorem, GD is also formalizable and provable in $\mathbf{EA_T}$ which confirms Gödel's claim that GD is a mathematically established fact.[24]

Theorem 4.2 (Reinhardt, 1986)**.** *Assume* $\mathbf{EA_T}$. *Then* GD *holds.*

Following Reinhardt, we should distinguish three levels of the mechanistic thesis.

(1) The weak mechanistic thesis (WMT): $\exists e(\mathbf{K} = Th(M_e))$;

(2) The strong mechanistic thesis (SMT): $\mathbf{K}\exists e(\mathbf{K} = Th(M_e))$;

(3) The super strong mechanistic thesis (SSMT): $\exists e\,\mathbf{K}(\mathbf{K} = Th(M_e))$.

Note that WMT is just the first disjunct which says that there is a Turing machine which coincides with the idealized human mind in the sense that the two have the same outputs. Note that SMT says that the idealized human mind knows that

[22] The informal proof of Claim One is as follows: Suppose $\mathbf{K}(Th(M_e) \subseteq \mathbf{Truth})$. Since it is knowable that $Th(M_e)$ is consistent, it is knowable that there is a true sentence of arithmetic which is not provable in $Th(M_e)$. So $Th(M_e) \subsetneq \mathbf{K}$.

[23] The informal proof of Claim Two is as follows: Suppose $Th(M_e) = \mathbf{K}$ for some e. Since $Th(M_e)$ is R.E. but Truth is not arithmetic, $\mathbf{K} \subsetneq \mathbf{Truth}$. So we can find some $\varphi \in \mathbf{Truth}$ but $\varphi \notin \mathbf{K}$ and $\neg\varphi \notin \mathbf{K}$.

[24] It is a little delicate to formalize GD in $\mathbf{EA_T}$ since $\mathbf{K}$ is formalized as an operator in $\mathbf{EA_T}$ and so we are prohibited from quantifying into it. For the details, we refer to Reinhardt (1986) and Koellner (2016; 2018a).

there is a Turing machine which coincides with the idealized human mind. Note that SSMT says that there is a particular Turing machine such that the idealized human mind knows that that particular machine coincides with the idealized human mind.

Suppose WMT holds. Then there exists an e^* such that in fact $\mathbf{K} = Th(M_{e*})$. It might seem at first that if we know that there is such an e^* then we will be able to find, in a computable way, the indices e such that $\mathbf{K} = Th(M_e)$. But this is an illusion, as demonstrated by Rice's Theorem, which we shall now explain.

In recursion theory, the sets $Th(M_e)$ are known as *computably enumerable* sets. Each such set is the domain of a partial computable function φ_e. Rice's Theorem states that for any class C of partial computable functions, $\{e : \varphi_e \in C\}$ is computable iff either $C = \emptyset$ or C is the class of all partial computable functions. Now consider the set of indices that we are interested in, namely, $\{e : \mathbf{K} = dom(\varphi_e)\}$, that is, $\{e : \varphi_e \in C\}$ where $C = \{\varphi_e : \mathbf{K} = dom(\varphi_e)\}$. It follows immediately from Rice's theorem that $\{e : \mathbf{K} = dom(\varphi_e)\}$ is not computable.

The following theorem shows that we can prove in $\mathbf{EA_T}$ that there does not exist a particular Turing machine such that the idealized human mind knows that that particular Turing machine coincides with the idealized human mind.

Theorem 4.3 (Reinhardt, 1985a)**.** $\mathbf{EA_T}$ + SSMT *is inconsistent.*

The following theorem shows that, from the viewpoint of $\mathbf{EA_T}$ it is possible that the idealized human mind is in fact a Turing machine. From Theorem 4.3, it just cannot know which one.

Theorem 4.4 (Reinhardt, 1985b)**.** $\mathbf{EA_T}$ + WMT *is consistent.*

From Theorem 4.4, the first disjunct is not provable in $\mathbf{EA_T}$. But Gödel did think that one day we would be in a position to prove the first disjunct, and what was missing, as he saw it, was an adequate resolution of the paradoxes involving self-applicable concepts like the concept of truth. Gödel thought that "[i]f one could clear up the intensional paradoxes somehow, one would get a clear proof that mind is not machine".[25]

The following technical theorem from Carlson shows that, from the point of view of $\mathbf{EA_T}$, it is possible that the idealized human mind knows that it is a Turing machine: it just cannot know which one.

Theorem 4.5 (Carlson, 2000)**.** $\mathbf{EA_T}$ + SMT *is consistent.*

Now we give a summary for the question whether Gödel's incompleteness theorems imply the first disjunct. The incompleteness theorems imply that

[25] This quotation is from Hao Wang's reconstruction of his conversations with Gödel (see Wang, 1996, p. 187).

$\neg\exists e\mathbf{K}(\mathbf{K} = Th(M_e))$. But from Theorem 4.4, it does not follow that $\neg\exists e(\mathbf{K} = Th(M_e))$; and from Theorem 4.5, it does not even follow that $\neg\mathbf{K}\exists e(\mathbf{K} = Th(M_e))$. The difference between $\exists e\mathbf{K}$ and $\mathbf{K}\exists e$ before $\mathbf{K} = Th(M_e)$ is essential. Assuming the principles embodied in $\mathbf{EA}_{\mathbf{T}}$, it is possible to know that we are a Turing machine (i.e. $\mathbf{K}\exists e(\mathbf{K} = Th(M_e))$); it is just not possible for there to be a Turing machine such that we know that we are that Turing machine (i.e. $\exists e\mathbf{K}(\mathbf{K} = Th(M_e))$).

Penrose proposed a new argument for the first disjunct in (1994, 2011). Penrose's new argument is the most sophisticated and promising argument for the first disjunct. It has been extensively discussed and carefully analyzed in the literature (see Chalmers, 1995; Feferman, 1995; Lindström, 2001; 2006; Shapiro, 1998; 2003; Gaifman, 2000; Koellner, 2016; 2018b, etc). The question of whether Penrose's new argument establishes the first disjunct is quite subtle. Penrose's new argument involves treating truth as type-free, and so for the analysis and formalization of Penrose's new argument, we need to employ type-free notions of truth. However, we now have many type-free theories of truth and there is no consensus as to which option is best. Koellner was the first to discuss Penrose's new argument in the context of type-free truth. And he shows that when one shifts to a type-free notion of truth then one can treat **K** as a predicate (as a contrast, in the context of **EA** and $\mathbf{EA}_{\mathbf{T}}$, **K** cannot be treated as a predicate).

In the literature, Koellner proposed the framework **DTK** which employs Feferman's type-free theory of determinate truth **DT** and some additional axioms governing **K** to the axioms of **DT**.[26] The following results about the system **DTK** are due to Koellner. From (Koellner, 2016; 2018b), **DTK** is consistent (see 2016, Theorem 7.14.1) and **DTK** proves GD (see 2016, Theorem 7.15.3). However, the particular argument Penrose gives for the first disjunct fails in the context of **DTK** (see 2018b, Theorem 4.1). Moreover, even if we restrict the first and second disjunct to arithmetic statements, **DTK** can neither prove nor refute either the first disjunct or the second disjunct (see 2016, Theorems 7.16.1–7.16.2). From the point of view of **DTK**, it is in principle impossible to prove or refute either disjunct. Koellner concluded that

> Since the statements that "the mind cannot be mechanized" and "there are absolutely undecidable statements" are independent of the natural principles governing the fundamental concepts and, moreover, are independent of any plausible principles in sight, it seems likely that these statements are themselves "absolutely undecidable". (Koellner, 2018b, p. 469)[27]

[26] For the details of the system **DT** and **DTK**, see (Koellner, 2016; 2018b).

[27] Koellner concluded in (2018b, p. 480) with a disjunctive conclusion of his own: "Either the statements that 'the mind cannot be mechanized' and 'there are absolutely undecidable statements' are indefinite (as the philosophical critique maintains) or they are definite and the above results and considerations provide evidence that they are about as good examples of 'absolutely undecidable' propositions as one might find".

In our previous discussion of GD, the first disjunct and the second disjunct, we identified absolutely undecidability with knowability of the idealized human mind and define that φ is absolutely undecidable if $\varphi \notin \mathbf{K}$ and $\neg\varphi \notin \mathbf{K}$. Under this framework, the second disjunct is equivalent to "**K** is not complete". Under the assumption that $\mathbf{K} \subseteq \mathbf{Truth}$, the second disjunct is equivalent to "$\mathbf{K} \subsetneq \mathbf{Truth}$". However, G1 only tells us that $\mathbf{Prov} \subsetneq \mathbf{Truth}$, and it does not tell us that $\mathbf{K} \subsetneq \mathbf{Truth}$.

Another natural informal definition of absolutely undecidability is: φ is absolutely undecidable if there is no consistent extension T of **ZFC** with well-justified axioms such that φ is provable in T. In this paper, we focus on whether Gödel's incompleteness theorem implies that the human mind cannot be mechanized. In philosophy of set theory, there are extensive discussions about whether there exists an absolutely undecidable statement in set theory. For a detailed discussion of the question of absolutely undecidability in set theory and especially whether the Continuum Hypothesis is absolutely undecidable, we refer to Koellner (2006).

5. Gödel's Undemonstrability of Consistency Thesis and the Definability of Natural Numbers

In Section 8, Krajewski (2020) discussed two consequences of Gödel's incompleteness theorem directly related to the Anti-Mechanist Arguments: Gödel's Undemonstrability of Consistency Thesis and the undefinability of natural numbers. For us, Krajewski's discussion on these two consequences is mainly philosophical and not very precise. In this section, we want to give a more precise logical analysis of Gödel's Undemonstrability of Consistency Thesis and the undefinability of natural numbers.

Let us first examine the definability of natural numbers. As a consequence of Gödel's incompleteness theorem, Krajewski (2020) claimed that we can not define the natural numbers in the sense that there is not a complete axiomatic system which fully characterizes all truths about natural numbers. We give some supplementary notes to make this point more precise.

Firstly, whether a theory about natural numbers is complete depends on the language of the theory. In the languages $L(\mathbf{0}, \mathbf{S})$, $L(\mathbf{0}, \mathbf{S}, <)$ and $L(\mathbf{0}, \mathbf{S}, <, +)$, there are, respectively, recursively axiomatized complete arithmetic theories (see Enderton, 2001, Section 3.1–3.2). For example, Presburger arithmetic is a complete theory of the arithmetic of addition in the language $L(\mathbf{0}, \mathbf{S}, +)$ (see Murawski, 1999, Theorem 3.2.2, p. 222). However, if a recursively axiomatized theory contains enough information about addition and multiplication, then it is incomplete and hence it must miss some truths about arithmetic. For example, any recursively axiomatized consistent extension of **Q** is incomplete. Thus, in Krajewski's sense, we can not define the natural numbers in any recursively axiomatized consistent extension of **Q**.

Secondly, if we discuss the definability of a set with respect to a structure, then the definability of natural numbers depends on the structure we talk about. It is well known that $\mathbb{N}$ is definable in $(\mathbb{Z}, +, \cdot)$ and $(\mathbb{Q}, +, \cdot)$ (Epstein, 2011, Chapter XVI), and $Th(\mathbb{N}, +, \cdot)$ is interpretable in $Th(\mathbb{Z}, +, \cdot)$ and $Th(\mathbb{Q}, +, \cdot)$. Since $Th(\mathbb{N}, +, \cdot)$ is undecidable,[28] by Theorem 2.1, $Th(\mathbb{Z}, +, \cdot)$ and $Th(\mathbb{Q}, +, \cdot)$ are all undecidable and hence not recursive axiomatizable. But $Th(\mathbb{R}, +, \cdot)$ is a decidable, recursively axiomatizable theory (even if not finitely axiomatizable) and $Th(\mathbb{R}, +, \cdot) =$ **RCF** (the theory of real closed field; see Epstein, 2011, p. 320–321). As a corollary, $\mathbb{N}$ is not definable in the structure $\langle\mathbb{R}, +, \cdot\rangle$ (if $\mathbb{N}$ is definable in $\langle\mathbb{R}, +, \cdot\rangle$, then $Th(\mathbb{N}, +, \cdot)$ is interpretable in $Th(\mathbb{R}, +, \cdot)$ and thus, by Theorem 2.1, $Th(\mathbb{R}, +, \cdot)$ is undecidable which leads to a contradiction). In summary, if we consider matters of definability relative to the base structure, then whether the set of natural numbers is definable depends on the base structure: $\mathbb{N}$ is definable in $(\mathbb{Z}, +, \cdot)$ and $(\mathbb{Q}, +, \cdot)$, but $\mathbb{N}$ is not definable in $\langle\mathbb{R}, +, \cdot\rangle$.

Now we examine Gödel's Undemonstrability of Consistency Thesis (i.e. G2). The intensionality of Gödel sentence and the consistency sentence has been widely discussed in the literature (e.g. Feferman, 1960; Halbach, Visser, 2014a; 2014b; Visser, 2011). Halbach and Visser examined the sources of intensionality in the construction of self-referential sentences of arithmetic in (2014a; 2014b) and argued that corresponding to the three stages of the construction of self-referential sentences of arithmetic, there are at least three sources of intensionality: coding, expressing a property and self-reference. Visser (2011) located three sources of indeterminacy in the formalization of a consistency statement for a theory *T*:

(I) the choice of a proof system;

(II) the choice of a way of numbering;

(III) the choice of a specific formula numerating the axiom set of *T*.

In summary, the intensional nature ultimately traces back to the various parameter choices that one has to make in arithmetizing the provability predicate. That is the source of both the intensional nature of the Gödel sentence and the consistency sentence.

For a consistent theory *T*, we say that G2 holds for *T* if the consistency statement of *T* is not provable in *T*. However, this definition is vague, and whether G2 holds for *T* depends on how we formulate the consistency statement. We refer to this phenomenon as the intensionality of G2. Both mathematically and philosophically, G2 is more problematic than G1. The difference between G1 and G2 is that in the case of G1 we are mainly interested in the fact that it shows that some sentence is undecidable if **PA** is ω-consistent. We make no claim to the effect that that sentence "really" expresses what we would express by saying "**PA**

[28] I.e. there does not exist an effective algorithm such that given any sentence φ in $L(\mathbf{PA})$, we can effectively decide whether $(\mathbb{N}, +, \cdot) \vDash \varphi$ or not.

cannot prove this sentence".[29] But in the case of G2 we are also interested in the content of the statement.

The status of G2 is essentially different from G1 due to the intensionality of G2. We can say that G1 is extensional in the sense that we can construct a concrete independent mathematical statement without referring to arithmetization and provability predicate. However, G2 is intensional and "whether G2 holds for *T*" depends on varied factors as we will discuss.

In the following, we give a very brief discussion of the intensionality of G2 (for more details, we refer to Cheng, in press). In this section, unless otherwise stated, we make the following assumptions:

(1) The theory *T* is a recursively axiomatized consistent extension of **Q**;

(2) The canonical arithmetic formula to express the consistency of *T* is $\mathbf{Con}(T) \triangleq \neg\mathbf{Pr}_T(\ulcorner \mathbf{0} \neq \mathbf{0} \urcorner)$;

(3) The canonical numbering we use is Gödel's numbering;

(4) The provability predicate we use is standard;

(5) The formula numerating the axiom set of *T* is Σ^0_1.

Based on works in the literature, we argue that "whether G2 holds for *T*" depends on the following factors:

(1) the choice of the base theory *T*;

(2) the choice of a provability predicate;

(3) the choice of an arithmetic formula to express consistency;

(4) the choice of a numbering;

(5) the choice of a specific formula numerating the axiom set of *T*.

These factors are not independent of each other, and a choice made at an earlier stage may have influences on the choices made at a later stage. In the following, when we discuss how G2 depends on one factor, we always assume that other factors are fixed as in the default assumptions we make and only the factor we are discussing is varied. For example, Visser (2011) rests on fixed choices for (1) and (3)–(5) but varies the choice of (2); Grabmayr (2020) rests on fixed choices for (1)–(2) and (4)–(5) but varies the choice of (3); Feferman (1960) rests on fixed choices for (1)–(4) but varies the choice of (5).

In the following, we give a brief discussion of how G2 depends on the above five factors. For more discussions of these factors, we refer to (Cheng, in press).

"Whether G2 holds for *T*" depends on the choice of the base theory. A foundational question about G2 is: how much of information about arithmetic is re-

[29] I would also like to thank the referee for pointing out this difference between G1 and G2.

quired for the proof of G2. If the base theory does not contain enough information about arithmetic, then G2 may fail in the sense that the consistency statement is provable in the base theory. Willard (2006) explored the generality and boundary-case exceptions of G2 under some base theories. Willard constructed examples of recursively enumerable arithmetical theories that couldn't prove the totality of successor function but could prove their own canonical consistency (see Willard, 2001; 2006). Pakhomov (2019) defined a theory $H_{<\omega}$ and showed that it proves its own canonical consistency. Unlike Willard's theories, $H_{<\omega}$ isn't an arithmetical theory but a theory formulated in the language of set theory with an additional unary function.

"Whether G2 holds for T" depends on the definition of provability predicate. Recall that T is a recursively axiomatizable consistent extension of $\mathbf{Q}$. Being a consistency statement is not an absolute concept but a role w.r.t. a choice of the provability predicate. Note that G2 holds for any standard provability predicate in the sense that if provability predicate $\mathbf{Pr}_T(x)$ is standard, then $T \nvdash \neg\mathbf{Pr}_T(\ulcorner \mathbf{0} \neq \mathbf{0}\urcorner)$. However, G2 may fail for some nonstandard provability predicates. Rosser provability predicate is an important kind of non-standard provability predicate in the study of meta-mathematics of arithmetic. Define the Rosser provability predicate $\mathbf{Pr}^R_T(\mathrm{x})$ as the formula $\exists y(\mathbf{Prf}_T(x,y) \wedge \forall z \leq y \neg \mathbf{Prf}_T(\dot{\neg}(x), z))$.[30] Define the consistency statement $\mathbf{Con}^R(T)$ via Rosser provability predicate as $\neg\mathbf{Pr}^R_T(\ulcorner \mathbf{0} \neq \mathbf{0}\urcorner)$. Then G2 fails for Rosser provability predicate: $T \vdash \mathbf{Con}^R(T)$.

"Whether G2 holds for T" depends on the choice of arithmetic formulas to express consistency. We have different ways to express the consistency of T. The canonical arithmetic formula to express the consistency of T is $\mathbf{Con}(T) \triangleq \neg\mathbf{Pr}_T(\ulcorner \mathbf{0} \neq \mathbf{0}\urcorner)$. Another way to express the consistency of T is $\mathbf{Con}^0(T) \triangleq \forall x(\mathbf{Fml}(x) \wedge \mathbf{Pr}_T(x) \rightarrow \neg\mathbf{Pr}_T(\dot{\neg} x))$.[31]

Kurahashi (2019) constructed a Rosser provability predicate such that G2 holds for the consistency statement formulated via $\mathbf{Con}^0(T)$ (i.e. the consistency statement formulated via $\mathbf{Con}^0(T)$ and Rosser provability predicate is not provable in T), but G2 fails for the consistency statement formulated via $\mathbf{Con}(T)$ (i.e. the consistency statement formulated via $\mathbf{Con}(T)$ and Rosser provability predicate is provable in T).

"Whether G2 holds for T" depends on the choice of numberings. Any injective function γ from a set of $L(\mathbf{PA})$-expressions to ω qualifies as a numbering. Gödel's numbering is a special kind of numberings under which the Gödel number of the set of axioms of $\mathbf{PA}$ is recursive. Grabmayr (2020) showed that G2 holds for acceptable numberings; But G2 fails for some nonacceptable numberings.[32]

Finally, "Whether G2 holds for T" depends on the numeration of T. As a generalization, G2 holds for any Σ^0_1 numeration of T: if $\alpha(x)$ is a Σ^0_1 numeration of T,

[30] $\dot{\neg}$ is a function symbol expressing a primitive recursive function calculating the code of $\neg\varphi$ from the code of φ.

[31] $\mathbf{Fml}(x)$ is the formula which represents the relation that x is a code of a formula.

[32] For the definition of acceptable numberings, we refer to (Grabmayr, 2020).

then $T \nvdash \mathbf{Con}_\alpha(T)$. However, G2 fails for some Π^0_1 numerations of T. For example, Feferman (1960) constructed a Π^0_1 numeration $\tau(u)$ of T such that G2 fails under this numeration: $T \vdash \mathbf{Con}_\tau(T)$.

REFERENCES

Beklemishev, L. D. (2010). Gödel Incompleteness Theorems and the Limits of Their Applicability I. *Russian Math Surveys*, *65*(5), 857–899.

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

Boolos, G. (1993). *The Logic of Provability*. Cambridge: Cambridge University Press.

Carlson, T. J. (2000). Knowledge, Machines, and the Consistency of Reinhardt's Strong Mechanistic Thesis. *Annals of Pure and Applied Logic*, *105*(1–3), 51–82.

Chalmers, D. J. (1995). Minds, Machines, and Mathematics: A Review of Shadows of the Mind by Roger Penrose. *Journal Psyche*, *2*.

Cheng, Y. (2019). *Incompleteness for Higher-Order Arithmetic: An Example Based on Harrington's Principle* (Springer series: Springerbrief in Mathematics). New York: Springer.

Cheng, Y. (2020). Finding the Limit of Incompleteness I. *The Bulletin of Symbolic Logic*. Retrieved from: https://arxiv.org/pdf/1902.06658.pdf

Cheng, Y. (in press). Current Research on Gödel's Incompleteness Theorem. *The Bulletin of Symbolic Logic*.

Enderton, H. B. (2001). *A Mathematical Introduction to Logic* (2nd ed.). Boston, MA: Academic Press.

Epstein, R. L. (With contributions by Lesław W. Szczerba). (2011). *Classical Mathematical Logic: The Semantic Foundations of Logic*. Princeton, New Jersey: Princeton University Press.

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

Feferman, S. (1995). Penrose's Gödelian Argument: A Review of Shadows of the Mind by Roger Penrose. *Journal Psyche*, *2*.

Feferman, S. (2009). Gödel, Nagel, Minds, and Machines. *The Journal of Philosophy*, *106*(4), 201–219.

Gaifman, H. (2000). What Gödel's Incompletness Result Does and Does Not Show. *The Journal of Philosophy*, *97*(8), 462–470.

Gödel, K. (1931). Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I. *Monatsh. Math. Phys*. *38*(1), 173–198.

Gödel, K. (1986). An Interpretation of the Intuitionistic Propositional Calculus. In S. Feferman et. al. (Eds.), *Collected Works, Volume I: Publications 1929–1936* (pp. 301–303). Oxford University Press.

Gödel, K. (1951). Some Basic Theorems on the Foundations of Mathematics and Their Implications. In S. Feferman et. al. (Eds.), *Collected Works, Volume III: Unpublished Essays and Lectures* (pp. 304–323). Oxford University Press.

Gödel, K. (1995). *Collected Works, Volume III: Unpublished Essays and Lectures*. New York: Oxford University Press.

Grabmayr, B. (2020). On the Invariance of Gödel's Second Theorem With Regard to Numberings. *The Review of Symbolic Logic*. Retrieved from: https://arxiv.org/pdf/1803.08392.pdf

Hájek, P., Pudlák, P. (1993). *Metamathematics of First-Order Arithmetic*. Berlin-Heidelberg-New York: Springer-Verlag.

Halbach, V., Visser, A. (2014a). Self-Reference in Arithmetic I (2014a). *Review of Symbolic Logic*, *7*(4), 671–691.

Halbach, V., Visser, A. (2014b). Self-Reference in Arithmetic II (2014b). *Review of Symbolic Logic*, *7*(4), 692–712.

Horsten, L. (1998). In Defense of Epistemic Arithmetic. *Synthese*, *116*(1), 1–25.

Horsten, L., Welch, P. (2016). *Gödel's Disjunction: The Scope and Limits of Mathematical Knowledge.* Oxford University Press.

Koellner, P. (2006). On the Question of Absolute Undecidability. *Philosophia Mathematica*, *14*(2), 153–188.

Koellner, P. (2016). Gödel's Disjunction. In L. Horsten, P. Welch (Eds.), *Gödel's Disjunction: The Scope and Limits of Mathematical Knowledge*. Oxford University Press.

Koellner, P. (2018a). On the Question of whether the Mind Can Be Mechanized. Part 1: From Gödel to Penrose. *Journal of Philosophy*, *115*(7), 337–360.

Koellner, P. (2018b). On the Question of Whether the Mind Can Be Mechanized. Part 2: Penrose's New Argument. *Journal of Philosophy*, *115*(9), 453–484.

Kotlarski, H. (2004). The Incompleteness Theorems After 70 Years. *Annals of Pure and Applied Logic*, *126*, 125–138.

Krajewski, S. (2020). On the Anti-Mechanist Arguments Based on Gödel Theorem. *Studia Semiotyczne*, *34*(1), 9–56.

Kurahashi, T. (2019). Rosser Provability and the Second Incompleteness Theorem. Retrieved from: https://arxiv.org/pdf/1902.06863.pdf

Leitgeb, H. (2009). On Formal and Informal Provability. In O. Bueno, Ø. Linnebo (Eds.), *New Waves in Philosophy of Mathematics* (pp. 263–299). London: Palgrave Macmillan.

Lindström, P. (1997). *Aspects of Incompleteness*. Cambridge University Press.

Lindström, P. (2001). Penrose's New Argument. *Journal of Philosophical Logic*, *30*, 241–250.

Lindström, P. (2006). Remarks on Penrose's New Argument. *Journal of Philosophical Logic*, *35*, 231–237.

Lucas, J. R. (1961). Minds, Machines, and Gödel. *Philosophy*, *36*, 120–124.

Lucas, J. R. (1996). Minds, Machines, and Gödel: A Retrospect. In P. J. R. Millican, A. Clark (Eds.), *Machines and Thought: The Legacy of Alan Turing* (vol. 1, pp. 103–124). Oxford University Press.

Montague, R. (1963). Syntactical Treatments of Modality, With Corollaries on Reflexion Principles and Finite Axiomatizability. *Acta Philosophica Fennica*, *16*, 153–167.

Murawski, R. (1999). *Recursive Functions and Metamathematics: Problems of Completeness and Decidability, Gödel's Theorems*. Heidelberg: Springer Netherlands.

Myhill, J. (1960). Some Remarks on the Notion of Proof. *Journal of Philosophy*, *57*(14), 461–471.

Nagel, E., Newman, J. R. (2001). *Gödel's Proof*. New York University Press.

Pakhomov, F. (2019). A Weak Set Theory That Proves Its Own Consistency. Retrieved from: https://arxiv.org/pdf/1907.00877.pdf

Pudlák, P. (1999). A Note on Applicability of the Incompleteness Theorem to Human Mind. *Annals of Pure and Applied Logic*, *96*, pp. 335–342.

Penrose, R. (1989). *The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics*. Oxford University Press.

Penrose, R. (1994). *Shadows of the Mind: A Search for the Missing Science of Consciousness*. Oxford University Press.

Penrose, R. (2011). Gödel, the Mind, and the Laws of Physics. In M. Baaz, Ch. H. Papadimitriou, H. Putnam, D. S. Scott, Ch. L. Harper (Eds.), *Kurt Gödel and the Foundations of Mathematics: Horizons of Truth* (pp. 339–358). Cambridge University Press.

Putnam, H. (1960). Minds and Machines. In S. Hood (Ed.), *Dimensions of Mind: A Symposium* (pp. 138–164). New York University Press.

Reid, C. (1996). *Hilbert*. Springer.

Reinhardt, W. N. (1985). Absolute Versions of Incompleteness Theorems. *Noûs*, *19*(3), 317–346.

Reinhardt, W. N. (1985). The Consistency of a Variant of Church's Thesis With an Axiomatic Theory of an Epistemic Notion. *Revista Colombiana de Matematicas*, *19*, 177–200.

Reinhardt, W. N. (1986). Epistemic Theories and the Interpretation of Gödel's Incompleteness Theorems. *Journal of Philosophical Logic*, *15*, 427–474.

Shapiro, S. (1985). Epistemic and Intuitionistic Arithmetic. In S. Shapiro (Ed.), *Intensional Mathematics* (vol. 113 of Studies in Logic and the Foundations of Mathematics, pp. 11–46). Amsterdam: North-Holland.

Shapiro, S. (1998). Incompleteness, Mechanism, and Optimism. *The Bulletin of Symbolic Logic*, *4*(3), 273–302.

Shapiro, S. (2003). Mechanism, Truth, and Penrose's New Argument. *Journal of Philosophical Logic*, 32(1), 19–42.

Smith, P. (2007). *An Introduction to Gödel's Theorems*. Cambridge University Press.

Smoryński, C. (1977). The Incompleteness Theorems. In J. Barwise (Ed.), *Handbook of Mathematical Logic* (pp. 821–865). Amsterdam: North-Holland.

Tarski, A., Mostowski, A., Robinson, R. M. (1953). *Undecidabe Theories* (Studies in Logic and the Foundations of Mathematics). Amsterdam: North-Holland.

Thomason, R. H. (1980). A Note on Syntactical Treatments of Modality. *Synthese*, *44*, 391–395.

Visser, A. (2011). Can We Make the Second Incompleteness Theorem Coordinate Free? *Journal of Logic and Computation*, *21*(4), 543–560.

Visser, A. (2016). The Second Incompleteness Theorem: Reflections and Ruminations. In L. Horsten, P. Welch (Eds.), *Gödel's Disjunction: The Scope and Limits of Mathematical Knowledge* (pp. 67–91). Oxford University Press.

Wang, H. (1996). *A Logical Journey: From Godel to Philosophy*. MIT Press.

Willard, D. E. (2001). Self-Verifying Axiom Systems, the Incompleteness Theorem and Related Reflection Principles. *Journal of Symbolic Logic*, *66*(2), 536–596.

Willard, D. E. (2006). A Generalization of the Second Incompleteness Theorem and Some Exceptions to It. *Ann. Pure Appl. Logic*, *141*(3), 472–496.

Williamson, T. (2016). Absolute Provability and Safe Knowledge of Axioms. In L. Horsten, P. Welch (Eds.), *Gödel's Disjunction: The Scope and Limits of Mathematical Knowledge* (pp. 243–253). Oxford University Press.

Zach, R. (2007). Hilbert's Program Then and Now. In D. Jacquette (Ed.), *Philosophy of Logic* (pp. 411–447). Amsterdam: North Holland.

STUDIA SEMIOTYCZNE (SEMIOTIC STUDIES), t. XXXIV, nr 1 (2020), s. 183–202
ISSN 0137-6608, e-ISSN 2544-073X
DOI: 10.26333/sts.xxxiv1.08
Article

ŠTĚPÁN HOLUB [*]

UNDERSTANDING, EXPRESSION AND UNWELCOME LOGIC

SUMMARY: In this paper I will attempt to explain why the controversy surrounding the alleged refutation of Mechanism by Gödel's theorem is continuing even after its unanimous refutation by logicians. I will argue that the philosophical point its proponents want to establish is a necessary gap between the intended meaning and its formulation. Such a gap is the main tenet of philosophical hermeneutics. While Gödel's theorem does not disprove Mechanism, it is nevertheless an important illustration of the hermeneutic principle. The ongoing misunderstanding is therefore based in a distinction between a metalogical illustration of a crucial feature of human understanding, and a logically precise, but wrong claim. The main reason for the confusion is the fact that in order to make the claim logically precise, it must be transformed in a way which destroys its informal value. Part of this transformation is a clear distinction between the Turing Machine as a mathematical object and a machine as a physical device.

KEYWORDS: mechanism, Gödel's theorem, Turing machine, hermeneutics.

The controversy surrounding the alleged refutation of Mechanism by Gödel's theorem is hard to approach. The discrepancy between the fact that the argument has been rigorously and unanimously rejected by logicians on one hand and the fact that proponents[1] are still defending it on the other hand, is striking. It indicates a deeper misunderstanding entrenched in the controversy. The verdict of

[*] Charles University, Faculty of Mathematics and Physics. E-mail: holub@karlin.mff.cuni.cz. ORCID: 0000-0002-6169-5139.

[1] By the term "proponents" (of the anti-Mechanist thesis) I will refer to thinkers who claim that Gödel's results refute Mechanism, mainly to John Lucas and Roger Penrose.

logicians was succinctly formulated by Hilary Putnam (1975a, p. 366): "misapplication of Gödel's theorem, pure and simple". The same critic later rejected a variant of the argument as a "sad episode in our current intellectual life" (Putnam, 1994). A more polite version of the same conclusion is the one by Stewart Shapiro (1998, p. 275): "My conclusion (perhaps slightly exaggerated) is that there is no plausible mechanist thesis on offer that is sufficiently precise to be undermined by the incompleteness theorems".

Nevertheless, the idea keeps provoking thinkers who again and again rush to add their take in the spirit of the opening sentence of John Lucas's original paper (1961, p. 112): "Gödel's Theorem seems to me to prove that Mechanism is false, that is, that minds cannot be explained as machines". Lucas (2011) himself remained unimpressed by the criticism: "Since many critics are unaware of the argument, and are unlikely to look back at papers published some time ago, it is worth articulating the argument afresh". Lucas is willing to reiterate his feeling and he obviously believes that his extant answer to objections is sufficient. Apparently, something more is needed than Stewart Shapiro's (1998, p. 273) "modest aim of forging connections between different parts of [the] literature and clearing up some confusions, together with the less modest aim of not introducing any more confusions".

In this paper I will attempt to explain why proponents ignore logical arguments, and I will argue that they in fact want to establish a philosophical point which is not directly related to Mechanism. Instead they have in mind a fundamental feature of understanding which is the core tenet of philosophical hermeneutics, namely the insurmountable gap between any intended meaning and any of its formulations. The intention always contains more than what is captured by the expression. The proof of Gödel's theorem is particularly interesting specimen of this gap, in this case between our understanding of natural numbers, and its expression by a logical theory like Peano Arithmetic. However, this is a metalogical observation which does not disprove Mechanism, since it does not exclude the possibility that the human mind can be simulated by a formal system which absolutely surpasses human understanding, and therefore lies in a realm where no hermeneutics can be applied. Whether admitting such a formal system is reasonable or not becomes a futile argument without clear criteria. The main mistake of the proponents is that they appeal to a logical result which, as such, has no bearing on the dispute. They do not realize that in order to even make their claim intelligible, it is necessary to translate it into a logical language, whereby the proper hermeneutic insight is lost.

One of the first aspects of the translation is the definition of the Turing Machine which is equivalent to the recursive formal system. The difference between Turing Machines and their instantiations is stressed in the first section. The second section summarizes the discussion after the anti-Mechanist claim is properly formulated in logical terms and explains why its validity cannot be established. The third chapter develops the argument that the main feature of human mind that proponents want to highlight can be best described in terms of philosophical hermeneutics. The chapter also explains why the hermeneutic feature cannot be suitably captured within the framework of logic. The fourth section analyzes the proof of Gödel's theorem in order to illustrate how an informal, that is, ordinary mathematical understanding of natural numbers is crucially responsible for its validity. This prompts considerations about the mutual dependence between formal and informal aspects of logic. The conclusion then points out the main lesson that can be learned from the curious discussion about Mechanism and Gödel's theorem, namely the need to respect methodical limitations of scientific disciplines.

1. Turing Machines and Robots

One of the standard reproaches against the anti-Mechanist thesis is that it is based on one of several problematic "idealizations" (Shapiro, 1998; Feferman, 2009). The main target is the "idealized human mind", whereas the idealization of a machine resulting in the concept of the Turing Machine is usually considered unproblematic, and the fact that the Turing Machine is not just a machine with infinite space and a flawless processor is easily underestimated.[2] However, only the idealized concept of the Turing Machine makes the proponents' argument possible because it corresponds to a recursive formal system in Gödel's proof. The "Turing Machine" is a mathematical entity in pretty much the same way as a natural number. The most important part of its definition is the transition relation which in turn is a finite set of quintuples, called instructions. The definition is then extended to a relation between "instantaneous descriptions", which, if the relation is deterministic, finally defines a (partial) map from natural numbers to themselves, called a recursively enumerable function. The way from "honest machines" to Turing Machines is therefore not a short and simple one. In the same way in which natural numbers are the mathematical conceptualization of discrete quantity, the Turing Machine is a mathematical conceptualization of a fully controlled process. We call such a process "mechanical" but that is only a metaphor.

It is worth noting that Turing (1937) originally used the word "computer" as a reference to a diligent and fully reliable clerk. Therefore, calling our electronic devices "computers" is just one, and an almost forgotten one, among anthropomorphic expressions (like "memory") we use for (electronic) devices. When we ask whether the human mind is a computer, it is therefore a kind of reversed metaphor, which actually asks whether the behavior of the mind can be exhaustively described as the activity of a diligent and reliable clerk. The precise sense of what "diligent and reliable" means is then mathematically captured by the notion of the Turing Machine. The nature of a "fully controlled process" is inde-

[2] Although already Shapiro correctly observed that the idealization performed in the definition of the Turing Machine is "similar to idealizations made throughout mathematics" (Shapiro, 2009, p. 275).

pendent of its realization, be it by the human mind or by a machine. The Turing Machine is therefore actually no machine in the strict sense. The frivolous use of metaphors concerning machines is an important contribution to the confusion surrounding these issues. When, for example, Paul Benaceraf (1967) calls the Turing Machine in question Maud, and notes that "she convinced herself [...] of her own consistency" but that she "she shouldn't go around blabbing it", we may see it as a refreshing stylistic feature. However, the concept of the Turing Machine is so relatively recent (compared to, say, the concept of natural number), and so deeply connected to the idea of a "tape", and a "processing head", that it is difficult for many, at least for a majority of university students, to fully appreciate the fact that the Turing Machine is actually a mathematical object, even after they become familiar with other mathematical concepts and start to understand, for example, that working with a five-dimensional vector space does not require any magical sensory ability.

While physical computers people build are instantiations of intended Turing Machines (if nothing goes wrong), the opposite is much less obvious, and that is where metaphors may betray us. For example, I do not see any good reason for Krajewski's claim (2020, p. 35) that we would "have no doubt" that the robot Luke, a fictional result of a long robotic evolutional process, is a Turing Machine. Actually, as shown above, such a claim does not even make good sense. Neither is it clear to me how Luke's program (which is the Turing Machine we speak about) would be "investigated by human computer scientists" (p. 40). Two completely different problems are conflated here. The first one is how to obtain the "program" from a physical device (including a brain) at our disposal. That is, how to describe the behavior of the device by a finite set of states, and by a similarly finite set of transition rules governing their evolution conditioned by a finite set of possible instantaneous inputs. This problem, in addition to being close to hopeless, is not even remotely related to Gödel's theorem. Only then comes the additional question, namely whether we can understand what the Turing Machine obtained in the first phase "does", that is, to derive some relevant properties of the partial recursive function it defines. Even this is a daunting task, but at least it is somewhat related to Gödel's work.

If we call life-simulating artificial products "robots", we can then say that the problem is an inadvertent identification of robots with Turing Machines. While the question whether the Turing Machine can become conscious is as nonsensical as whether a sufficiently large natural number can, the question whether robots can eventually acquire mind is completely open, or at least it is a question which has hardly anything to do with formal logic.[3] Hilary Putnam dedicated several

[3] The anti-utopia drama *R.U.R.* in which Karel Čapek coined the word "robot" depicts the creation of robots as an invention on the chemical level. The robots are used as a labor force, and the question of computation is not particularly stressed. One of the "humanizing" aspects is the deliberate introduction of pain into their functioning, and the distinctive human feature, which robots eventually develop, is love, not understanding of Gödel sentences.

papers to the relation between minds, robots and Turing Machines, where he makes a similar point many times. Even in papers where he defended the thesis that “we are Turing Machines” he makes clear that the identity has to be understood as a “functional isomorphism” depending on a description of conscious life in terms of a finite set of discrete psychological states. What is at stake is a “functional organization”, not “physical realization” (Putnam, 1975a, p. 373). Moreover, reflecting later on the implied condition of the existence of discrete states describing human experience, Putnam admitted—citing reasons one is tempted to describe as common sense—that his earlier “point of view was essentially wrong” (Putnam, 1975b, p. 298).

2. Why Proponents Are Wrong

Keeping in mind that we speak about recursive functions, not about robots, the intuitive appeal of the question is undoubtedly reduced for a non-mathematician or even a mathematician who is not a logician. Perhaps, its appeal should be reduced for the proponents themselves. In any case, the very meaning of the question now requires better clarification. What could it mean that minds can, or cannot, be explained as Turing Machines? The question must be reformulated in terms of the mind’s output. Namely, the question becomes whether the set of all arithmetical propositions the human mind can in principle prove is or is not recursive. When pressed about the use of the incompleteness theorem in their claim, the proponents are therefore eventually forced to resort to a purely syntactic competition between the mind and the machine. The machine and the human mind will each produce sentences in a given formal system, and the human mind will always win by producing a true sentence (the Gödelian one) which the machine never will, unless the system is inconsistent. As Krajewski stresses (2020, p. 11), the content of the competition can be even reduced to establishing the solvability of Diophantine equations.[4] Since the precise conditions of this competition remain chronically unclear,[5] the focus turns to specifications of the idealized human mind. This necessarily leads to a construction of some abstract concept, ultimately mathematical but often accompanied by some playful theo-

[4] I found confusing, in this respect, the numerous remarks Krajewski makes about alleged circularity. For example, he says: “we should beware of a circularity: if we simply assume that the mind, which is self-conscious, does not operate according to […] rules, then we assume what we are supposed to prove by Lucas’s argument, and the whole business with Gödel’s theorem is superfluous” (2020, p. 19). Why so? Lucas’s argument is that it can be shown beyond doubt from Gödel’s theorem that the mind can outperform any Turing Machine in the field of solving Diophantine equations. This is independent of what we assume about the mind otherwise.

[5] Lucas’s (2011) metaphor of the dispute against the mechanist in terms of the Oxford First and Second Public Examinations is just one example of how unclear it is.

logical terminology.[6] The discussion is already loaded by two dangerous ambiguities concerning machines (Turing Machines vs. robots) and the human mind (understanding vs. output).

Three basic technical facts govern the discussion. First, the most basic problem for the anti-Mechanist application of Gödel's theorem is the impossibility of proving the consistency of the considered system within the system itself (the impossibility is shown by the second incompleteness theorem). It is therefore not sufficient for proponents of human superiority to construct the Gödelian sentence, they first have to be able to show that the system is consistent, which is far from granted. This fact was quickly pointed out by many critics (it is also behind Putnam's "misapplication" remark), and it became one of the main points of contention. The second difficulty for the anti-Mechanist claim is that the construction of the independent Gödel sentence is itself algorithmic. That is, it can be performed by a suitable algorithm, although not the one corresponding to the examined theory. This leads to an infinite chase between Turing Machines, each new one "out-Gödeling" the previous one and being "out-Gödeled" by the next one. The third and technically most involved fact is a partial and final concession to proponents, called "Gödel's disjunction". It claims that either "mind is not a machine", that is, the set of sentences knowable by the "idealized human mind" is nonrecursive, or, if after all such a set is recursive, then the corresponding Turing Machine cannot be known, which in particular means that there are "absolutely unknowable" mathematical truths. This observation dates back to Gödel's own reflections on the matter which are often ridiculed for their perceived naïve "Platonism", but which nevertheless show both prudence and perspicacity concerning logical facts. Gödel's disjunction has proven to be a solid logical fact. Most important, it turns out that the second possibility, which represents a version of Mechanism, cannot be excluded by logical means. The technical layer of the literature on this provides a large variety of advanced and very interesting results in this direction, effectively warranting Shapiro's (cited above) "slightly exaggerated" informal conclusion. Moreover, the conclusion is shown not to be exaggerated at all in particular by the results presented in recent papers by Peter Ko-

[6] See the title of Benaceraf's paper (1967) or the skeptical remark of Peter Koellner (2018b, p. 476) about the "angelic mind". Shapiro (1998, p. 273) mocks this terminological manner when he writes: "A descriptive title for this paper would be 'Gödel, Lucas, Penrose, Turing, Feferman, Dummett, mechanism, optimism, reflection, and indefinite extensibility'. Adding 'God and the Devil' would probably be redundant". On the other hand, Peter Vopěnka entertained seriously the idea that the concept of god (or God) and his capabilities with respect to infinity helps to explain different conceptions of mathematics. The antic gods, corresponding to Christian angels, are able to see easily as small (or large) quantities as they wish, however always with the possibility to go deeper. The Christian God is on the contrary able to see the whole set of natural numbers or the absolute geometric point in one shot. This is obviously a variant of potential and actual infinity. However, Vopěnka both suggests that medieval theology directly influenced modern mathematics, and tries to use the theological explanation as a common sense basis for the non-standard analysis and for its practical use (cf., for example, Vopěnka, 2010).

ellner (2018a; 2018b). Both the strength and the weakness of such technical results is that they are, by definition, results about some formalized versions of the Mechanist claim. Although the details are sophisticated, the nature of the results is fairly straightforward. Since provability has its precise technical meaning, it remains to identify formal counterparts for truth and knowability. This requires the introduction of predicates or operators T and K, to formulate suitable axioms for them, and then to show, by standard (or rather advanced) logical means corresponding facts, namely the relative (in)consistency of certain scenarios. As indicated, all these results are devastating for the proponents in the sense that carefully formulated versions of Mechanism informed by Gödel's disjunction are logically consistent (provided Peano Arithmetic is) in all situations one can think of.

To provide those logical achievements here in more detail is both unnecessary and insufficient for the simple reason that the proponents themselves seem to ignore them by plainly dismissing the whole glorious technicality in favor of alleged informal evidence against the second option of Gödel's disjunction. The discussion could be closed here and shifted to a different kind of philosophical investigation of the mind. The trouble is that the proponents want to base their philosophical argument on, of all things, Gödel's theorem. They insist that their original insight, if properly understood, is valid despite the objections.[7]

We may try to summarize the whole controversy as follows. Proponents assume (implicitly and sometimes explicitly) a self-evident capacity of the human mind ("getting hold", "twigging", "truth-divining", see note 7), which can briefly be called understanding. They further see the ability to understand as an obviously non-mechanical attribute. This is essentially what Descartes tried to say in his oft-quoted anti-Mechanist argument,[8] or what John Searle illustrates by his Chinese room argument. Descartes apparently considered the test of the presence of understanding to be a matter of course, which is not the case anymore for us who know modern computers. In any case, we simply know (or feel)

[7] Relevant quotes are, for example:

"There is a way of arguing that commends itself to those possessed of minds, who get the hang of the Gödelian argument, and twig that they can apply it, suitably adapted, in each and every case that crops up. Mechanists may refuse to see the general case, and, acknowledging only knock-down arguments, will have to be knocked down each time they put forward a detailed case: minds can generalise, and will realise that defeat for the Mechanists is always inevitable" (Lucas, 2011).

"As to the very dogmatic Gödel-immune formalist who claims not even to recognize that there *is* such a thing as mathematical truth, I shall simply ignore him, since he apparently does not possess the truth-divining quality that the discussion is all about" (Penrose, 1999, p. 582).

[8] "For it is highly deserving to remark, that there are no men so dull and stupid, not even idiots, as to be incapable of joining together different words, and thereby constructing a declaration by which to make their thoughts understood; and that on the other hand, there is no other animal, however perfect or happily circumstanced, which can do the like" (Descartes, 1637, Part V).

we are conscious and express meanings, and there is no real argument about this. What becomes unclear is whether the existence of understanding can be conclusively proven exclusively on the syntactic level, that is, on the level of produced signs. This yields the question whether syntactic rules can simulate understanding successfully, that is, whether the same syntactic output can be obtained without the corresponding understanding. Here an apparently equally obvious proposition arises, namely that Gödel's incompleteness theorem proves conclusively the impossibility of such a simulation. This is the core anti-Mechanist thesis. The latter proposition is nevertheless wrong, since the second possibility in Gödel's disjunction remains unrefuted: it may be the case that the entire output corresponding to the (human) understanding is successfully simulated by purely syntactic rules, namely rules that (forever) transcend the (human) understanding in question. This is the state of the art from the technical point of view, which, however, makes nobody happy. Opponents cannot concede the thesis while proponents understandably feel that the objection misses the point. Lucas's objection could be formulated as follows: The syntactic rules mentioned above must make some sense, namely as rules. It is irrelevant, Lucas can insist, whether the human mind can or cannot understand them, in any case they are understandable "in principle", understandability is part of their being rules. Consequently, in order to save the point, proponents are forced to adopt some kind of metaphysical commitment concerning formal systems and the capacity of human mathematical understanding, which, however, have no clear backing in Gödel's theorem.

3. What Proponents Want to Say

The proponents were lured into an incorrect logical claim by the necessity to formulate their claim as a thesis that permits logical proof, which in turn implied dubious metaphysical assumptions. I want to suggest that the real point the proponents are after is something different, namely the inexhaustibility of understanding by expression, of meaning by syntax. Let me start by illustrating the uncertain relation between understanding and Turing Machines (or formal systems) first with an example of a finite structure like chess, and then with the question of consistency.

From the point of view of the present anti-Mechanist argument, chess is a trivial case of a finite directed graph of legal positions with edges representing moves. Every possible claim about chess is trivially decidable by an exhaustive search. On the other hand, it is safe to say that as long as human competitive chess will exist, we shall continue to speak about the understanding of a position in chess. In order to assess how such an understanding relates to computations done by a Turing Machine, let us compare a brute force algorithm with the sophisticated engines we can use today that define an evaluation function and optimize it within a large, but still limited search space. The evaluation function incorporates a formalization of the understanding of chess by top players, or, it is blindly inferred from a huge number of matches (in cases such as AlphaZero deep learning program). The evaluation function is the closest parallel to a "computer's understanding" of the game, and it is what practical artificial intelligence is all about. Nevertheless, if we ignore questions of computational complexity (which have no significant place in the anti-Mechanist controversy), the tricky nature of evaluation function becomes irrelevant. The brute force algorithm, which contains no advanced intelligence and could be written by any decent undergraduate student, becomes unbeatable. This is a rather trivial illustration of the fact, that by "understanding" we mean something else then blind syntactic ability. If we want to interpret "artificial intelligence" as an "understanding" possessed by Turing Machines, we either have to consider computational complexity, or to explain why understanding of finite (albeit very large) structures is substantially different from understanding of infinite ones.

A touchstone for what the role of understanding is within formal logic is the question of consistency. Aristotle, in his original formulation of the principle of non-contradiction, argued that contradiction must be excluded because it destroys meaning.[9] It is completely unclear what somebody says, or whether he says anything at all, if the same claim is asserted and denied in the same time and the same sense. The care with which the sameness of the two claims is stressed underlines how we usually deal with an inconsistency. We either try to repair it on the formal level of expression (as a typo), or, when the misprint is excluded, we try to search for a deeper distinction which would make the apparent sheer contradiction comprehensible. This is more than "overcoming the contradictions by pointing to the metaphorical character of expressions" as Krajewski suggests in one place (2020, p. 22); unless "metaphor" is understood not as a "mere metaphor" but as a substantial feature of any meaningful speech. Let us consider the seminal example of a set of all sets that are not elements of themselves. There may be an argument about whether the very expression "being its own element" makes sense. The answer will depend on what exactly we mean by "incidence", that is, by "being an element of". We may try to capture the exact meaning by various ways of reflection, for example by some kind of Husserlian "eidetic variation". Formal logic proposes to investigate the question on the syntactic level of propositions that include the word "incidence". We are invited to pretend that we have no idea at all what the sign $\in$ means. We just understand how it can be manipulated (note for further purposes that even this is a kind of understanding). Eventually, we discover a sentence that can be derived, according to the rules, as well as its negation. What shall we do? Formally speaking, we just discard the theory. On practical level, some kind of "correction" takes place, so often invoked in the anti-Mechanist controversy. The set in question certainly

[9] "If on the other hand it be said that 'man' has an infinite number of meanings, obviously there can be no discourse; for not to have one meaning is to have no meaning, and if words have no meaning there is an end of discourse with others, and even, strictly speaking, with oneself; because it is impossible to think of anything if we do not think of one thing…" (*Metaphysics*, IV, 1006b). See my paper (Holub, 2004) in Czech for a discussion of Aristotle's approach.

cannot in the same time and in the same sense be and not be its own element, independently of what incidence means. That makes no sense. However, it does not imply that a set cannot be its own element. Although making the formula "$x \in x$" itself contradictory is one possible (and standard) solution, it is an overcautious one. There are theories which allow sets to be elements of themselves. The collection of sets that are not elements of themselves is then just not itself a set. If we consider the last conclusion paradoxical, then we have not taken the formalization seriously. During the formalization, we were asked to forget completely that variables are supposed to refer to "collections". In fact, there is a standard technical (meta)term for this kind of collection, namely a proper class. If this is not a proof of a specific mathematical sense of humor, it is at least a proof of a pragmatic approach which cares as little as possible about formal contradictions, and instead is driven by understanding.

In contrast to the essentially infinite nature of both Turing Machine computation and the consistency requirement, the likely original motivation of the proponents is relevant already for human understanding in its finite form. We have to abandon the misleading and unclear idea of simulation and focus instead on the tension between expression and its meaning. This happens to be the starting point of an area of philosophy as alien to formal logic as *philosophical hermeneutics*. According to its main exponents, the core tenet of philosophical hermeneutics is *verbum interius*,[10] or the *surplus of meaning*,[11] the fact that the meaning intended by the speaker or writer never perfectly matches the linguistic expression. This precludes a direct approach to the intended meaning for an interlocutor or reader, making an interpretation necessary. Moreover, there is no pure "original intention", independent of the expression, for the speaker either. In order to fix any meaning, it is necessary to express it. The need for interpretation therefore applies to all thinking which becomes, in Plato's words, an inner dialogue of the mind. The dialectics of understanding and expression is grounded in the unique perspective of the author and the unique context of the locution as opposed to the stability of the expression, which allows others to share the intended meaning, as well as the authors to return, possibly with a surprise, to their own previous thoughts. The gap between expression and meaning is revealed by a reflection on the expression, and the comparison with meaning it allows. The expression is a transparent medium leading directly to the meaning in the case of a successful understanding. We become aware of the expression when the understanding is disrupted. The expression then loses its transparency, becomes visible as an independent reality, and an interpretation is needed in order to reestablish understanding. Such an interpretation adds new expressions that may help to elucidate the original meaning but, at the same time, they themselves may be-

[10] See the foreword to Grondin's (2011), where the author quotes his discussion with H. G. Gadamer.

[11] See, for example, Ricoeur's (1976).

come unclear. The process then continues, until an understanding needed for practical needs of the particular situation is achieved.

Hermeneutics shares with logicism the suspicion concerning a direct approach of consciousness to itself, in some kind of transcendental reflection not mediated through any expression. Paul Ricoeur replaces the self-transparent Cartesian *cogito* with a *cogito brisé*, a broken consciousness. Formal logic is a deliberate strategy to make the expression fully "opaque", fully devoid of meaning. Its full focus is on the syntactic rules. We remarked above that even in this case an understanding of the formal system *qua* formal system is required (for example, understanding of how well-formed formulas can be obtained). The formal system thus becomes a mathematical object in an ordinary sense, which substitutes for the original one (like natural numbers) and which can be informally, or again formally, investigated. However, the investigation should receive no guidance from the original, motivating understanding, lest be misled by it. It was Frege's and Hilbert's hope that restricting understanding in this radical way will eventually yield a better grasp of the original meaning. The hope is that syntactic or logical rules, while being simpler to control, will nevertheless fully substitute for the suspended meaning. This hope was frustrated by Gödel. Even in mathematics, the understanding always means more than what its formulation says explicitly.

4. Informal Mathematics in Gödel's Proof

The proof of Gödel's theorem reveals very clearly the above described hermeneutic principles through the relation between formal expressions of Peano Arithmetic on one side, and the informal mathematical understanding on the other side. I will show this by an analysis of the technical content of the proof. The main goal of this analysis is to trace informal mathematical aspects of the proof. "Informal mathematics" should be understood as ordinary mathematics, which in fact uses formalism quite heavily, but which is nevertheless not formal in the logical sense. In other words, "informal" means mathematical but at the same time meta-logical.[12]

Gödel's incompleteness theorem claims that any recursive formal theory that is sufficiently strong contains a sentence such that neither the sentence nor its negation is provable in the system. The theory in question can be some theory designed to capture our understanding of natural numbers, for example Peano Arithmetic.

The proof of the theorem is based on three technical ingredients.[13] The first one is the famous Gödel numbering which establishes a correspondence (possi-

[12] I think this is what Gödel (1931, p. 176) has in mind when he speaks about "metamathematische Überlegungen".

[13] An excellent self-contained exposition of the main structure of the argument, spanning only five pages, can be found in the first chapter of Smullyan (1992, pp. 5–9).

bly one-to one) between formal expressions in the language of the theory, and natural numbers. This is a crucial step since in this way formulas are transposed from the level of language to the level of objects the language is supposed to describe. This allows us to eventually interpret the theory in a way that yields some information about the theory itself. It should be stressed, however, that the correspondence between numbers and formulas is realized on the meta-level, by the mathematician writing the proof. Moreover, the existence or meaningfulness of the very structure of natural numbers that we use to encode formulas is of course in no way guaranteed *a priori* by the theory that is designed to describe them. We rely on our pre-formal (in the logical sense) meta-understanding.

The second main ingredient of Gödel's theorem is the diagonalization, which is also the core of other related cornerstones of modern mathematics, such as the existence of algorithmically undecidable problems, and the concept of higher infinities beyond countable infinity. The simplicity of the idea deserves to be stressed and kept in mind. In fact, it is recommended to contemplate the basic nature of the diagonal argument as an antidote whenever one is tempted by the "mystical charm" (Krajewski, 2020, p. 15) of Gödel's theorem or related results. The finite version of the diagonal argument provides an elegant constructive form of the fact that the number of sequences of length n is larger than n (provided there are at least two distinct symbols). It is always straightforward to exhibit a particular missing sequence, namely the "negated diagonal", that is, the sequence whose i-th element is a symbol distinct from the i-th element of the i-th sequence of the list. This idea can be extended to any countably infinite list of infinite sequences, yielding Cantor's proof for the uncountability of the continuum.[14]

The third main ingredient of Gödel's theorem is expressibility, the ability to describe certain important features of the language in terms of the language itself. Here we essentially use the above introduced encoding. More careful formulation should therefore be that the encoding is extended to more complex linguistic expressions (ultimately to formal proofs), and the numbers that represent expressions with desired properties are captured by suitable formal expressions. Specifically, the theory must be able to formulate "n-th expression to which number n is substituted" (realizing the diagonal idea), and, most prominently, "m is a proof of n-th expression". Once more, m actually is a number, which encodes a formal proof of the n-the expression. Construction of formulas representing the above properties is the technically difficult part of the proof,

[14] Let me remark that Cantor's theorem can serve either as a starting point for doubts about actual (as opposed to potential) infinity, or as the entrance gate to "Cantor's paradise" of Set Theory. Set Theory then postpones its moment of reflection to the problem of the "universal diagonal" of the collection of sets that are not elements of themselves, which is famously contradictory if accepted naively. Set Theory could thus be characterized as the grey area created by the diagonal argument and extending between full acceptance of actual infinity and the rejection of contradiction.

requiring a logician of Gödel's greatness.[15] Again, it must be stressed that by proof we mean a formal proof here, that is, a sequence of formulas obtained by successive elementary derivation steps. This observation can hardly be overestimated. In fact, the difference between formal and informal proof is probably the most contentious aspect of the debate. Just as the Turing Machine is no machine although it can be instantiated by one, formal proof does not prove anything although it is designed to reflect a full-blooded proof and can be interpreted as such.

What exactly is required from the formal theory to be able to express notions needed for the incompleteness theorem can be investigated in several ways. The standard mathematical way a theory is shown to be too weak, and therefore complete and decidable, is quantifier elimination. This is related to the fact, we pointed out above, that finite structures are decidable trivially. Undecidability always arises due to the presence of quantified formulas, formulas which dare to claim something about all objects. Quantifier elimination reveals the weakness of a theory by showing that each such formula is in fact equivalent to one without quantifiers. The theory is shown to be too weak to be able to say something universal.

Seen from this perspective, the gap between the decidable Presburger Arithmetic and the undecidable Peano Arithmetic becomes curious. The difference between them is the presence of multiplication. It turns out that speaking about natural numbers in terms of addition only does not allow anything to be said universally. Krajewski (2012) considers the appearance of undecidability when multiplication is added to be one example of "emergence" in mathematics, as a fact that remains irreducibly surprising even for an expert. Let me attempt a speculative explanation of this phenomenon.[16] It may be argued that multiplication is the place where natural numbers start to apply to themselves. While the basic role of natural numbers is to count objects (be they apples or abstract units), in multiplication the counted objects become numbers themselves. No more five times an apple, instead five times four.[17] The four is suddenly not only a quantitative property of a particular assemblage, it becomes a proper object which itself deserves to be counted. This can be therefore seen as the required threshold of "self-reflection".[18]

Using the above three ingredients, Gödel is able to find an independent sentence, a sentence which can neither be proved nor disproved in the formal system. Formally speaking, we have a pair of formulas (the sentence and its negation),

[15] There is, of course, the difference between the difficulty of a proof, and the difficulty of discovering it. In today's form the full proof can be included in an undergraduate course. Gödel himself (1931, p. 173) calls the independent sentences he derives in his famous paper: "Relativ einfache Probleme aus der Theorie der gewöhnlichen ganzen Zahlen".

[16] I am indebted to a remark by Kateřina Trlifajová for this idea.

[17] Of course, both the multiplicand and multiplier must be allowed to be universally quantified. Multiplication by a given individual number can be expressed as a sum.

[18] I wonder whether this speculation can be somehow supported by the analysis of quantifier elimination.

such that both of them are unprovable, that is, they cannot be obtained from other specific formulas (called axioms) in a prescribed way. This purely formal fact does not sound very interesting. Its real importance depends on the interpretation we give to the formulas. More specifically, we interpret sentences as claims about natural numbers that can be true (or false). Also, we interpret formal proof as an object that faithfully captures ordinary mathematical reasoning, which, in turn, depends on the truth preserving quality of certain reductions.

Finally, we are convinced that the theory in question (Peano Arithmetic to start with) is consistent.[19] This is a particular case of the way mathematicians tend to deal with possible inconsistency, which we have discussed above. The consistency of Peano Arithmetic is a particularly bold assumption, and some serious mathematicians have even sincerely doubted that it is the case.[20] The point is that Peano Arithmetic contains infinitely many axioms within the scheme of induction. In particular, it contains the induction claim for arbitrarily large and complex formulas, even for those we shall never be even able to read, let alone to understand what they say.[21] However, even if it turned out that Peano Arithmetic is contradictory, it would not, as Krajewski correctly observes (2020, p. 22), necessarily disturb ordinary mathematics in any significant way. We can imagine that an extremely huge proof of contradiction would somehow miraculously appear somewhere on the internet. It would be fairly easy to verify, using computers, that the contradiction is genuine. Nevertheless, it would involve a lot of extremely complicated instances of the scheme of induction. Lucas believes that we would eventually be able to sort things out, an example of the depth of his optimism. From the practical point of view, however, it would just mean that some of the involved axioms should be prohibited. Undoubtedly, a new field of research would be created to investigate which one, but ordinary mathematicians would be, at best, just more conscious of what kind of induction they use.

Nevertheless, let us believe with Lucas that “we” (whoever that is) are “in principle” (whatever that means) able to understand and even to verify the validity of all axioms. Combining this with the truth preserving quality of formally logical inferences, we are ready to claim that no arbitrarily large derivation within the whole monstrous theory can lead to a contradiction simply because such a contradiction could, “in principle”, be translated into firm evidence of the fact that zero equals one. Finally, granted all this, we are able to contemplate the insufficiency of our theory (Peano Arithmetic) to exhaust our intuition.

[19] More precisely, we believe that it is ω-consistent, which excludes the possibility that a provable universally quantified formula is at the same time disprovable for all numerals.

[20] See, for example, (Nelson, 2006). The intransigence of Nelson’s views has certainly been compromised by his mistaken announcement that he actually proved the inconsistency of Peano Arithmetic.

[21] Nelson (2006) points out further difficulties related already to induction on relatively simple formulas, which illustrate that our belief in natural numbers is far from self-evident.

It is not that difficult to understand why this magnificent proof leads Lucas to celebrate "getting the hang of the argument" and "twigging that we can apply it" or Penrose to brag about a "truth-divining quality" (see remark 7) that are supposed to establish the superiority of creativity over rule-following.[22] But even if we enjoy joining Lucas in his exaltation of the scintillating human mind, we have to return to our question about the exact contribution of Gödel's theorem here. We have to do so because we have seen how rather than support optimism, the validity of the theorem turns out to depend on it. At best, its proof is one among many occasions to experience mathematical understanding at work. It also makes clear that our understanding is not exhausted by theorems provable in Peano Arithmetic, or in any theory for which we are able to perform a similar construction, provided that the theory is consistent. But it does not exclude the possibility that our mathematical understanding is governed by a formal system for which we are unable to carry out the proof, since we are unable to understand it. We have seen that the main difference between proponents and critics is whether they care about the latter qualifications. While critics expected that the argument will deal with them, proponents instead took them for granted. But then Gödel's theorem is just an instance of mathematical understanding, which can be philosophically investigated but whose specific content provides no particular philosophical contribution. The argument is flat, and it is understandable that Lucas wants to turn the page.[23] Frege's original objective was to explain the conceptualization involved in mathematics by means of logic. The fact that in Gödel's theorem logicism defeats itself, as it were, by its own means is undoubtedly an epochal result. On the other hand, the discussion about the central objective of logicism has gone on since the sixties within neo-Fregeanism and neologicism.[24] Lucas's paper could have been part of that discussion, less famous but philosophically more substantial, had it started with "Gödel's Theorem seems to me to prove that natural numbers cannot be described purely logically…"

The failure of the project induced by Gödel's theorem represents a vindication of (Kantian) intuition. However, as soon as formal logic becomes an established mathematical discipline, it can be cultivated independently of its philosophical foundations, as can any other mathematical discipline, like arithmetic or geometry. Gödel's theorem may then lead unprepared students astray to nonsensical conclusions of the following kind. The failure of the attempt to capture fully the whole formal truth about natural numbers, seen from the point of view of formal logic, casts doubts not primarily on the logic itself but rather on our original intuition about natural numbers. Is there something like the standard model at

[22] See: "[A]lthough Gödel cannot make us scintillate, he does show that scintillation is conceptually possible. He shows us that to be reasonable is not necessarily to be rule-governed, and that actions not governed by rules are not necessarily random" (Lucas, 2011).

[23] For the same reason, it is unclear why Krajewski's "Theorem of Inconsistency of Lucas" should be described as "unexpected" (2020, p. 32).

[24] See for example (Kolman, 2005) for a polemic against the neologicism.

all? Is the standard model somehow distinguished among other models? Is it possible to explain the standard model in set theoretical terms? Which set theoretic model is appropriate and why? A working logician sometimes seems to feel much more comfortable when speaking about nonstandard models, and some even believe that Gödel's incompleteness theorem actually shows that there is nothing like the standard natural numbers.[25] However, the proof of the theorem, as we stressed, is based on the interpretation of formal sentences as claims about natural numbers. What natural numbers? If the logical analysis deconstructs our concept of natural numbers, then the deconstruction itself is undermined as far as it depends on the interpretation of natural numbers. Where do we stand then? The space for sophistry and wild mystical comments is open in this aporia (and the opportunity is amply seized). For example, it is a basic "ordinary mathematical" conclusion that the Gödelian formula, which declares itself to be unprovable, indeed is unprovable, and therefore true. If it were provable, then it would be true, and therefore unprovable (since that is what it says under the interpretation), which is an (informal, ordinary mathematical) contradiction.[26] Does this argument work anymore when we are not sure about the standard interpretation? Since the formula is independent of the axioms of the theory under investigation, there is a (nonstandard) model in which it is provable. Does Model Theory magically allow (formal or even informal) contradiction?

These are questions that require a calm reflection on the technical ingredients of Gödel's theorem listed above, combined with keeping in mind that, when proving Gödel's theorem, we are doing "ordinary mathematics" with ordinary natural numbers. Formulas are just sequences of symbols with no magic power to destroy our mathematical understanding. These formulas are actually objects of our mathematical understanding as much as natural numbers, which is particularly apparent in the encoding of formulas by numbers. Our knowledge, then, is fully dependent on the interpretation we give to formulas as claims about natural numbers, and on the truth preserving quality of derivation. The sentence is true, in natural numbers, because otherwise we would obtain an inconsistency in our understanding of natural numbers and of truth derivation. The alleged proof of the sentence would be inconsistent with the claim the sentence makes about its own unprovability. The latter, recall, depends on the encoding. The sentence claims that there is no number with certain subtle properties expressed by a complicated formula. However, the alleged proof of the sentence, if encoded, will yield a number with exactly those properties. In order to see this, we have to "get

[25] Concerning set theoretic models let me mention a paper by Paul Benacerraf (1965, p. 73) which happens to conclude: "They think that numbers are really sets of sets while, if the truth be known, there are no such things as numbers; which is not to say that there are not at least two prime numbers between 15 and 20".

[26] See already Gödel's original paper: "Aus der Bemerkung, dass [R(q); q] seine eigene Unbeweisbarkeit behauptet, folgt sofort, dass [R(q); q] richtig ist, denn [R(q); q] ist ja unbeweisbar (weil unentscheidbar). Der im System PM unentscheidbare Satz wurde also durch metamathematische Überlegungen doch entschieden" (1931, p. 176).

the hang" of the proof, in Lucas' words. Finally, the Gödelian sentence together with all provable sentences of the original theory form a consistent system, which therefore has a model. This model is a mathematical structure, which looks much like natural numbers, but it contains an element which represents the proof of the Gödelian formula in the original theory. Nevertheless, this brings about no inconsistency, since the element representing the (non-existing) proof of the Gödelian formula is not a standard number, therefore it yields no sequence of formulas we would accept as a proof in "ordinary mathematics". The firm basis of all this in informal mathematics is obvious.

5. Conclusion

The troubled dispute about Mechanism inspired by Gödel's theorem is an instructive example of difficulties resulting from the lack of respect for methodical limitations of different scientific areas. Three levels are at play: mathematical logic (formal arithmetic), (informal) mathematics of natural numbers, and the philosophical reflection on both these disciplines. Since the anti-Mechanist claim has a philosophical nature, after its forced transition to the field of formal logic it suffers from not sufficiently distinguishing between technical results on the one hand, and the philosophical reflection on their significance on the other. While moving between "levels" and "meta-levels" is well established within formal logic, it typically happens within the discipline itself. Model Theory builds models of one theory using another one, obtaining thereby essentially relative results. Asking philosophical questions transcending the discipline as a whole is often seen with suspicion by logicians, if not directly dismissed as a delusion.[27] Frege (1998, p. XII) was well-aware of this predicament when he was skeptical about prospects of his own work.[28] However, the anti-Mechanist thesis is precisely a reflection on philosophical consequences of purely technical results. A seemingly paradoxical principle applies to such attempts. In order to estimate the relevance of philosophical, informal conclusions drawn from a technical theorem,

[27] Peter Koellner (2018b, p. 476), for example, dismisses concepts of "absolute provability" and "knowability by the idealized human mind" as "not sharp enough for our questions […] to have definite sense and determinate truth-values", but nevertheless continues: "With the above discussion in place, I would like to once again suspend the above skeptical considerations and assume, for the sake of argument, that the concepts of 'absolute provability' and 'knowability by the idealized human mind' are definite".

[28] "Sonst sind die Aussichten meines Buches freilich gering. Jedenfalls müssen alle Mathematiker aufgegeben werden, die beim Aufstossen von logischen Ausdrücken, wie "Begriff", "Beziehung", "Urtheil" denken: *metaphysica sunt, non leguntur!* und ebenso die Philosophen, die beim Anblicke einer Formel ausrufen: *mathematica sunt, non leguntur!* und sehr wenige mögen das nicht sein. Vielleicht ist die Zahl der Mathematiker überhaupt nicht gross, die sich um die Grundlegung ihrer Wissenschaft bemühen, und auch diese scheinen oft grosse Eile zu haben, bis sie die Anfangsgründe hinter sich haben. Und ich wage kaum zu hoffen, dass meine Gründe für die peinliche Strenge und damit verbundene Breite viele von ihnen überzeugen werden".

it is absolutely necessary to investigate the nature of the technicality involved. In other words, the conceptualization involved in the formal logic has to be prominently taken into account. Attempts to draw informal conclusions from a formal argument understood informally, are, in my opinion, the essence of the imprecision in thinking that causes a big part of the sadness of the corresponding episodes of our intellectual life.

The tension between formality and informality creates a difficulty which has its impact already on the superficial level of the external organization of relevant papers. They typically contain long sections of technical explanations using formal language.[29] The technical achievements then have to be interpreted, translated into ordinary language and their relevance has to be established. The technical language also typically contains lots of terms with suggestive non-technical meanings, terms that tend to permeate the informal discussion although the non-technical meaning may very poorly reflect the term's technical function. Real numbers, to take a trivial example, are no more real than natural numbers. While nobody is likely to draw philosophical consequences from the term "real number", the term "provable" in our context turned out to be significantly less safe. We have discussed the example of the word "machine".

I have attempted to show that in the Mechanist controversy the ordinary human mind, and its capacity to understand, is the first casualty of the battle which is officially waged for its sake. It simply turns out that mathematics, or at least formal logic, has no good tools to capture the cherished superiority of understanding.[30] The anti-Mechanist argument is a misguided effort to vindicate the capability of a particular human mind by means of the idealized one. We have seen how heavily the proof of Gödel's theorem depends on the informal understanding of arithmetic. The proponents make a quixotic attempt to translate the evidence for the superiority of the informal understanding to the formal level. They want to use the failure of logicism to prove at least this failure in the logically water-proof way. Gödel's incompleteness theorems are a kind of touchstone for the ambition of formal logic to substitute syntax for semantics for good. Lucas and Penrose are tragic heroes of this fight. They are driven by the obvious superiority of meaning over the syntax. However, they make a foolish choice of attempting to prove this superiority by the very means of syntax. While the Mechanist wants to reduce meaning to the pure manipulation of symbols, Lucas and Penrose want to vindicate the superiority of meaning—by pure manipulation of symbols. We are however well advised to spend the finite resources of our creative mind on something more reasonable than syntactic competitions with Turing Machines.

[29] The remark by Paul Benacerraf (1967, p. 13) often applies: "I trust that the following exposition will prove too elementary to be of any interest to those who are familiar with the logical facts, and too compressed for those who are not. For the sake of future reference, however, it must be done".

[30] Cf. (Feferman, 2009, p. 213): "[I]t is hubris to think that by mathematics alone we can determine what the human mind can or cannot do in general".

REFERENCES

Benacerraf, P. (1967). God, the Devil, and Gödel. *Monist*, *51*(1), 9–32.

Benaceraf, P. (1965). What Numbers Could Not Be. *Philosophical Review*, *74*(1), 47–73.

Descartes, R. (1637). *Discourse on the Method*. Leiden. Retrieved from: http://www.gutenberg.org/files/59/59-h/59-h.htm

Feferman, S. (2009). Gödel, Nagel, Minds, and Machines. *Journal of Philosophy, 106*(4), 201–219.

Frege, G. (1998). *Grundgesetze der Arithmetik*. Olms Verlag.

Gödel, K. (1931). Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I. *Monatshefte für Mathematik und Physik*, *38*, 173–198.

Grondin, J. (2011). *Einführung in die philosophische Hermeneutik*. Darmstadt: Wissenschaftliche Buchgesellschaft.

Holub, Š. (2004). Aristotelův princip sporu ve čtvrté knize Metafyziky. *Reflexe*, *25*, pp. 71–80.

Klein, J. (1969). *Greek Mathematical Thought and the Origin of Algebra.* Cambridge, Mass.: M.I.T. Press.

Koellner, P. (2018a). On the Question of Whether the Mind Can Be Mechanized, I: From Gödel to Penrose. *Journal of Philosophy*, *115*(7), 337–360.

Koellner, P. (2018b). On the Question of Whether the Mind Can Be Mechanized, II: Penrose's New Argument. *Journal of Philosophy*, *115*(9), 453–484.

Kolman V. (2005). Lässt sich der Logicismus retten. *Allgemeine Zeitschrift fur Philosophie*, *30*(2), 159–174

Krajewski, S. (2020). On the Anti-Mechanist Arguments Based on Gödel's Theorem. *Studia Semiotyczne*, *34*(1), 9–56.

Krajewski, S. (2012). Emergence in Mathematics? *Studies in Logic, Grammar and Rhetoric*, *27*, 95–105.

Lucas, J. (1961). Minds, Machines and Gödel. *Philosophy*, *36*(137), 112–127.

Lucas, J. (2011). *The Gödelian Argument: Turn Over the Page*. Retrieved from: http://users.ox.ac.uk/~jrlucas/Godel/turn.html

Nelson E. (2006) *Warning Signs of a Possible Collapse of Contemporary Mathematics*. Retrieved from: https://web.math.princeton.edu/~nelson/papers/warn.pdf

Penrose, R. (1999). *The Emperor's New Mind*. Oxford: Oxford University Press.

Putnam, H. (1975a). Minds and Machines. In: H. Putnam (Ed.), *Mind, Language and Reality: Philosophical Papers* (vol. 2, pp. 362–385). Cambridge: Cambridge University Press.

Putnam, H. (1975b). Philosophy and Our Mental Life. In: H. Putnam (Ed.), *Mind, Language and Reality: Philosophical Papers* (vol. 2, pp. 291–303). Cambridge: Cambridge University Press.

Putnam, H. (1994). The Best of All Possible Brains? Review of Roger Penrose, *Shadows of the Mind*. *New York Times Book Review*, *144*, 7.

Ricoeur, P. (1976). *Interpretation Theory: Discourse and the Surplus of Meaning*. Fort Worth, Texas: Texas Christian University Press.

Shapiro, S. (1998). Incompleteness, Mechanism, and Optimism. *The Bulletin of Symbolic Logic*, *4*(3), 273–302.

Smullyan, R. M. (1992). *Gödel's Incompleteness Theorems*. Oxford: Oxford University Press.

Turing, A. M. (1937). On Computable Numbers, with an Application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society*, *s2-42*(1), 230–265.

Vopěnka, P. (2010), *Calculus infinitesimalis, pars prima*. Praha: OPS.

STUDIA SEMIOTYCZNE (SEMIOTIC STUDIES), t. XXXIV, nr 1 (2020), s. 203–232
ISSN 0137-6608, e-ISSN 2544-073X
DOI: 10.26333/sts.xxxiv1.09
Article

DAVID KASHTAN [*]

DIAGONAL ANTI-MECHANIST ARGUMENTS

SUMMARY: Gödel's first incompleteness theorem is sometimes said to refute mechanism about the mind. §1 contains a discussion of mechanism. We look into its origins, motivations and commitments, both in general and with regard to the human mind, and ask about the place of modern computers and modern cognitive science within the general mechanistic paradigm. In §2 we give a sharp formulation of a mechanistic thesis about the mind in terms of the mathematical notion of computability. We present the argument from Gödel's theorem against mechanism in terms of this formulation and raise two objections, one of which is known but is here given a more precise formulation, and the other is new and based on the discussion in §1.

KEYWORDS: mechanism, mind, computability, incompleteness theorems, computational theory of mind, the cogito, diagonal arguments, Gödel, Descartes, Tarski, Turing, Chomsky.

Mechanism about *X*, roughly, is the view according to which *X* can be understood in terms of a machine. Descartes famously held the doctrine of mechanism with regard to everything but the human mind. More recently, some writers have argued that Gödel's famous theorem about the impossibility of a complete computable axiomatization of arithmetic shows that the human mind is not amenable to a mechanistic explanation.[1] Gödel's theorem is a likely candidate for the job of combatting mechanism, first, because it is very famous; second, because the notion of computability is the modern approach to mechanism about the mind,

[*] Hebrew University of Jerusalem, Edelstein Center for the Philosophy of Science. E-mail: david.kashtan@mail.huji.ac.il. ORCID: 0000-0002-4237-1809.

[1] Most notably (Lucas, 1961; Penrose, 1989; 1994). I refer the reader to (Krajewski, 2020) for a brief recap of the dialectic.

and Gödel's theorem is, or at least entails, a limitative result on it; and third, because it is a diagonal argument, and diagonal arguments seem to be exactly the right tool to wield against a thesis such a mechanism.

However, Gödel's theorem is a precisely formulated and decisively proven mathematical theorem, and mechanism about the mind is a vague and messy philosophical question. Any attempt to bring a mathematical theorem to bear on a philosophical question should be viewed with suspicion. Shapiro, for example, asserts that "there is no plausible mechanistic thesis on offer that is sufficiently precise to be undermined by the incompleteness theorems" (1998, p. 275). The problem, according to Shapiro, lies in the many idealizations that are involved in applying the theorems to humans and to machines. But if Shapiro's condemnation is correct, this is hardly comforting to the mechanist, who should want to resist the Gödelian argument by virtue of being right, not by the vice of being vague. The goal of this paper is to formulate a mechanistic thesis sharply enough that it stands a chance both of being refuted by and of resisting the Gödelian argument.

The purpose of §1 is to get a handle on mechanism about the mind. Starting with Descartes, we review the central epistemological motivations for mechanism, distinguish between programmatic and metaphysical mechanism, and inspect Descartes' reason for denying mechanism about the human mind. Then, we ask whether the advent of computability theory and modern computing machines would have made Descartes change his mind. §2 is about the Gödelian anti-mechanistic argument. Based on the discussion of §1, a sharp criterion is offered for deciding the metaphysical mechanistic thesis, in terms of the computability of sets of mental representations. The classic Gödelian anti-mechanist argument is formulated with reference to this criterion, and two objections to it are raised, one of which is new. In the final subsection a sketch of an alternative diagonal anti-mechanist argument, based on Tarski's indefinability theorem, is given.[2]

1. Mechanism and Anti-Mechanism

The bare term "mechanism", or mechanism about some thing *X*, can be glossed as the claim that *X* is, often despite appearances, essentially a machine. We are interested specifically in mechanism about humans—the question whether humans are essentially machines. Gödel's theorem is thought to have bearing on this question because of its relation to computability theory. The machines in question are, therefore, the special class of computing machines. However, the origins of mechanism, and of mechanism about the mind, lie long before the modern theory of computing machines, in the philosophy of science of several

[2] Michael Goldboim and Balthasar Grabmayr read an earlier draft and helped reduce the number of errors. Those that remain are mine alone. I have also benefited greatly from conversations on the themes of the paper with the aforementioned, as well as with Eli Dresner, Philip Papagiannopoulos and Carl Posy.

important early modern thinkers, most notably Descartes. Descartes' is an interesting case because he both endorsed mechanism about animals and rejected mechanism about humans. By examining his reasons, we can form an idea of why mechanism is attractive, as well as of why it is not compelling. In addition, we may ask whether Descartes' anti-mechanism was not tied to the particular kinds of machines that he knew, and whether the advent of computing machines would have caused him to change his mind.

1.1. Cartesian Mechanism

A classical machine, or mechanical system, roughly, is a finite collection of basic corporeal objects that can move in space and interact through contact, i.e. collision or pressure, and together achieve some desired effect. The properties of the set of basic parts, the sizes and shapes of the objects, and their spatial configuration, we call the basis of the system. From the basis, using the mechanical laws of motion and force, one can calculate the effect; and conversely, if one is interested in a certain effect, one can set up a basis that will achieve it, in other words one can engineer an effect. What allows engineering is the fact that mechanical systems are "bottom-up": that the basis is describable, perceivable and manipulable independently of the effect, and that the effect is "generated" from the basis according to determinate laws.

Mechanism in science or natural philosophy is, first of all, an empirical research program according to which natural phenomena should be studied as though they are effects of mechanical systems. A mechanistic explanation of a phenomenon consists in hypothesizing a basis, and showing that the phenomenon is indeed generated from it by the laws of mechanics. The epistemological virtue of mechanistic explanations is that, since the basis of a mechanical system is describable independently of the effect, they make a positive, self-standing assertion about reality. Such an assertion is straightforward (though not always technically possible) to test, and, in principle, allows nature to be manipulated with design. They, therefore, yield a kind of engineer's, or maker's, knowledge. The contrast is with explanations in terms that are abstract, or that are describable only top-down, in terms of their explanandum, like Molière's *virtus dormitiva*. The basic claim of mechanism is that only positive theories can count as explanations, whereas abstract or top-down theories are explanatorily vacuous.[3]

An explanation is positive rather than vacuous, I propose, when it postulates a basis that has an independent criterion of existence and identity. It should be possible to determine, at least in principle, whether the basis of a hypothesized mechanical system exists in reality without appealing to the properties of the

[3] For comprehensive and detailed accounts of mechanism, especially in its Cartesian brand, see (Gaukroger, 2002; 2007; 2010), also the papers in (Gaukroger, Schuster, & Sutton, 2000). For mechanistic explanation as maker's knowledge, see (Funkenstein, 1986, p. 290).

effect; otherwise the explanation is circular. For Descartes, the basic existent is inert matter, where "inert" means that the spatial extension of material objects, including their motion in space and their collisions with one another, is all there is. Inert matter is a basic existent because spatial shapes, in principle at least, are vividly, distinctly and publicly perceived; and they are bottom-up in the sense that a spatial extension is the "sum" of its parts; the parts are independent of the whole, but not the other way around. Consequently, a mechanical system can be exhaustively described in purely geometrical terms. Mechanistic explanation becomes a kind of geometrical construction.[4]

Mechanism as a research program is thus the call to explain natural phenomena in terms of mechanical systems, and ultimately in terms of geometrical constructions. But aside from being a research program, mechanism is sometimes asserted or denied as a metaphysical thesis. Roughly, metaphysical mechanism about a natural phenomenon *X* is the thesis that *X* is a classical machine, or that the true theory of *X* is a mechanistic theory. Spatially extended matter, on this view, is not only epistemologically virtuous in being vividly perceived or imagined, it is also metaphysically substantial. Metaphysical mechanism about a phenomenon *X* is the claim that *X*, metaphysically, is extended substance, or *res extensa*.[5]

Descartes famously held that animals were, metaphysically, mere machines.[6] This thesis may sound banal to us, but in Descartes' time it was paradoxical and even revolutionary. Supposedly, what made it so unlikely in the eyes of Descartes' predecessors was the seemingly unbridgeable disparity between the behaviors of machines and of animals. In particular, there was the fact that animals and their physiology exhibit spontaneous and organized movement, whereas machines typically do no more than transform external force that is applied to them, and, therefore, cannot move "on their own". This led pre-Cartesian natural philosophy to postulate an intangible life force animating the bodies of animals. The problem with this kind of theory, in the eyes of a mechanist, is that it gives no independent information about the nature of this life force, no way to con-

[4] See (Sepper, 2000; McLaughlin, 2000) for some elaboration. Descartes' notion of the geometrically (as opposed to mechanically) constructible is wider than that of the ancients, but does not cover arbitrary curves. I am unsure whether Descartes' "geometrical" conception of matter is restricted to his geometrically constructible curves or whether it extends to arbitrary curves.

[5] Descartes doesn't, as far as I know, distinguish explicitly between mechanism as a research program and mechanism as a metaphysical thesis.

[6] This formulation is a little misleading. Descartes thought that animal bodies, including human bodies, are machines. As we will see, he did not think the same about minds. Animals did not, and humans did, have minds, so it is in this sense that *non*-human animals are *mere* machines. See (Cottingham, 1978) for more about this.

struct it in the imagination or calculate its properties. In other words, it is a vacuous theory.[7]

Descartes' endorsement of mechanism about animals was motivated by two circumstances. First, around Descartes' time, mechanistic theories of animal physiology were being developed and were achieving remarkable empirical success. Descartes himself proposed extensive theories of this kind, ranging from an account of the heart and blood circulation system, to theories of feelings and the imagination, which Descartes considered part of physiology. The second circumstance was some recent advances in technology, which allowed the construction of self-moving machines, or clockwork automata, operated by a spring or a hydraulic mechanism. Such machines were often used for recreational purposes, and given the shape of a human or an animal. Their "capacity" for self-movement would make them startlingly life-like in the eyes of Descartes' contemporaries, a fact which served to dull the edge of the perceived disparity between animals and machines.[8]

Descartes' metaphysical mechanism about animals was motivated by the science and technology of his time, but neither the empirical success of mechanistic theories nor the advent of new machines can establish a metaphysical thesis. Descartes' own physiological theories turned out to be largely incorrect. For example, though he enthusiastically accepted Harvey's momentous discovery of the circulation of the blood, Descartes rejected the attendant theory of the movement of the heart in terms of muscular expansion and contraction, and favored an account, incorrect as we now know, in terms of "ebullition" (Anstey, 2000, p. 421f). Surely, we don't want to say that an incorrect theory can establish a metaphysical thesis. Likewise, as lifelike as moving statues can get, we know perfectly well that the mechanism behind their movement has nothing in common with the mechanisms behind animal movement. It will be false, then, to say that Descartes' metaphysical mechanism about animals is in any way proven, or even strictly speaking confirmed, by the science and engineering of his time, though it was certainly suggested or motivated by them.

Still, it is not unreasonable to say that the relevance of the two motivating circumstances does spill over a little from the context of discovery to the context of justification. The fact that Descartes' theories were incorrect is less important than the fact that they were mechanistic, which is to say positive, and that they were plausible. It showed that, in principle, mechanistic physiology had a chance to succeed, even if Descartes' own theory happened to be incorrect. Likewise, the existence of moving machines, though they do not simulate the true mechanism of animal movement, shows that self-movement is mechanically possible,

[7] See (Ben-Yami, 2015, Chapter 4) for a less ahistorical discussion of the claim that pre-Cartesian science denied mechanism because the machines it knew did not move on their own.

[8] See Part Five of the *Discourse* (Descartes, 2006) for a summary of Descartes' mechanistic theory of the blood circulation system and his statement of mechanism about animals. Ben-Yami (2015) gives an extended discussion of Descartes' physiology.

and this opens the way to positive speculations about the actual mechanism. We can say, then, that although the science and engineering available to Descartes were a long way off from *proving* metaphysical mechanism about animals, they did provide *positive grounds* for it. Arguably, that's the best metaphysics can hope for anyway.

1.2. Universal Mechanism?

At least as famously, or infamously, as he endorsed mechanism about animals, Descartes rejected mechanism about humans, specifically about the human mind. In the next subsection, we will review his reasons. First, let's see what goes wrong with a seemingly quick and easy argument for mechanism about humans: the argument from universal mechanism.

In the previous subsection, we distinguished between two ways in which mechanism about a phenomenon *X* can be maintained: (a) Metaphysical mechanism is the theoretical claim that the *true* explanation of *X* is mechanistic; (b) Scientific mechanism is the programmatic call to *seek* mechanistic explanations for *X*. Now given the characterization of mechanism sketched above, one may argue that the phrase "mechanistic explanation" is redundant, since an explanation that is not positive, in the required sense, and therefore mechanistic, is no explanation at all. Let's agree to assume this, that is, that all adequate explanations are mechanistic. In addition, one may wish to deny the possibility that some natural phenomena are not amenable to explanation at all. Let's assume this as well, without discussion. From these two assumptions it is tempting to conclude a kind of *universal mechanism*—the claim that every phenomenon has a mechanistic explanation. From this, metaphysical mechanism about the human mind seems to follow immediately.

There are two main problems with this line of reasoning which it will be instructive to uncover. The first concerns the logical form of the inference. On the face of it, we have here a run-of-the-mill universal instantiation: From mechanism about all phenomena we infer mechanism about the particular phenomenon of the human mind. However, such an inference holds only if the instance is in the range of the quantifier, in this case if the human mind is a natural phenomenon that stands to be explained. The problem is that what counts as a natural phenomenon is not a simple and non-negotiable matter. To state the issue clearly, let's distinguish between the *data*, which is immediately given (by the senses, say), and the *phenomenon*, which is that which we need to explain. The phenomenon is extracted, or constructed, from the data, by various conceptual operations we can call *idealizations*, which consist primarily of extending the scope of the phenomenon beyond what has actually been perceived in the data, and of cleaning it up of factors that supposedly belong to the measuring procedure or to external factors, and not to the phenomenon itself. Which idealizations are to be applied is an issue that can be negotiated, and different decisions affect the domain of the quantifier in the statement of universal mechanism.[9] Thus, whether we are prepared to accept the inference from universal mechanism to mechanism about the human mind ultimately depends on whether or not we accept the mind as a phenomenon to be explained.

One way the explanatory burden of mechanism can be reduced is by "eliminating" some would-be phenomenon. For example, pre-Cartesian natural philosophy considered vital processes in animals a phenomenon to be explained. What the data contained, however, was not the vital processes themselves, but observations of seemingly organized spontaneous movement in animal bodies. Mechanistic physiology does not explain vital processes in mechanistic terms ("reduce" them to mechanics), it rather rearranges the data so that vital processes cease to count as a phenomenon (they are "eliminated"). The data is kept the same, but it is idealized differently, into a phenomenon of spontaneous movement, which yields more easily to explanation in terms of inert matter.[10] In a similar fashion, the inference from universal mechanism to mechanism about the mind can be avoided if we take the mind out of the domain of the quantifier. Then, there is simply nothing there to explain. Now, certainly the exclusion of recalcitrant data from the domain of the explanandum flirts dangerously with question-begging. However, since some idealization of the data is anyway unavoidable, ultimately the legitimacy of elimination turns on whether it can be motivated independently of the recalcitrance of the data, and on whether what is left to explain is interesting enough.

The second problem with the argument from universal mechanism is that it is too cheap. Scientific mechanism is predicated on the distinction between positive and vacuous explanations, and on the rejection of the latter from science. Metaphysical mechanism is a metaphysics guided and supported by scientific mechanism. Descartes' metaphysical mechanism about animals, for example, was grounded in positive (though incorrect) physiological theories and actual engineering techniques. But the inference we are now considering proposes that we accept mechanism about the mind on the basis of a general principle, in complete absence of any positive theory of the mind, or of any machine that can simulate it. Such an inference goes against the very grain of mechanism. This doesn't make it a logical fallacy, but it should make us uneasy about accepting its conclusion. The mechanism we end up with is a vacuous doctrine, and we should not be satisfied with it.

We have cited two reasons why mechanism about the mind should not be accepted on the basis of the argument from universal mechanism. Descartes, however, goes farther and rejects universal mechanism altogether.

[9] See (Bogen & Woodward, 1988) for the classical modern statement of the distinction between data and phenomenon, and (Woodward, 2011) for a summary of the ensuing discussion.

[10] This is, at least, how Gaukroger presents things in (Gaukroger, 2007, p. 323ff).

1.3. Cartesian Anti-Mechanism

In the *Discourse*, after having stated his thesis that animals are mere machines, Descartes goes on to say:

> [I]f any such machines resembled us in body and imitated our actions insofar as this was practically possible, we should still have two very certain means of recognizing that they were not, for all that, real human beings […]. The first is that they would never be able to use words or other signs by composing them as we do to declare our thoughts to others. For we can well conceive of a machine made in such a way that it emits words, and even utters them about bodily actions which bring about some corresponding change in its organs […] but it is not conceivable that it should put these words in different orders to correspond to the meaning of things said in its presence, as even the most dull-witted of men can do. (Descartes, 2006, p. 56)[11]

In this passage Descartes puts forth a test for deciding that a humanoid machine is not a genuine human. The claim is that language is a reliable indicator of the presence of mind, and that no machine can simulate human linguistic competence. Note, that Descartes is comfortable with machines voicing sentences, even as a response to stimulation; but that would stop short of genuine linguistic capacity, which consists, first, in the ability to form indefinitely many sentences, and second, in the fact that these sentences are used in accordance with their meaning, hence, that they are meaningful. In other words, we should not consider a mechanistic theory to be a theory of the mind, if it does not account for the syntactic and semantic aspects of language use.[12]

Unfortunately, Descartes doesn't explicitly say why he thinks linguistic capacity resists a mechanistic explanation. Here's a conjecture. Although there is no difficulty in imagining a mechanistic theory that accounts for sound and voice,[13] there is no way to calculate the syntactic and semantic properties of an utterance from its acoustic or even phonological properties alone. Semantic phenomena cannot even be described, let alone explained, in phonological terms. Consequently, there is a basic incongruity between the explanatory resources of mechanism and the phenomenon of linguistic competence, an incongruity that makes genuinely linguistic machines unimaginable for Descartes. Nor does Descartes think that the option of eliminating the mind is open to us (below we'll see why). Linguistic capacity, and with it the mind, presents an ineliminable phenomenon which the explanatory resources of mechanism simply have no chance of accounting for.

[11] In the text, Descartes mentions another test for humanity, which (Gunderson, 1964, p. 199) calls "the action test". I shall not address it here.

[12] See (Gunderson, 1964) for an extensive discussion.

[13] This was, in fact, one of the earliest mechanistic theories, by Beeckman, see (Cohen, 1984, Chapter 4).

It has been suggested that the problem here has more to do with what Descartes can and cannot imagine than with any objective incongruity between machines and the mind. Descartes was familiar with a certain type of machine, and his imagination, remarkable though it was, was inevitably limited to that type. Recall how the pre-Cartesian anti-mechanists, according to the story as we've told it, had difficulty imagining that animals were machines because the machines they knew could not move about on their own. This limitation to the imagination was removed by the appearance of self-moving statues. Similarly, the development of new machines with previously unforeseen features, namely modern digital computers, might provide positive ground for mechanism about the mind. This issue will be taken up presently.[14]

But apart from the perceived incongruity between the linguistic phenomenon and the explanatory resources of mechanism, Descartes also had a more properly philosophical argument for his anti-mechanism, the *cogito*. Although the *cogito* is not expressly presented as an anti-mechanistic argument, its anti-mechanistic import is easy to establish. Briefly, since mechanism for Descartes is metaphysically limited to *res extensa*, it suffices to find one thing which is not *res extensa* in order to refute universal mechanism, even in its vacuous form. The *cogito*'s twin conclusions are, first, that the self[15] exists, and second, that it is a kind of thinking substance, or *res cogitans*, and not *res extensa*. This self is identified with the mind, and it follows that the mind cannot be a machine.

The *cogito* is an interesting case because it resembles a diagonal argument, but on closer inspection it isn't.[16] It resembles a diagonal argument (a) in the form of its conclusion, and (b) in the structure of the argument, as follows. (a) Like many diagonal arguments, the *cogito* (on its anti-mechanistic reading) purports to refute a completeness claim by producing an "outsider" element. For example, Cantor's diagonal proof of the indenumerability of the real numbers refutes the claim that there is a complete enumeration of the reals by producing, for each enumeration, an outsider. In Descartes, the completeness claim is that all things are *res extensa*, and the outsider element is the human mind, or the thinking self. (b) Diagonal arguments typically construct an outsider element by applying a procedure involving self-reference and negation to all members of the putatively complete class. Cantor shows how to construct, for a given enumeration, a real number based on the negation, in the relevant sense, of all members of the enumeration. Descartes' procedure is to doubt the reality of all extended substances; but when he arrives at his own self, he finds that the procedure fails:

[14] The claim that technological advancements might affect our assessment of Descartes' anti-mechanism is suggested by the discussion in (Ben-Yami, 2015, p. 126f).

[15] Or the *I*, or whatever. The *cogito* is awkward to report in the third person.

[16] I defer a more detailed treatment of the *cogito* for another occasion. See (Slezak, 1983; 1988) for a different take on the *cogito*'s being a diagonal argument, (Sorensen, 1986) for a critique.

> But I have convinced myself that there is absolutely nothing in the world, no sky, no earth, no minds, no bodies. Does it now follow that I too do not exist? No: if I convinced myself of something [or thought anything at all] then I certainly existed. (Descartes, 1996, p. 25)[17]

The thinking self is discovered when we try to include it in the domain of our skeptical procedure and fail. However, only the thinking aspect is immune from doubt in this way. It follows that a non-extended object exists.

However, on closer inspection neither the form of the argument in the *cogito*, nor the form of its conclusion, are those typical of diagonal arguments. We show this, again, (a) for the form of the conclusion, and (b) for the structure of the argument. (a) Diagonal arguments typically show the existence, not of an absolute outsider, but of a method to generate an outsider given a particular completeness claim. For example, Cantor does not show that there is a real number which absolutely cannot be enumerated. That's absurd. Rather, he shows how to find, for every enumeration, a real that's outside of it, even if it does belong to some other enumeration. By contrast, Descartes' *res cogitans* is meant to be an absolute outsider, not being captured by any mechanistic system.[18] (b) Diagonal arguments construct the outsider using a negative procedure on the members of the putatively complete system. The identity of Cantor's outsider element for an enumeration *E* is a function of all elements of *E*, negating, as it were, every one of them, and thereby establishing its distinctness from them all. By contrast, in the *cogito*, no diagonal element is constructed. The stage in the *cogito* that resembles the diagonal procedure, quoted above, is the one in which doubt is applied to all things (step A: "I convinced myself that there was nothing at all in the world, no sky, no earth, no minds, no bodies"), including, tentatively, the self (step B: "did I therefore not also convince myself that I did not exist either?"), but unsuccessfully in the latter case (step C: "certainly I did exist, if I convinced myself of something"). Here, what sanctions the inference from self-doubt (B) to self-affirmation (C) is the fact that doubting in general implies the existence of the self, regardless of the object of doubt. But doubting in general was performed already in step (A). Therefore (C) could have been inferred directly from (A). The act of self-doubt (B), ostensibly the diagonal heart of the *cogito*, doesn't play any logical role in the argument. It is primarily an expository device, serving to highlight, but not to establish, the existence of the self. In diagonal arguments, by contrast, the diagonal construction is an essential step of the inference.

[17] The interpolated part is from the French version.

[18] The difference is rooted in the respective claims that diagonal arguments and the *cogito* purport to refute. Diagonal arguments typically refute existential claims. In Cantor's case, the claim is not of the form "every real number is thus and so", but "*there is an enumeration such that* every real number is thus and so". By contrast, Descartes' *cogito* purports to refute the claim "every existent is extended", or something along these lines. This is why a diagonal argument yields only a relative existence claim, and the *cogito* purports to yield an absolute existence claim.

The conclusions from this brief discussion are as follows. First, we see that the *cogito* resembles a diagonal argument, but turns out upon scrutiny not to be one. Second, the aspect in which it fails to be a diagonal argument is exactly the point at which it loses much of its force, since the thinking *I* has not been given a definite enough constitution in order to count as a genuine existent.[19] If this is correct, then we get the impression, wildly anachronistic though it may sound, that Descartes is here groping for a diagonal argument, in a hunch that this is the kind of argument that can refute mechanism about the mind.

1.4. Turing's Computing Machines

With Descartes' pseudo-diagonal anti-mechanist argument out of the way, we can come back to the question, raised in the middle of the previous subsection, whether computing machines can provide positive support for mechanism about the mind in a way that classical machines could not. In order to begin to answer this, we have to state clearly what distinguishes the two kinds of machines.

Today, the term "computing" already implies "machine", but originally the notions were only indirectly related. Computation was just another name for calculation. We get an intuition about what calculation is by looking at a simple case, the grade-school algorithm for addition. The fundamental way to add two numbers, e.g. 13 and 28, is to produce collections, of fingers, say, with the corresponding cardinalities, and then count the members of their union. But counting is not always a good option, and for practical purposes we usually turn to methods that exploit properties of the numerical notation. The positional notation system for numbers, for example, allows us to perform sums of arbitrarily large numbers in terms of the iteration of the operation of summing up two single-digit numbers, in our example case first 3 and 8, and then 1 and 2. Since the possible sums of two single-digit numbers are few, they can be memorized or written down in a small instruction table. When we appeal to such a memorized or written table, reference to the numbers themselves effectively drops out. The table simply instructs us, when we see the digits "3" and "8", to write "1" below them and mark a carry above the next column. What is in play here are the digits, not the numbers 3 and 8, since these latter were not part of the original problem at all. The procedure is iterated for every position of the numerals, resulting in a string of digits, in our example case "41", which we then interpret as referring to a number.

[19] This point deserves elaboration, which, for reasons of space I shall not provide. Briefly, the problem is that the *I* that is discovered in step A is abstract, or vacuous, in something like the sense of the previous subsection, and, therefore, cannot be the basis of a genuine existence claim of the kind that the *cogito* aims to establish. This difficulty is, I believe, recognized by Kant in his discussion of the "*I think* that accompanies all my representations" in the *Critique of Pure Reason*. See esp. the discussion in §16 of the B-deduction (B131ff), where the necessity of the *I think* is affirmed, and in the Paralogisms of Pure Reason (A341/B399ff), where the substantiality of the *I* is denied.

Strictly speaking, therefore, the numbers themselves are only present at the entry and exit points of the calculation, but are completely absent during the calculation itself. The process of calculation is mechanical, in a sense quite close to the one used in relation to machines and mechanism. What it applies to are (usually) marks on paper, and it is sensitive only to their visible geometrical properties. The operations that we perform (writing further symbols) are also definable in terms of geometrical properties. In other words, calculation is performed almost purely within the bounds of *res extensa*. There are two exceptions: the interpretation of the symbols at the input and output points. The meaning of symbols cannot be counted part of their *res extensa* aspect, and yet without acknowledging their meaningfulness, we can't think of our procedure as calculation over and above the mere producing of marks on paper.[20]

What the mechanical nature of calculation provides is epistemic security. Since each step can be written down, surveyed in a glance and compared with the table of instructions, any mismatch will be evident and public. Many philosophers, most notably Leibniz, have entertained the hope of enjoying the epistemic virtues of calculation in domains beyond mathematics. Such a project requires a notation system in which the properties of the subject matter are reflected in the form of the symbols. The formalized languages of modern logic, for example Frege's *Begriffschrift* and Hilbert's deductive systems, can be seen as systems of generalized calculation, designed to be applied to any subject matter. However, these systems implement one particular method of calculation, and the question of a general analysis of the concept of calculation is left open.

Historically, the need for such a general analysis became pressing with the appearance of Gödel's incompleteness theorems. Gödel showed that any formal system, the syntax of which can be captured by recursive functions, and which can represent recursive functions in an appropriate sense, is incomplete. Hilbert's arithmetical system fulfilled these conditions, and was thereby proven incomplete.[21] However, since recursive functions were only one particular form of calculation, there was doubt (at least, Gödel doubted) whether the theorems would apply to any formal calculus whatsoever. This doubt was resolved by Turing's general analysis of calculability, in terms of imaginary computing machines. Turing's machines operate on written symbols, and their simple and highly scalable design enables a mathematical definition of various classes of relations on strings, most importantly the class of computable and computably enumerable (c.e.) relations. It was proven that the class of recursive functions on strings is exactly the class of Turing-computable functions on strings, which showed that Gödel's results hold for all systems with computable syntax. Tu-

[20] This "formal symbol manipulation" account of computing has come under attack in recent decades, e.g. in (Smith, 2002; Fresco, 2014, who provide many further references). However, this literature is mostly concerned with the concept of physical computation, especially in the context of computational cognitive theory, so it is irrelevant here. (It will become more relevant in the next subsection).

[21] Frege's system had, of course, other problems.

ring's account was accepted (in particular by Gödel) as definitive of the concept of computability.[22]

With respect to this history we ask: (a) What role does the notion of a machine play in Turing's account? (b) What convinced Gödel of the account's validity and generality? Regarding question (a), we note that calculation by itself, as exemplified above in the long addition algorithm, makes no reference to machines, though it does depend on the notion of the mechanical. We note also that of the general accounts of calculation that preceded Turing's or were independent of it—Church's lambda calculus, Gödel's and Kleene's general recursive functions, Post's production systems—none mentioned machines in any way. In fact, if we look closely at Turing's account, we see that the reference it makes to machines is, strictly speaking, superfluous. The actual Turing machine construction is nothing but a straightforward generalization of the long addition algorithm, arrived at by simplifying the instruction table to very basic operations. There is no compulsion to describe it as a machine, and we can just as well describe it in psychological terms (as Post did). On the face of it, then, the answer to question (a) is that the notion of a machine plays no role at all in the content of Turing's analysis of computability.

However, when we ask about the perceived conceptual superiority of Turing's account over the other accounts (question (b)), it is hard to avoid the impression that it is due precisely to Turing's appeal to the notion of machine. The point of this appeal is to convince the reader that the procedure described remains squarely within the bounds of the mechanical, or *res extensa*, and, therefore, guaranteed to have the epistemic virtues associated with paradigmatic calculation. The fact that a machine can "perform" Turing's generalized algorithms promises that only geometrical properties of the objects of calculation are appealed to, and in particular that no semantic properties are exploited. This is, however, ascertainable even without the machine trope, for example if we interpret Turing's algorithms as instructions for human computers. The machine trope ultimately has just an auxiliary expository role in Turing's argument.[23]

In order for the machine trope to fulfill its expository role, the machines in question have to be exactly the classical mechanical machines, the ones that Descartes was thoroughly acquainted with. There is no impediment to realizing

[22] This story is told in many places, for example in (Sieg, 2013).

[23] For Gödel's well-known endorsement of Turing's account, see for example the 1951 essay in (Gödel, 1986): "The most satisfactory way [of arriving at a definition of computable function of integers] is that of reducing the concept […] to that of a machine with a finite number of parts, as has been done by the British mathematician Turing".

See (Sieg, 2013, §1) for more quotes by Gödel to a similar effect. See also (Sieg, 2001) for a more general discussion. Gödel does mention machines already before the appearance of Turing's work. In his 1933 paper (Gödel, 1986), inference rules in formalized languages are characterized as: "[P]urely formal, i.e. refer only to the outward structure of the formulas, not to their meaning, so that they could be applied by someone who knew nothing about mathematics, or by a machine".

Turing's machines with gears, chains and a crank rather than scanners, printers and tapes. Indeed, but for their scalability, Turing's machines would be much simpler to engineer than the average moving statue. But if computing machines are not essentially different from mechanical machines, why do we have the impression that they stand a better chance than the latter of simulating human cognition? Conversely, why did Descartes fail to see that classical machines are capable of such simulation?

The answer to this puzzle has been given in our discussion of the long addition algorithm above. There, we mentioned two points in the procedure of calculation that went beyond the merely mechanical: these were the input and output points, at which the meanings of the symbols were appealed to. If we ignore these semantic limit points, the computer, whether human or machine, cannot properly be said to compute, but only to be moving bits of *res extensa* about (or rather, to be bits of *res extensa* moving about). It is the person writing the input on, and reading the output off, the tape who is performing the computation, using the machine as they would a slide rule or a sophisticated abacus.[24] In order to treat machines as genuinely computing, we have to revise our notion of a machine. We now include the semantic limit points of the computation procedure within our notion of computation, and accordingly, we include the interpretation of the symbols on the machine's tape as belonging to the machine. Thus, we distinguish between, on the one hand, Turing's machines, which are strictly mechanical and used by Turing as an expository trope in his classic account of (human) calculation; and on the other, Turing machines which are to Turing's machines as a closed interval is to an open one—they include the limit points. Turing machines are not just bits of *res extensa* moving about, for their symbols are not mere geometrical shapes—they are symbols properly so-called, i.e. bits of *res extensa* that designate things. Only in this sense can we say that the machine returning, say, the string "41" upon receiving the strings "13" and "28", computes addition. It is the concept of a Turing machine, and not a Turing's machine, that stands a chance at simulating human cognition.

Recall, Descartes' worry was that the semantic aspect of linguistic competence was incongruent with the explanatory resources of mechanism. On our new understanding of computing machine, the semantic layer is built-in. Should Descartes be satisfied? Probably not. The new machines might be said to compute, but they also explicitly go beyond the purely mechanical. It is therefore not correct to say that we have shown, by appealing to computing machines, that mechanism about the mind as Descartes understood it is tenable. Rather, we have changed the subject.

[24] See the first three sections of (Papayannopoulos, 2020) for a more detailed statement of roughly this view.

1.5. The Cognitive Inversion

The main empirical enterprise that makes use of the theory of computability as the basis for a mechanistic outlook is the computational theory of mind. This theory emerges from the concept of computation through two conceptual twists.

First, as we recounted above, the notion of computation originally referred to a species of conscious human activity, not something specifically related to machines. Admittedly, due to the mechanical or rote character of computation, no emphasis needed to be placed on "conscious". But calculating is something a person does intentionally, not something done in the background. In Turing, I claimed, the appeal to the idea of a machine was just an expository trope. However, once it was shown in theory what such machines could do, it was a (relatively) small step to building them in practice. The success of this project was so overwhelming that the word "computing" and its cognates came to be associated exclusively with machines. This is a conceptual twist. Turing exploited the fact that both human computation and machines are mechanical, in order to argue that his mathematical model captures human computation. Actual computing machines do the conceptually opposite—they exploit the mechanical character of human computation in order to take over the rote part, leaving humans the sole job of interpreting the results.

Once actual physical computers became available and familiar, it was a (relatively) small step to using them as a model for human unconscious cognitive activity. In digital computers we distinguish between the hardware, which is the machine itself, and the software, which is, roughly, a representation of the design of the machine which abstracts from any physical implementation. This distinction allows us to implement several abstract machines on a single physical machine.[1] Analogously, one is tempted to view the brain as a piece of naturally developed computer hardware, on which various software programs are implemented. The software corresponds to the human mind. Cognitive phenomena are then explained in terms of software implemented in the brain. In this way the notion of computation, which started out as pertaining to the human mind, made its way back to the mind after having been appropriated by machines. And it came back transformed, being now postulated to underlie the whole of human unconscious cognitive makeup, rather than being one type of conscious human activity.[2]

The idea of a computing machine thus provides a basis for a methodlogically sound and empirically fruitful science of the human mind. Does this provide support for metaphysical mechanism about the mind in the same way that mechanistic physiological theories provided support for metaphysical mechanism about animals?

[1] For complications regarding the hardware/software distinction, see (Duncan, 2017).

[2] See (Gardner, 1987) for a detailed history of cognitive science (with relation to the present paragraph, see pp. 16ff, 40f, 138ff, 384ff).

Ideally, a full mechanistic theory of the mind would show how cognitive phenomena follow from the physical description of the brain in the same way that the behavior of a digital computer follows from its physical makeup. However, currently at least, we are nowhere near such a full derivation. In practice, cognitive science makes progress by abstracting from the physical implementation, the hardware, of the mind, and studying just the software, the system of conscious and unconscious mental processes and representations that underlie human capacities and behavior. For example, Chomsky explains the cognitive phenomenon of linguistic competence by postulating a specialized language faculty, described as an "abstract linguistic computational system" which is "an internal component of the mind/brain" (Hauser, Chomsky, & Fitch, 2002, p. 1570f). By "abstract" what is meant, supposedly, is that the physical implementation of the computational system is abstracted from.

This strategy, of abstracting from the physical implementation, has implications for our question. Computation, as we have understood it in the previous subsection, is the mechanical manipulation of strings of symbols, such that the symbols are visually, or anyway sensibly, individuated, in other words that they are *res extensa*. This bottom-up character is what endows computation with its epistemic virtue, and also what connects it with the doctrine of mechanism. With computing machines, things become a little more complicated because we add a non *res extensa* layer, the level of interpretation; but the bottom-up character of computation is kept, because the symbols are still manipulated strictly mechanically. However, in the mind there is nothing that corresponds to the visible strings of symbols of conscious computation, and the abstraction from hardware eliminates all reference to the physical substrate of the computing machine. The representations that the computations of cognitive science operate on are individuated top-down, by their systematic contribution to the computation of the phenomena, or in other words, functionally. Such functionalism does not, perhaps, invalidate cognitive science as a science; but it seems to waive the mechanistic demand for positivity.[3]

Another way to state the problem is this. As far as I know, Chomsky never fully specifies what is meant by "/" in "mind/brain" in the quote two paragraphs above. The intention is probably to flag the fact that, although abstract representations are immaterial, no metaphysical dualism is thereby implied, since we expect them to be reduced to neural terms at some time in the future. On this understanding, the "brain" in "mind/brain" is a promissory note to the effect that cognitive science (in this case linguistics) will, at some point, be reconciled with mechanism proper. Read in this way, the phrase "mind/brain" is an implicit endorsement of mechanism about the mind. However, the promissory note is a rain check, not motivated by any existing positive theory of the relation between

[3] See (Miłkowski, 2013) for a relatively nuanced discussion of mechanism about the mind in the context of cognitive theories and the so-called new mechanism. Miłkowski's view of computation is not the one presented in the previous subsection.

linguistic capacity and anything in the brain; it is motivated solely by the widespread conviction that everything cognitive has its seat in the brain. This conviction is too widespread to doubt, at least in certain prominent circles, but this is not the same as being a positively discovered empirical fact. To the extent that it purports to be more than just a methodological injunction to seek explanations of cognitive facts in the brain, in other words, to the extent that it presumes to be a metaphysical thesis, it is a vacuous mechanism, in the sense of §1.2.[4]

In this subsection and the previous we considered, all too briefly, the notion of a computing machine and its application to the empirical study of the mind. The question was whether computing machines can provide positive ground to mechanism about the mind in the same way that moving statues and mechanistic physiological theories grounded mechanism about animals for Descartes. On the one hand, the new conception of computing machine, and the cognitive science built upon it, give up on many of the features that made mechanism attractive—the inertness of matter, the bottom-up derivation of phenomena, etc. On the other hand, the empirical success of cognitive theory, as well as the engineering achievements in the field of computing machines, show that the new conception is stable and fruitful. I leave the issue undecided. I turn now to consider whether we cannot find a principled argument against mechanism in Gödel's incompleteness theorem.

2. The Gödelian Argument

Our foregoing exposition of computational mechanism about the human mind was not sufficiently precise for Gödel's theorem to be applied to it. In the present section, our tasks are, first, to provide a sharp(ish) formulation of mechanism; second, to give a correspondingly sharp rendering of Lucas's famous Gödelian anti-mechanist argument; and finally, to topple this argument from several angles. In closing, I shall sketch a diagonal argument that I think stands a better chance.

2.1. Mechanism About the Mind

Mechanism about the mind, on the construal sketched in §1.5 above, is the claim that every natural aspect of human cognition, or in other words every cognitive phenomenon, is the result of a computational system that is part of the mind. Cognitive science studies many phenomena that don't manifest themselves through language, but they are arguably not relevant to the issue at hand, and

[4] This is not to say that no connection has been made between linguistic theory and the brain sciences. See in particular the findings reported in (Grodzinsky & Santi, 2008) and papers cited therein. However, it is clear that these findings are very far from sufficient to metaphysically ground Chomskyan theory in the brain sciences. They should rather be considered the fruit of Chomskyan mechanism, where the latter is viewed as a research program, not a metaphysical thesis.

I put them aside. We therefore think of a cognitive phenomenon as a set of sentences uttered or assented to by speakers, and of a mechanistic explanation as an algorithm that enumerates the set.

We assume that natural languages are fully intertranslatable, and identify them all with a single language L, which we assume for convenience is a formalized first-order predicate language. In addition, we assimilate to L the language in which the mental computations are carried out, the Language of Thought, as it were. Such assumptions seem to be implicit in much of the practice of (linguistically oriented) cognitive science. Let S_L be the set of sentences of L. Cognitive phenomena are associated with certain subsets of S_L, which I shall call their yield. Scientific mechanism is the call, given a cognitive phenomenon A, to look for an algorithm that enumerates its yield S_A. Metaphysical mechanism is the claim that the mind really is computational, which we express as follows:

Metaphysical Mechanism: Let L be the language of the mind. If $S_A \subseteq S_L$ is the yield of a natural human cognitive phenomena A, then S_A is computably enumerable (c.e.).

By our assumptions, S_L itself is infinite and c.e. (in fact, computable). Being infinite, it will have non-c.e. subsets. On pain of trivializing the question, mechanism therefore cannot be equated with the claim that all subsets of S_L are c.e. We need some means of characterizing the class of subsets that are of interest, i.e. that correspond to cognitive phenomena.

Cognitive science is an empirical discipline, one that studies phenomena as they are given in observation and experiment. Let's further restrict the experimental paradigm to that in which sets of sentences (or judgments about sentences) are collected from subjects, whether they occur naturally or through elicitation. In practice, experiment and observation can only give rise to finite sets of perceived sentences. Again, on pain of trivializing the question, we cannot assume that the yields of phenomena are finite, since finite sets are trivially computable. To bridge the gap between the finitude of the sets actually given, and the infinitude of the sets yielded by cognitive phenomena, we call on our previous distinction (§1.2) between data and phenomena, and on our notion of idealization whereby the latter is constructed from the former. This allows us to consider infinite subsets of S_L as cognitive phenomena. Clearly, much hangs on which idealizations are allowed.

In order to have an actual case before our eyes, let's think again about Chomskyan linguistics, and in particular the cognitive phenomenon of linguistic competence. The premise of linguistics is that there is a language faculty which computationally enumerates, or "generates", the set of sentences that speakers accept as grammatical in their language. It is the job of the linguist to find the generating algorithm.[5] The phenomenon of linguistic competence is associated with an

[5] See, e.g., (Chomsky, 1957; 1975).

infinite set S of sentences. The data available to the linguist has to be finite.[6] S is therefore the product of idealizations performed on a finite data set D. The operations involved are roughly two: extrapolation to an infinite set, and cleaning up the data into a well-behaved collection that exhibits enough regularity to be studied scientifically.

Let's look closely at an (unrealistic) example of idealization. Let s_J be the sentence:

(1) John is very, very tall.

Let $s_J\{n\}$ be the result of replacing "very, very" in s_J with a string of n times "very". Since data sets are finite, no data set will contain $s_J\{n\}$ for every n. However, clearly we could procure a data set D_J such that $s_J\{n\} \in D_J$ for every n less than some integer k, i.e. with no gaps below k. It seems reasonable to extrapolate D_J to a set S_J, such that $s_J\{n\} \in S_J$ for every n. Note that, clearly for some integer l, sentences $s_J\{n\}$ longer than l will not appear in any data set. They will simply be too long. But this shouldn't discourage us from accepting S_J, since we can reasonably ascribe the absence to factors that lie outside the language faculty proper, for example to constraints on memory or on patience.

In addition, and especially if D_J is drawn from a corpus of naturally occurring speech rather than elicited speaker judgments, there may be sentences in D_J that we refrain from carrying over to S_J. Naturally occurring speech is the product of many heterogeneous factors, the language faculty being just one. The subject may be distracted midsentence, or interrupted, or there might be another reason for us to decide that an observed sentence does not reflect a genuine product of our language faculty. In this way not only will there be many sentences in S_J that were not directly given in D_J, but also sentences that were given can be filtered out of the phenomenon to be explained. S_J is, therefore, both extrapolated and pruned, relative to the data set D_J.

This doesn't mean that anything goes. There have to be constraints on which extrapolations and which prunings are legitimate, constraints which I shan't attempt to specify precisely. Instead, let me give an example of an illegitimate idealization. Consider the following experiment. The subject is presented with a natural number n, and is asked to form a grammatical sentence with n words. This is a task that linguistically competent subjects can perform with ease, for example by giving $s_J\{n - 3\}$ as an answer (for $n > 2$, of course). Let D_B be the set of sentences actually collected in the experiment, a finite set. Now let B be some non-c.e. set of natural numbers with an initial segment identical with the set of lengths of sentences in D_B. Finally, let S_B be a set of sentences such that $D_B \subset S_B$, and such that if $s \in S_B$ and n is the length of s, then $n \in B$. In other words, B is the set of lengths of sentences of S_B, and S_B is an extrapolation of D_B, conditioned by

[6] See (Pullum & Scholz, 2010) for a critique of the assumption that an infinite set has to be assumed.

B. It follows that B is Turing-reducible to S_B—all we need to do is count the lengths of sentences in S_B—and therefore that S_B is non-c.e. But S_B was extrapolated from the (imagined but) plausible data set D_B. If this extrapolation is a legitimate idealization, then S_B is the yield of some cognitive phenomenon, and thus a counterexample to computational mechanism.

Clearly, S_B is not a legitimate idealization of D_B. I will not attempt a statement of general conditions on legitimate idealization, but the condition that this example suggests is that extrapolation has to preserve and continue trends existing in the original set. The concept of a trend and its continuations is not a very precise or determinate notion, but since B was chosen arbitrarily, it obviously does not fit the bill. In the case of D_B, the idealization is conditioned by a set that is completely external to the mind, so the fact that it is not c.e. is not a counterexample to mechanism after all. More generally, when we idealize a data set into a phenomenon, the principle that guides the idealization must somehow reflect the situation in the mind.

2.2. Gödel's Theorem and Its Proof

With this sharpened statement of mechanism in hand, let's turn our attention to the Gödelian argument. First, let's review Gödel's theorem and proof.

Let L_T be a formalized language.[7] A set $T \subseteq S_{L_T}$ is a theory if it contains the logical axioms and is closed under the logical inference rules; it is a formalized theory if it is, in addition, c.e.; and it is said to contain arithmetic if it contains the Peano axioms (L_T is therefore implied to contain the language of arithmetic).[8] T is consistent if $T \neq S_{L_T}$, equivalently if for no L_T sentence α: $\alpha, \ulcorner\neg\alpha\urcorner \in T$.[9]

Theorem G1: *If T is a consistent formalized theory that contains arithmetic, then we can compute from (the algorithm that enumerates) T a sentence $g_T \in S_{L_T}$ such that g_T, $\ulcorner\neg g_T\urcorner \notin T$.*

Proof: Let $c(x)$ be a mapping from numbers to L_T sentences, called the coding scheme. $\ulcorner\overline{n}\urcorner$ is the numeral in L_T that refers to the number n.

[7] As before, I limit consideration to standard first-order languages.

[8] It is possible to state the theorem also for weaker conditions, but this will not affect the argument.

[9] The corner quotes are used in order to form names of expressions by concatenating symbols with other names of expressions. "α" in the text is a variable over sentences; "$\ulcorner\neg\alpha\urcorner$" is a function taking a sentence and returning its negation. The source is (Quine, 1940, p. 33ff), though I also allow constant names (e.g., "g_T" below) to occur in corner quotes. This is not exactly the same as the (more common) use of corner quotes to signify Gödel codes.

Lemma 1 (Reflection): *There is a unary formula $PRV(x)$ of L_T, such that for every n:*

$c(n) \in T$ *if and only* $\ulcorner PRV(\bar{n})\urcorner \in T$.

Lemma 2 (Diagonalization): *There is a number k (which can be computed from T) such that:*

$\ulcorner \neg c(k)\urcorner \in T$ *if and only if* $\ulcorner PRV(\bar{k})\urcorner \in T$.

From the two lemmas it immediately follows that:

(2) $c(k) \in T$ *if and only if* $\ulcorner \neg c(k)\urcorner \in T$.

We put g_T for $c(k)$. By consistency of T, the theorem follows. □

The sentence g_T effectively says of itself that it is not in T. Consequently:

Corollary: *Under the hypothesis of the theorem, g_T is true.*

2.3. The Gödelian Anti-Mechanist Argument

The Gödelian argument applies theorem G1 to the sharp statement of mechanism given in §2.1. The language in question will be the general language of cognition L. Call a set $T \subseteq S_L$ Gödelian if it is a consistent formalized theory that contains arithmetic. For each Gödelian set T, by G1 and its corollary, we have a true sentence $g_T \notin T$. Let $S_G = \{g_T\colon T \text{ is Gödelian}\}$. Clearly no superset of S_G is Gödelian. Otherwise put:

Fact: A consistent superset of S_G that contains arithmetic is not c.e.

For the anti-mechanist it therefore suffices to find a cognitive phenomenon A such that:

(a) S_A contains arithmetic,

(b) S_A is consistent,

(c) $S_G \subseteq S_A$.

Condition (a) points to an immediate suspect. On the model of Chomsky's approach to language, we posit a human cognitive faculty C, which accounts for our arithmetical competence—our ability to recognize the truth of arithmetical sentences. Clearly, S_C contains arithmetic in the appropriate sense.

Why assume, as per condition (b), that S_C is consistent? On the face of it this seems false. After all, individuals often make mistakes and change their minds about arithmetical statements, resulting in inconsistencies in the accumulated set of accepted sentences. However, it is clear the arithmetical judgments that people actually make, the data set, do not fully reflect their cognitive arithmetical faculty, if such there is. Following the Chomskyan practice outlined in §2.1, we allow S_C to be an extrapolated and pruned extension of the set of arithmetical sentences that are actually asserted. First, the finite data set (the set of actually uttered arithmetical judgments) is extrapolated into an infinite set (this was already implicitly assumed for condition (a)). Second, it is cleaned up by pruning it of inconsistencies and, perhaps, of falsehoods generally. The Chomskyan procedure for arithmetical competence is therefore assumed to result in a sound, or at least consistent, set S_C.[10]

In order to show condition (c), that $S_G \subseteq S_C$, the anti-mechanist appeals to the corollary to G1, in which g_T is proven for arbitrary T. The reasoning here is that g_T is mathematically proven (though not, of course, in a formal system), which is to say recognized as true. Since C was characterized as the ability to recognize as true arithmetical sentences, we have $g_T \in S_C$, and since T was arbitrary, we have $S_G \subseteq S_C$.[11]

Since conditions (a, b, c) hold, by our Fact above it follows that S_C is not c.e. By our characterization of Metaphysical Mechanism (§2.1), it follows that mechanism is false. This is the Gödelian anti-mechanist argument. We now turn to its refutation.

The arguments for all three conditions contain serious problems. First, in proving that condition (c) holds, we assumed that all g_T's are proven true. For this we have relied on the corollary to G1. However, the corollary carries over the hypothesis of the theorem, namely that T is Gödelian, and in particular, consistent. The sentence g_T for a particular T is only proven by the corollary if T is consistent. But it is no claim of the anti-mechanist argument that we can see whether a given arithmetical theory is consistent or not. Nor is this a plausible premise to add to the argument. But without it, it is not the case that we have proved the g_T's, not even informally, and, therefore, it is not the case that $S_G \subseteq S_C$. What we have proven is the set of conditionals $S_G^* = \{\ulcorner\text{if } T \text{ is consistent, then } g_T\urcorner\}$, for T c.e. and containing arithmetic. But S_G^* is certainly c.e., and so are many of its supersets. It was, therefore, not shown that S_C is non-c.e.[12]

Though this objection seems conclusive, the other problems I shall mention are arguably more illuminating philosophically. The first problem is with the appeal to arithmetic in condition (a). The second is with the idealization performed in the appeal to competence in condition (b).

[10] Compare (Shapiro, 1998, p. 275).

[11] Compare this with the moves in (Lucas, 1961) and (Penrose, 1989).

[12] An early statement of this objection is in (Putnam, 1960). See also (Bowie, 1982) and (Krajewski, 2020).

2.4. Against Arithmetic

The first point to become aware of is that, strictly speaking, it is not arithmetical competence that is doing the work in the Gödelian argument. By a specifically arithmetical competence we mean, I assume, the kinds of specifically arithmetical reasoning that we perform in order to reach arithmetical conclusions. In formalized theories, "specifically arithmetical reasoning" means logical reasoning from arithmetical axioms. Establishing an arithmetical theorem by means of, say, a set-theoretic proof can hardly count as an exercise of our pure arithmetical ability, but seems clearly to go beyond it and rely on additional resources. From the other direction, it seems that our specifically arithmetical competence is all but idle in our knowledge of the truth of arithmetical sentences such as "$2 = 2 \vee 2 \neq 2$". It is, therefore, not just the character of the theorem proved that determines which cognitive competence is responsible for its knowledge, but also the character of the proof.

Assume, contrary to fact, that theorem G1 does allow us to prove, as a corollary, all members g_T of the set S_G. Formally, the Gödelian argument would then go through, since we would have proven all members of a non-c.e. set. However, it would be wrong to say that we proved them using our arithmetical competence. The reasoning that led us to accept the corollary to G1 has nothing to do with arithmetic. It is justified only by the equivalence of g_T with its own unprovability, which is a metamathematical, not an arithmetical, fact. Granted, g_T itself is, in principle, an arithmetical sentence; but its specific arithmetical content is completely abstracted from in the proof, and is anyway dependent on the coding function $c(x)$, from the precise content of which we have also abstracted. All we rely on in the proof of G1 and its corollary are the metamathematical properties of g_T. Thus, even if the Gödelian argument had been valid formally, it would not show that arithmetical competence is non-computational.

The reason that the Gödelian anti-mechanist appealed to arithmetical competence is that Gödel's theorem is ostensibly about arithmetic. Looked at more carefully, however, we see that the connection between G1 and arithmetic is not that straightforward. Technically, the role that arithmetic plays in G1 is the fact that arithmetical theories represent, in the proof-theoretic sense, all recursive relations between numbers.[13] Given the well-known connections between recursive relations on numbers and computable relations on strings, this means that arithmetical theories can represent computable relations between strings. This is the point at which G1, in its classic formulation, makes contact with the notion of computability. The diagonalization function and the provability predicate are, respectively, a computable and a c.e. relation on strings, whence our Lemmas 1 and 2 in the proof-sketch above.

[13] In what follows I'll be loose and say "represent recursive (computable) relations" as shorthand for "strongly/weakly represent recursive/r.e (computable/c.e.) relations".

Once we highlight this, however, it becomes clear that the coding function and the use of recursive relations are just a detour. Theories are sets of strings, and provability in a theory T (for us, the predicate "$x \in T$") is a property of strings. In order to state Lemma 1 in the above proof (repeated here for convenience), we appealed to the coding function $c(n)$ on one side of the equivalence, and to the predicate PRV on the other:

(3) $c(n) \in T \Leftrightarrow \ulcorner PRV(\bar{n}) \urcorner \in T$

Neither the coding function, nor the possibility of mentioning predicates, are part of Gödel's formalized arithmetical object-theory. This is clear from the fact that the language of the object-theory doesn't, in the general case, contain reference to strings. Both the coding function and the term referring to the provability predicate are defined in Gödel's unformalized metalanguage. If we were to formalize the metalanguage, we would have to include strings in its domain, alongside numbers. However, once we can refer to strings, all reference to numbers can be dropped. The syntactic concepts, in particular the provability predicate, are originally defined in terms of strings.[14] The form of Lemma 1 is simpler when stated for theories T that contain string theory instead of number theory:

Lemma 1* (Reflection): *There is a unary formula $PRV^*(x) \in L_T$, such that for every sentence $s \in L_T$,*

$s \in T \Leftrightarrow \ulcorner PRV^*(\bar{s}) \urcorner \in T$.[15]

Lemma 1* is a product of the fact that $PRV^*(x)$ is a c.e. relation, and that string theories proof-theoretically represent such relations. Since theories are understood as sets of strings, reference to strings is anyway unavoidable, unlike reference to numbers and, hence, Lemma 1* is a more basic statement of the fact expressed by the original Lemma 1 above. The references to coding and to recursive relations in Lemma 1 are just a detour through arithmetic that allows us to

[14] Gödel himself, in the original paper (Gödel, 1953), refers to the more or less string-theoretic syntactic definitions in Łukasiewicz and Tarski (Tarski, 1956). Łukasiewicz and Tarski define the syntactical notions using set-closure definitions, which are the explicit higher-order counterparts of recursive definitions. Being higher-order, they are not computable. One of Gödel's crucial contributions in (1953) was to show how sequences can be coded into single numbers, allowing him to give, in the case of the syntactic notions, explicit counterparts of recursive definitions without going higher-order. The detour through arithmetic was thus necessary for Gödel at the time, in order to be able to code sequences. However, string sequences too can be coded in terms of single strings, though this technique was probably not available to Gödel. See (Quine, 1936; 1944) for work in string theory (concerned with elementary, not computable, relations). See (Grzegorczyk, 2005) for a development of Gödel's results in a string-theoretic setting.

[15] $\ulcorner \bar{s} \urcorner$ is the name of the string s.

apply Lemma 1* to a special case. In fact, G1 can be applied to any theory which represents c.e. relations, whether it deals with strings, numbers, sets or what have you. The corresponding form of the theorem is:

Theorem G1*: *If T is a consistent formalized theory that represents computable relations, then we can compute from T a sentence* g_T *such that* $g_T, \ulcorner\neg g_T\urcorner \notin T$.

The upshot is that arithmetic is not part of the essential subject matter of G1 at all. G1 is a metamathematical theorem about formal proof systems in general, if they capture computable relations.

This suggests that arithmetical competence is the wrong cognitive phenomenon to use in a Gödelian anti-mechanist argument. It is simply not the right setup for an application of G1. We might conjecture some other kind of competence, more in tune with the essential content of G1; but whereas a natural arithmetical competence is somehow easy and smooth to postulate, there is no obvious natural cognitive competence that corresponds to the subject matter of G1 in the way required. Should we say we have a natural metamathematical competence? Or some general epistemological faculty? It is not clear that any reasonable form of mechanism has to be committed to this.

The upshot of the foregoing is that, even if the formal objection of the previous section did not apply (*per impossibile*), still the argument would not achieve its purpose, since it does not show that our arithmetical competence is non-c.e. One comment before we move on. From the foregoing discussion one might get the impression that Gödel's theorem is only accidentally connected with arithmetic, and that it is only a historical accident that the theorem has been discovered through its application to arithmetic as a special case. Inasmuch as Gödel's theorem is ultimately about theories, not about numbers, and theories are made up of strings, this impression is correct. However, the connection between arithmetic and string-theories is not just a historical accident. As we know, both the theories and the structures of strings and of numbers are very similar, and any result about the one can be expressed in terms of the other.[16] Indeed, numbers have a philosophical advantage over strings in that they constitute a single natural domain, whereas when considering strings concretely we have to fix a particular alphabet arbitrarily. We can say that the domain of numbers distills the invariant element in string domains. That would be a sense in which G1 is about arithmetic after all. In any case, it is not immediately about arithmetic, and making the connection clear and explicit will require further work.

2.5. Against Competence

The second philosophical problem with the Gödelian argument is with the notion of competence. In particular, the idealization performed in generating the

[16] See (Corcoran, Frank, & Maloney, 1974; Svejdar, 2008) for more on this.

phenomenon S_C from the set of actually uttered arithmetical sentences (see condition (b) above) is not legitimate.

Recall, following Chomsky, we have allowed the yield of cognitive phenomena, our "competences", to be extrapolated and pruned from data sets. But not every extrapolation and pruning would work. We cannot, for example, idealize away from the phenomenon a pattern of usage found in the data just because it doesn't conform to the theory we are inclined to accept. Nor can we idealize into the phenomenon something that didn't exist in the data set, unless we can be convinced that it reflects something in our cognition, and that some factor related to performance blocks it from being manifested in the data set. For example, we cannot condition our idealization on some decidedly external factor, like the set B from §2.1.

With regard to the Gödelian argument, the question is whether the finite and inconsistent set of humanly asserted arithmetical sentences, the arithmetical data set, can legitimately be idealized into an infinite consistent set. The analogy with Chomskyan grammatical competence is misleading. Granted, Chomsky sometimes speaks of grammatical competence in terms of (tacit) grammatical knowledge, and in our case too we speak of arithmetical knowledge. But there is an important sense in which grammar doesn't behave like knowledge at all. Knowledge, as usually understood, describes a belief that could have been false, and happens to be true. For example, if I know that $2 + 2 = 4$, this implies that I could have falsely believed that $2 + 2 > 4$. And my belief counts as knowledge partly because, in point of fact, $2 + 2 = 4$ is the case. Nothing corresponds to this in grammatical competence. There is no external domain of independent facts to which grammatical knowledge needs to correspond. No one is ever mistaken with respect to their tacit beliefs about grammar. It makes no sense to say that the sentence:

(4) John is very, very tall,

is both ungrammatical (for someone) and generated by that person's grammar. There is no external norm against which "knowledge" of language can be assessed.

Compare this now to the case of arithmetic, and to the idealization of arithmetical competence. The kind of knowledge that arithmetical competence is supposed to furnish its bearer with is genuine knowledge, one that refers to an independent reality. Unlike in the case of tacit grammatical knowledge, there is a sense in which we can say that we could have been wrong, that our arithmetical competence could have yielded $2 + 2 > 4$, in contradiction to actual fact. In arithmetic, there is, unlike in grammar, an external standard to which knowledge is compared.

This, I submit, makes the idealization of linguistic competence, relied on in the Gödelian argument above, illegitimate. If, in constructing a phenomenon, we base ourselves on something we know is external to the mind, like the set B of

§2.1, then the result cannot be counted a natural human cognitive competence, and mechanism about the mind makes no pronouncement about it. The mind might be thoroughly computational in the sense that the belief system of an individual is c.e., and yet the set of beliefs which constitute knowledge depends on factors external to the computational description.

This problem holds especially clearly if the assumption is that the arithmetical faculty is sound (not just consistent), and arguably this is the assumption that the anti-mechanist needs. But even if we only assume consistency, there is still here a reliance on an external standard, this time the truth of logical sentences. The issue is simply transferred to the question of logical competence, and here it is again soundness that is at stake.

To sum up, the Gödelian argument is mistaken in its assumption that we can treat arithmetical competence as having a consistent yield. This was one of the main premises of the argument, and so the argument fails.

2.6. The Tarskian Argument

In this section we have reviewed the Gödelian anti-mechanist argument and found, not only that it contains a formal fallacy, but also that its basic premise, the juxtaposition of human arithmetical competence with formalized systems, is deeply misguided. To conclude the paper, let me briefly sketch an anti-mechanist argument that I think has better prospects. This argument is based on Tarski's indefinability theorem, so it is also a kind of diagonal argument. Let's review, first, the theorem and its proof. Say that a language *M*, expressing the truth predicate for a language *L*, is *L*'s *metalanguage*. *L* is then the object-language. A language is *semantically closed* if it is its own metalanguage, $L = M$.

Theorem T1: *No language is semantically closed.*

Premise 1 (Reflection, Convention T): If $TRUE(x)$ is a truth predicate for *L* in a metalanguage *M*, then for every $s \in S_L$, the following sentence holds:

$$\ulcorner TRUE(\bar{s}) \leftrightarrow tr(s) \urcorner,$$

where $\bar{s}$ is the *M* name of *s*, and $tr(s)$ is the *M* translation of *s*.

Premise 2: If $L = M$, then there is a sentence $k \in S_L$, such that the following sentence holds:

$$\ulcorner TRUE(\bar{k}) \leftrightarrow \neg k \urcorner.$$

From the two premises (and the fact that $\ulcorner k \leftrightarrow tr(k) \urcorner$ when $L = M$), we the following for the case that $L = M$:

(5) $\ulcorner k \leftrightarrow \neg k \urcorner$.

By reductio, the theorem follows.

The Tarskian anti-mechanist argument has the following premises. First, though Tarski's own concern was with formalized languages, today, mainly following Davidson, one often uses the general structure of Tarski's theories in the other direction, i.e. assuming truth to be understood and understanding the truth-conditional statements as providing a semantics for the language under consideration. We apply this approach to the semantics of the language of thought *L* (Fodor & Pylyshyn, 2014). The second premise of the argument is that the language of thought *L* can express any scientific theory. This is forthcoming if we accept that scientists cognize the theories that they put forth, and, therefore, that their language of thought should be able to express them. The third and final premise is that a full mechanistic theory of the mind needs to contain a semantic theory for *L*. For if it doesn't, then it can't with justice be seen as giving the content of the mental states of human subjects, and the characteristic property of the mind is its ability to entertain contents. However, together the three assumptions make *L* semantically closed, and this is impossible by theorem T1. The consequence is that no full mechanistic theory of the mind is forthcoming.

REFERENCES

Anstey, P. (2000). Descartes' Cardiology and Its Reception in English Physiology. In J. Schuster, S. Gaukroger, J. Sutton (Eds.), *Descartes' Natural Philosophy* (pp. 420–444). London, New York: Routledge.

Ben-Yami, H. (2015). Descartes' Philosophical Revolution: A Reassessment. London: Palgrave-Macmillan.

Bogen, J., Woodward, J. (1988). Saving the Phenomena. *Philosophical Review*, *97*(3), 303–352.

Bowie, G. L. (1982). Lucas' Number Is Finally Up. *Journal of Philosophical Logic*, *11*(3), 279–285.

Cantor, G. (1890). Ueber Eine Elementare Frage Der Mannigfaltigketislehre. *Jahresbericht Der Deutschen Mathematiker-Vereinigung*, *1*, 72–78.

Chomsky, N. (1957). *Syntactic Structures*. Berlin: Walter de Gruyter.

Chomsky, N. (1975). *The Logical Structure of Linguistic Theory*. Berlin: Springer Science+Business Media.

Cohen, H. F. (1984). *Quantifying Music: The Science of Music at the First Stage of Scientific Revolution 1580–1650*. Berlin: Springer Science & Business Media.

Corcoran, J., Frank, W., Maloney, M. (1974). String Theory. *Journal of Symbolic Logic*, *39*(4), 625–637.

Cottingham, J. (1978). 'A Brute to the Brutes?': Descartes' Treatment of Animals: Discussion. *Philosophy*, *53*(206), 551–559.

Descartes, R. (1996). *René Descartes: Meditations on First Philosophy: With Selections From the Objections and Replies* (2nd ed.). Cambridge University Press.

Descartes, R. (2006). *A Discourse on the Method of Correctly Conducting One's Reason and Seeking Truth in the Sciences*. Oxford University Press.

Duncan, W. D. (2017). Ontological Distinctions between Hardware and Software. *Applied Ontology*, *12*(1), 5–32.

Fodor, J. A., Pylyshyn, Z. W. (2014). *Minds Without Meanings: An Essay on the Content of Concepts*. MIT Press.

Franks, J. (2010). Cantor's Other Proofs That R Is Uncountable. *Mathematics Magazine*, *83*(4), 283–289.

Fresco, N. (2014). *Physical Computation and Cognitive Science*. Springer.

Funkenstein, A. (1986). *Theology and the Scientific Imagination From the Middle Ages to the Seventeenth Century*. Princeton University Press.

Gardner, H. (1987). *The Mind's New Science: A History of the Cognitive Revolution*. New York: Basic Books.

Gaukroger, S. (2002). *Descartes' System of Natural Philosophy*. Cambridge University Press.

Gaukroger, S. (2007). *The Emergence of a Scientific Culture: Science and the Shaping of Modernity 1210–1685*. Oxford University Press UK.

Gaukroger, S. (2010). *The Collapse of Mechanism and the Rise of Sensibility: Science and the Shaping of Modernity, 1680–1760*. Oxford University Press.

Gaukroger, S., Schuster, J., Sutton, J. (2000). *Descartes' Natural Philosophy*. London, New York: Routledge.

Gödel, K. (1953). *Kurt Gödel: Collected Works*. Oxford University Press.

Gödel, K. (1986). *Kurt Gödel: Collected Works: Volume III: Unpublished Essays and Lectures*. Oxford University Press.

Grodzinsky, Y., Santi, A. (2008). The Battle for Broca's Region. *Trends in Cognitive Sciences*, *12*(12), 474–480.

Gunderson, K. (1964). Descartes, La Mettrie, Language, and Machines. *Philosophy*, *39*(149), 193–222.

Hauser, M. D., Chomsky, N., Tecumseh Fitch, W. (2002). The Faculty of Language: What Is It, Who Has It, and How Did It Evolve? *Science*, *298*(5598), 1569–1579.

Lucas, J. R. (1961). Minds, Machines and Gödel. *Philosophy*, *36*(137), 112–127.

McLaughlin, P. (2000). Force, Determination and Impact. In J. Schuster, S. Gaukroger, J. Sutton (Eds.), *Descartes' Natural Philosophy* (pp. 81–112). London, New York: Routledge.

Miłkowski, M. (2013). *Explaining the Computational Mind*. MIT Press.

Papayannopoulos, P. (2020). Computing and Modelling: Analog vs. Analogue. *Studies in History and Philosophy of Science Part A*.

Penrose, R. (1989). *The Emperor's New Mind*. Oxford University Press.

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

Pullum, G. K., Scholz, B. C. (2010). Recursion and the Infinitude Claim. *Recursion in Human Language*, *104*, 113–38.

Putnam, H. (1960). Minds and Machines. In S. Hook (Ed.), *Dimensions of Minds* (pp. 138–164). New York University Press.

Putnam, H. (1963). Degree of Confirmation' and Inductive Logic. In P. A. Schilpp (Ed.), *The Philosophy of Rudolf Carnap* (pp. 761–783). La Salle: Open Court.

Quine, W. V. (1940). *Mathematical Logic*. Harvard University Press.

Sepper, D. L. 2000. Figuring Things Out: Figurate Problem-Solving in the Early Descartes. In J. Schuster, S. Gaukroger, J. Sutton (Eds.), *Descartes' Natural Philosophy* (pp. 228–248). London, New York: Routledge.

Shapiro, S. (1998). Incompleteness, Mechanism, and Optimism. *Bulletin of Symbolic Logic*, *4*(3), 273–302.

Sieg, W. (2001). Calculations by Man and Machine: Conceptual Analysis. *Reflections on the Foundations of Mathematics (Essays in Honor of Solomon Feferman)*, *15*, 387–406.

Sieg, W. (2013). Gödel's Philosophical Challenge (to Turing). In B. J. Copeland, C. J. Posy, O. Shagrir (Eds.), *Computability: Gödel, Church, Turing, and Beyond* (pp. 183–202). MIT Press.

Slezak, P. (1983). Descartes's Diagonal Deduction. *The British Journal for the Philosophy of Science*, *34*(1), 13–36.

Slezak, P. (1988). Was Descartes a Liar? Diagonal Doubt Defended. *The British Journal for the Philosophy of Science*, *39*(3), 379–388.

Smith, B. C. (2002). The Foundations of Computing. In M. Scheutz (Ed.), *Computationalism: New Directions* (pp. 23–58). MIT Press.

Sorensen, R. A. (1986). Was Descartes's Cogito a Diagonal Deduction? *The British Journal for the Philosophy of Science*, *37*(3), 346–351.

Svejdar, V. (2008). Relatives of Robinson Arithmetic. In M. Peliš (Ed.), *The Logica Yearbook* (pp. 253–263). London: College Publications.

Tarski, A. (1956). *Logic, Semantics, Metamathematics*. Oxford: Clarendon Press.

Woodward, J. F. (2011). Data and Phenomena: A Restatement and Defense. *Synthese*, *182*(1), 165–179.

STUDIA SEMIOTYCZNE (SEMIOTIC STUDIES), t. XXXIV, nr 1 (2020), s. 233–242
ISSN 0137-6608, e-ISSN 2544-073X
DOI: 10.26333/sts.xxxiv1.10
Article

V. ALEXIS PELUCE [*]

ON MARTIN-LÖF'S CONSTRUCTIVE OPTIMISM

SUMMARY: In his 1951 Gibbs Memorial Lecture, Kurt Gödel put forth his famous disjunction that either the power of the mind outstrips that of any machine or there are absolutely unsolvable problems. The view that there are no absolutely unsolvable problems is optimism, the view that there are such problems is pessimism. In his 1995—and, revised in 2013—*Verificationism Then and Now*, Per Martin-Löf presents an illustrative argument for a constructivist form of optimism. In response to that argument, Solomon Feferman points out that Martin-Löf's reasoning relies upon constructive understandings of key philosophical notions. In the vein of Feferman's analysis, one might be object to Martin-Löf's argument for either its reliance upon constructivist (as opposed to classical) considerations, or for its appeal to non-unproblematically mathematical premises. We argue that both of these responses fall short. On one hand, to be critical of Martin-Löf's reasoning for its constructiveness is to reject what would otherwise be a scientific advance on the basis of the assumption of constructivism's falsehood or implausibility, which is of course uncharitable at best. On the other hand, to object to the argument for its use of non-unproblematically mathematical premises is to assume that there is some philosophically neutral mathematics, which is implausible. Martin-Löf's argument relies upon his third law, the claim that from the impossibility of a proof of a proposition we can construct a proof of its negation. We close with a discussion of some ways in which this claim can be criticized from the constructive point of view. Specifically, we contend that Martin-Löf's third law is incompatible with what has been called "Poincaré's Principle of Epistemic Conservation", the thesis that genuine increase in mathematical knowledge requires subject-specific insight.[1]

KEYWORDS: optimism, pessimism, Martin-Löf, Gödel's disjunction

[*] City University of New York, Department of Philosophy, The Graduate Center. E-mail: vpeluce@gradcenter.cuny.edu. ORCID: 0000-0002-7440-3641.

[1] The author is grateful for the helpful comments of Alfredo Roque Freire, Alberto Naibo, Romina Padro, David Santamaría, Sandra Shapshay, Yale Weiss, and an anonymous reviewer. The author also thanks Sergei Artemov for his discussion and supervision.

Optimism and Pessimism

In his 1951 Gibbs Memorial Lecture *Some Basic Theorems on the Foundations of Mathematics and Their Implications*, Kurt Gödel argued for his famous disjunction: "Either […] the human mind (even within the realm of pure mathematics) infinitely surpasses the powers of any finite machine, or else there exist absolutely unsolvable diophantine problems of the type specified" (Gödel, 1951, p. 310).

While both disjuncts have received much attention (see, for example, Lucas, 1961; Penrose, 1989; 1994; Horsten & Welch, 2016), the former has been the primary focus of philosophical discussion surrounding Gödel's Disjunction. In this article, we focus on the latter. The view that there are no absolutely unsolvable diophantine problems is optimism. The alternative thesis is pessimism, that there exist absolutely unsolvable diophantine problems.

What does Gödel have in mind by an absolutely unsolvable diophantine problem? A diophantine equation is one in which only integer solutions and coefficients are used. A diophantine problem is the question of whether a given diophantine equation has a solution. Notably, one can verify if a given assignment to the variables is a solution in a finite number of steps. The question of whether or not a given equation has a solution we then understand as the disjunction of the claim that there is a solution to that equation and its negation. This, in turn, is the traditional way of understanding the law of excluded middle as the formalization of optimism, which we see in L.E.J. Brouwer, for example (Brouwer, 1908, p. 109).

The question of whether or not there exist absolutely unsolvable problems has its proximal roots in 19th century philosophy of science. Emil du Bois-Reymond famously closed his 1872 Berlin address, as translated by Andrea Reichenberger: "As regards the riddle of the nature of matter and force and how they are able to think, we must resign ourselves once and for all to the far more difficult verdict: 'Ignorabimus'" (Reichenberger, 2019, p. 53).

du Bois-Reymond's view here is that there are portions of reality that are inaccessible to us. While these are not the sort of mathematical problems we have been discussing, the view here is that there are some metaphorical "corners" that we will never be able to see around.

Opposite the pessimism of du Bois-Reymond we find the optimism of David Hilbert. In his 1900 address to the International Congress of Mathematicians in Paris, he claimed:

> This conviction of the solvability of every mathematical problem is a powerful incentive to the worker. We hear within us the perpetual call: There is the problem. Seek its solution. You can find it by pure reason, for in mathematics there is no 'ignorabimus'. (1902, p. 445)

On this view, the mathematical "worker" is motivated by the knowledge that their task is not some thankless Sisyphean task but rather one they can actually complete.

Martin-Löf’s Constructive Optimism

In his 1995 and, revised in 2013, *Verificationism Then and Now*, Per Martin-Löf presents a case in favor of optimism. Making use of several laws for which he provides philosophical justification, he argues:

> [T]here are no absolutely undecidable propositions. And why does this follow from [the third law, the claim that if a proposition cannot be known to be true then it can be known to be false]? Well, suppose that we had a proposition which could neither be known to be true nor be known to be false. Then, in particular, it cannot be known to be true, so, by the third law, it can instead be known to be false. But that contradicts the assumption that the proposition could not be known to be false either. (2013, pp. 12–13)

Imagine that a given proposition is absolutely undecidable, which is just to say that the associated problem is unsolvable in the sense we used above. In terms of knowledge, given that it is in fact absolutely unsolvable, this means that it cannot be known to be true and it cannot be known to be false. But, if a proposition cannot be known to be true, then, Martin-Löf argues, it can be known to be false. This is in virtue of his third law. The thought is that if it is impossible that *a* is a proof of *A* for any *a*, then we can conclude a refutation of *A*. But, if we have a refutation of the proposition in question, then the problem is not absolutely unsolvable, which contradicts our original assumption. Therefore, there are no absolutely undecidable propositions. Call the above articulation of optimism constructive optimism.

There is a clear step worth examining in more detail, that from the impossibility of knowing the truth of the proposition we can move to the possibility of knowledge of its falsehood. This, however, will be the focus of the second half of this paper. Let us first turn to a different sort of objection to Martin-Löf’s argument. Solomon Feferman, in “Are there Absolutely Unsolvable Problems? Gödel’s Dichotomy”, comments:

> Indeed, Per Martin-Löf has proved exactly that, in the form: There are no propositions which can neither be known to be true nor known to be false [...]. However, this is established on the basis of the constructive explanation of the notions of “proposition”, “true”, “false”, and “can be known”. (2006, p. 147)

Feferman continues:

> For the non-constructive mathematician, Martin-Löf’s result would be translated roughly as: “No propositions can be produced of which it can be shown that they can neither be proved constructively nor disproved constructively”. For the non-constructivist this would seem to leave open the possibility that there are absolutely unsolvable problems *A* “out there”, but we cannot produce ones of which we can show that they are unsolvable. (2006, p. 147)

Feferman's point here is that while Martin-Löf's argument succeeds at establishing optimism for the constructivist, it falls short of establishing optimism *tout court*. He goes on to present examples of problems that are "absolutely unsolvable from the standpoint of practice" (Feferman, 2006, p. 149).

Feferman argues that the non-constructive mathematician can evade Martin-Löf's target conclusion of optimism by reinterpreting it in a way that fits within a non-constructive worldview. If pessimism or optimism is to be established *tout court*, the reasoning would go, it must be done so independent of a constructive philosophy of mathematics. This can be interpreted in two ways, however. The first emphasizes the constructivist portion of Martin-Löf's reasoning. The second emphasizes the philosophical, where this is understood as something non-mathematically neutral, content of Martin-Löf's argument. For the remainder of this section, we discuss the first interpretation. The second interpretation is the focus of the following section.

The first interpretation emphasizes that Martin-Löf employs constructive understandings of key notions, and that these admit of non-constructive interpretations. We point out only that just because an unorthodox thesis can be given an interpretation that coheres with the orthodoxy does not mean that it should. Even the strictest Quinian should admit that in some cases genuine development first appears as unorthodox. Moreover, it is clear that in some situations translation from the unorthodox is responsible for the loss of relevant content. This is arguably what happens in the non-constructive interpretation of constructivism. In interpreting intuitionistic logic in **S4**, we substitute a constructive understanding of truth for the notion of a proof of something that holds classically. While doing so provides a way of explaining intuitionism to the classical modal logician, it does so at the expense of what is arguably the most foundational notion within intuitionism. In this case also, the practice of recasting constructive contributions as mere features within a broader classical panorama threatens to make unavailable what might otherwise be seen as a genuine scientific advance in Martin-Löf's constructive optimism.

The Axiomatic Method

Based on our discussion of Martin-Löf's argument for constructive optimism and Feferman's response, the question arises: can optimism or pessimism be established on purely mathematical grounds? That is, can we decide this question in some manner that is not seized upon by philosophy?

But what would it be to have such an understanding of mathematics? Perhaps examining a distinction on uses of axiomatics in Gödel can help us to get clear on whether or not it is plausible that there is a way to thus separate off the philosophy. Also in his 1951 lecture, Gödel distinguishes between the proper and merely hypothetico-duductive uses of axiomatics. He claims:

> [The inexhaustibility of mathematics] is encountered in its simplest form when the axiomatic method is applied, not to some hypothetico-deductive system such as geometry (where the mathematician can assert only the conditional truth of the theorems), but to mathematics proper, that is, to the body of those mathematical propositions which hold in an absolute sense, without any further hypothesis […].
>
> Of course, the task of axiomatizing mathematics proper differs from the usual conception of axiomatics insofar as the axioms are not arbitrary, but must be correct mathematical propositions, and moreover, evident without proof. (1995, p. 305)

As Gödel emphasizes to his audience, of course there are "widely divergent" ways of saying just what counts as mathematics proper. One suggestion might be to use this distinction to try to find a sense of mathematics not seized upon by philosophy.

What happens when we consider proper mathematics axiomatically? To do so is to limit the application of the axiomatic method to a specific domain, setting aside the specific sort of Platonist view that any consistent system has application, as in (Balaguer, 1998). We see this, for example, in the emphasis on contentual reasoning in Sergei Artemov's *Provability of Consistency* (2019), where considerations of meaning filter out gerrymandered uses of formalism. This, of course, is motivated by a philosophy of mathematics and clearly is not philosophically neutral in a general sense.

Let us instead consider the hypothetico-deductive use of axiomatics. By this, after all, Gödel meant reasoning conditionally with axioms irrespective of how they relate to mathematical reality, however that is explained. If the philosophically neutral way of understanding mathematics corresponds to Gödel's hypothetico-deductive use of axiomatics, then this suggests an account of absolute provability as provable in a given hypothetico-deductive system.

There are two objections to this proposal. The first is that it is easy to see that absolute provability as understood in this way would trivialize the notion. In his *Inexhaustibility: A Non-Exhaustive Treatment* (2004), Torkel Franzén makes exactly this point:

> That a formalization of a mathematical statement is provable in a formal theory does not itself imply that the statement can be proved in the ordinary mathematical sense, that is, that an argument establishing the statement as a mathematical theorem can be given. As an extreme instance, any statement is provable in a theory in which it is taken as an axiom, but this tells us nothing about whether or not the statement can be proved in the ordinary sense. (p. 8)

Anything that can be taken as an axiom is provable in some hypothetico-deductive axiomatic system. Hence if we take absolute provability to be provability in some hypothetico-deductive system, then this trivializes the notion of absolute provability. Second, such a thesis on absolute provability would not give us an account of what counts as mathematical in a way that is not encroached upon by philosophy. To think that what is expressible in hypothetico-deductive formal systems just is this mathematical core itself is of course not philosophically neutral by any means.

There are ways of amending the proposal that absolute provability just is hypothetic-deductive provability. Perhaps we want to consider provability to the standards of a mathematical community (Franzén, 2004), demand that the axioms be objects of knowledge or possible objects of knowledge given certain stipulations (Williamson, 2016, pp. 247–248), that axioms be "deemed plausible" (Clarke-Doane, 2013, p. 469). In any such case though it is clear that insofar as we are appeal to a ground for a given circumscription of proper mathematics that we appeal to philosophical considerations.

While we do not claim that using Gödel's discussion of the axiomatic method is the only way of attempting to find a notion of mathematics not encroached upon by philosophy, the prospects for something else in this vein look dim. The criticism that Martin-Löf's argument is objectionable insofar as it makes use of philosophical notions can seemingly be leveled thus against any account in the literature. For this reason, it seems like it would be misguided. A paragraph from Saul Kripke's *Is there a Problem about Substitutional Quantification?* (1976) makes such a point, though in a different context. He writes:

> Logical investigations can obviously be a useful tool for philosophy. They must, however, be informed by a sensitivity to the philosophical significance of the formalism and by a generous admixture of common sense, as well as a thorough understanding both of the basic concepts and of the technical details of the formal material used. It should not be supposed that the formalism can grind out philosophical results in a manner beyond the capacity of ordinary philosophical reasoning. There is no mathematical substitute for philosophy. (1976, p. 416)

An answer to the question of pessimism or optimism does not seem to be the sort of thing that can be achieved by recourse to mathematics that is not seized upon in some sense by philosophy. Instead, the answer to this question is a consequence of an account of absolute provability, which is itself a thoroughly philosophical undertaking. Martin-Löf's reasoning cannot be faulted for essentially attempting to do just this.

The Third Law

We suggested that two sorts of responses to Martin-Löf's argument were unsatisfying. The first was that his argument relied upon constructive notions, which admit of a non-constructive interpretation. This, we argued, is to overlook the revolutionary character of constructivism, and in this way was less of an objection than a dismissal. The second was that Martin-Löf's argument relied upon non-mathematical groundwork. We argued, however, that we should be skeptical about the possibility of finding an account of mathematics that is not somehow permeated by philosophy. Setting the above discussion aside, when we introduced Martin-Löf's argument we did flag a premise for later discussion.

This was Martin-Löf's third law, that from the fact that it was not possible to know a given proposition, we can conclude positive evidence for the negation of that proposition.

How is this justified? Take a given proposition *A*. By *a:A*, we mean that *a* is a proof of *A*. By the claim that *A* cannot be known, Martin-Löf understands that "the situation *a:A* cannot arise, for any *a*" (Martin-Löf, 2013, pp. 11, 13). He continues:

> Now, from this negative piece of information, I have to get something positive, namely, I have to construct a refutation, and a refutation of *A* is a hypothetical proof of [*falsum*] from *A*, or, equivalently, a function which takes a proof of *A* into a proof of [*falsum*]. The argument is this: we simply introduce a hypothetical proof of [*falsum*] from *A*, call it *R*. (2013, p. 13)

The thought is that we have negative information that it is impossible that for any *a*, it holds that *a:A*. We get the positive refutation of *A* by constructing a hypothetical proof of *falsum* from *A*.

But what does it mean to say that it is impossible to know that *A*, or alternatively, that the situation *a:A* cannot occur for any *a*? Martin-Löf is clear about what he means by possibility. He writes: "[By] the notion of possibility, I have nothing more to say, except that it is the notion of logical possibility, or possibility in principle, as opposed to real or practical possibility, which takes resources and so on into account" (2013, p. 9). Since Martin-Löf is here discussing applications of introduction and elimination rules, it seems clear that his "possibility in principle" or "logical possibility" will have to do with what can be reached by transformations of this sort.

A first objection is that this view assumes that the rules articulated by some specific system express what it really is for something to be possible in principle. After all, if our set of rules is somehow suspect, it would be strange to assume they were even in the position to lead us securely to a mathematical insight. But perhaps we can make this point sharper. In his *Science and Method* (2012), Henri Poincaré discusses the dynamic nature of the concept of solution:

> Many times already men have thought they had solved all the problems, or at least that they had made an inventory of all that admit of solution. And then the meaning of the word solution has been extended; the insoluble problems have become the most interesting of all, and other problems hitherto undreamt of have presented themselves. (2012, p. 370)

The thought here is that the horizon of what seems possible at a given period is consistently surpassed, and that it is in these instances that we see genuinely interesting development. After this, we revise what we thought was possible.

There is perhaps a stronger objection, though, to Martin-Löf's third law. We can concede that the specific laws chosen of the system in question actually do characterize logical possibility in the relevant sense. Nonetheless, the third law involves the passage from a logical insight to a mathematical one insofar as we go from a fact about logical impossibility to a positive refutation of a claim. There is something non-constructive about this. To get clear on this, we look again to the thought of Poincaré. Michael Detlefsen dubs "Poincaré's Principle of Epistemic Conservation" the thesis that "there can be no increase in genuine knowledge of a specific mathematical subject without an underlying increase in subject-specific insight into (i.e. intuitional grasp of) that subject" (Detlefsen, 1990, pp. 501–502). While logical reasoning is characteristically general, mathematical understanding arguably involves subject-specific insight. But Martin-Löf's third law, insofar as it relies on a notion of logical impossibility, takes us from a general claim about what can be done, in this case with the application of introduction and elimination rules, to the existence of a positive mathematical insight. While perhaps in some cases the realization that a proof of some proposition is logically impossible will lead to a specific proof of the refutation of that claim (consider the case in which the proposition in question is one about the capabilities of the rules that characterize this notion of possibility), to assume that this holds generally is far stronger. Indeed, that there could be a general recipe for getting mathematical insights from logical ones is exactly the sort of thing that contradicts Poincaré's Principle.

Perhaps we can fix a notion of possibility in a different way. For example, we might consider what an actual agent can do or what an idealized agent can do. For the remainder of this section, we argue that in both cases we need not endorse the third law. The first sort of case suggests considering empirical agents who fall short based on lack of resources or similar circumstances. While this is not the sort of agent Martin-Löf has in mind when discussing possibility, it is worth examining nonetheless. In terms of possible worlds, this is discussed in the context of the condition that if A holds in all worlds accessible from a given world in a Kripke model, then at that world A is known, in Sergei Artemov's *Knowing the Model* (2016, p. 4). Let A be some arithmetical truth unknown to a particular individual. Assume that they have seen no proofs of A. It would seem strange for that individual to conclude the negation of A from the agent's limited information. Alternatively, consider a B that is refutable. The agent, of course, has seen no proof of B. Even if they correctly conclude the negation of B, their reasoning is too hasty; something is missed when they move from a claim about their own abilities to an actual refutation of B. It thus seems that if one wants to understand the relevant notion of possibility invoked in the third law as "possibility for an empirical agent", they need not endorse the third law.

The above has to do with the sorts of mistakes empirical agents are prone to make. What if we were to idealize away from these concerns? Even if we consider the subject divorced from such empirical limitations, it would not follow that the agent would be fully aware of their own capabilities in the sense required for a version of the third law. For example, in his *On the Fourfold Root of the Principle of Sufficient Reason* (1997), Arthur Schopenhauer argues:

> [T]he subject knows itself only as a willer, not as a knower. For the ego that represents thus the subject of knowing, can itself never become representation or object, since, as the necessary correlative of all representations, it is their condition. (1997, p. 208)

Schopenhauer's subject is clearly not a limited empirical subject. Nonetheless, the subject inasmuch as it is a subject cannot know itself, because to know itself would be to treat itself as an object qua object of knowledge. The subject is explicitly ignorant of itself as a knower and must be so for the above reason. Again, it would seem strange from the subject to conclude from the impossibility of knowing that it is a knower to the conclusion that it is not a knower. Here we see then that even when a subject is considered in a way that has been idealized away from empirical limitations, the third law need not be accepted.

Conclusions

In this paper we discussed Martin-Löf's argument for constructive optimism. Therein, he argues for a form of optimism based a view of absolute provability as knowability. We presented criticisms of some objections to Martin-Löf's argument. Then, we put forth a novel criticism of Martin-Löf's argument based on his third law. His third law is a general rule describing the passage from a point about the impossibility of a proof, where this is understood in terms of logical possibility, to the positive existence of a refutation. This, we argued, ran afoul of Poincaré's Principle. A possible emendation of the third law was to interpret the notion of possibility therein in a way that does not appeal to logical possibility. An intuitive thought, especially in the constructivist context, is to think of possibility in terms of the abilities of agents. But even in this case, we continued, the third law need not hold. While the reasoning against the third law was straightforward when the agent under consideration has the limitations of an empirical agent, the Schopenhauerian subject provided an example of a case in which there are important reasons that even an idealized subject's abilities might not be completely transparent to them. In both instances, we observe failure of the third law as a general principle of reasoning.

REFERENCES

Artemov, S. (2016). Knowing the Model. Retrieved from: https://arxiv.org/pdf/1610.04955.pdf

Artemov, S. (2019). The Provability of Consistency. Retrieved from: https://arxiv.org/pdf/1902.07404.pdf

Balaguer, M. (1998). *Platonism and Anti-Platonism in Mathematics*. Oxford: Oxford University Press.

Clarke-Doane, J. (2013). What is Absolute Undecidability? *Noûs*, *7*(3), 467–481.

Detlefsen, M. (1990). Brouwerian Intuitionism. *Mind*, *99*(369), 501–534.

Feferman, S. (2006). Are There Absolutely Unsolvable Problems? Gödel's Dichotomy. *Philosophia Mathematica*, *14*(2), 134–152.

Franzén, T. (2017). *Inexhaustibility: A Non-Exhaustive Treatment*. Cambridge: Cambridge University Press.

Gödel, K. (1995). Some Basic Theorems on the Foundations of Mathematics and Their Implications. In S. Feferman et al. (eds.), *Collected Works: Volume III: Unpublished Essays and Lectures* (pp. 304–323). Oxford University Press.

Hilbert, D. (1902). Mathematical Problems: Lecture Delivered Before the International Congress of Mathematicians at Paris in 1900. *Bulletin of the American Mathematical Society*, *8*, 437–479.

Horsten, L. & Welch, P. (2016). *Gödel's Disjunction: The Scope and Limits of Mathematical Knowledge*. Oxford: Oxford University Press.

Kripke, S. A. (1976). Is There a Problem About Substitutional Quantification? In J. McDowell, G. Evans (Eds), *Truth and Meaning: Essays in Semantics* (pp. 324–419). Oxford: Oxford University Press.

Lucas, J. R. (1961). Minds, Machines and Gödel. *Philosophy*, *36*(137), 112–127.

Martin-Löf, P. (2013). Verificationism Then and Now. In M. van der Schaar (Ed.), *Judgement and the Epistemic Foundation of Logic* (pp. 3–14). New York: Springer.

Penrose, R. (1989). *The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics*. Oxford: Oxford University Press.

Penrose, R. (1994). *Shadows of the Mind: A Search for the Missing Science of Consciousness*. Oxford: Oxford University Press.

Poincaré, H. (2012). *The Value of Science: Essential Writings of Henri Poincaré*. New York: Modern Library.

Reichenberger, A. (2019). From Solvability to Formal Decidability: Revisiting Hilbert's "Non-Ignorabimus". *Journal of Humanistic Mathematics*, *9*(1), 49–80.

Schopenhauer, A. (1997). *On the Fourfold Root of the Principle of Sufficient Reason*. Chicago: Open Court.

Shapiro, S. (1998). Incompleteness, Mechanism, and Optimism. *Bulletin of Symbolic Logic*, *4*(3), 273–302.

Williamson, T. (2016). Absolute Provability and Safe Knowledge of Axioms. In L. Horsten, P. Welch (Eds.), *Gödel's Disjunction: The Scope and Limits of Mathematical Knowledge* (pp. 243–253). Oxford: Oxford University Press.

STUDIA SEMIOTYCZNE (SEMIOTIC STUDIES), t. XXXIV, nr 1 (2020), s. 243–266
ISSN 0137-6608, e-ISSN 2544-073X
DOI: 10.26333/sts.xxxiv1.11
Article

PAULA QUINON [*]

THE ANTI-MECHANIST ARGUMENT BASED ON GÖDEL'S INCOMPLETENESS THEOREMS, INDESCRIBABILITY OF THE CONCEPT OF NATURAL NUMBER AND DEVIANT ENCODINGS

SUMMARY: This paper reassesses the criticism of the Lucas-Penrose anti-mechanist argument, based on Gödel's incompleteness theorems, as formulated by Krajewski (2020): this argument only works with the additional extra-formal assumption that "the human mind is consistent". Krajewski argues that this assumption cannot be formalized, and therefore that the anti-mechanist argument—which requires the formalization of the whole reasoning process—fails to establish that the human mind is not mechanistic. A similar situation occurs with a corollary to the argument, that the human mind allegedly outperforms machines, because although there is no exhaustive formal definition of natural numbers, mathematicians can successfully work with natural numbers. Again, the corollary requires an extra-formal assumption: "**PA** is complete" or "the set of all natural numbers exists". I agree that extra-formal assumptions are necessary in order to validate the anti-mechanist argument and its corollary, and that those assumptions are problematic. However, I argue that formalization is possible and the problem is instead the circularity of reasoning that they cause. The human mind does not prove its own consistency, and outperforms the machine, simply by making the assumption "I am consistent". Starting from the analysis of circularity, I propose a way of thinking about the interplay between informal and formal in mathematics.

KEYWORDS: the Lucas-Penrose argument, the Church-Turing thesis, Carnapian explications, natural numbers, computation, conceptual engineering, conceptual fixed points, conceptual vicious circles, deviant encodings, structuralism.

[*] Warsaw University of Technology, Faculty of Administration and Social Sciences. International Center for Formal Ontology. E-mail: paula.quinon@pw.edu.pl. ORCID: 0000-0001-7574-6227.

1. Introduction

The Lucas-Penrose anti-mechanist argument against computability of the human mind in a nutshell states the following. According to Gödel's incompleteness theorems, a (sufficiently rich) consistent theory that can prove its own consistency does not exist. However, mathematical practice shows that Gödel-type results are commonly proven by human mathematicians. In consequence, says the argument, human mathematicians are not describable as formal proof systems, nor are they reducible to performing algorithms.

In (2020), Krajewski criticises the Lucas-Penrose argument by claiming that Gödel's incompleteness theorems standing alone (as it is in the Lucas-Penrose case) are not sufficient for formulating the claim that the human mind is non-computational. The anti-mechanist argument based on Gödel's incompleteness theorems needs to be enriched by an extra-formal assumption. For instance, an assumption that the theory constituting the human mind is consistent.

In order to provide an additional context to his investigations, Krajewski (2020), highlights the analogy between the claim that Gödel's incompleteness theorems imply the non-computational nature of the human mind, and the claim that "we [humans] cannot give a definition of the natural numbers as we understand them" (p. 49). The analogy goes as follows: in order to make a successful anti-mechanist argument based on Gödel's incompleteness theorems, one needs to assume—in addition to the formal counterpart—that the theory constituting the human mind is consistent. The fact that Gödel's argument can be iterated for increasingly rich theories is not sufficient for formulation of the anti-mechanist argument. The possibility to iterate increasingly rich theories, which all have a Gödel's sentence, and none of which proves its own consistency, is a formal process and as such can be executed by purely formal means. Thus, it does not say anything about computability or non-computability of the human mind. In order to be able to formulate the anti-mechanist argument, one needs to assume—for instance—that the human mind is consistent. Analogously, each definition of a natural number ends up in a vicious circle of definitions, or—as Krajewski says

> [O]ur axioms [both the first-order (**PA1**) and the second-order Peano Arithmetic (**PA2**)] define numbers only when taken together with some background knowledge or apparatus that makes possible our intuitive grasp of numbers [such as the intuition that the first-order Peano's Arithmetic is complete or the intuition that there exists the set of all natural numbers being referred to in the background of the second-order Peano's Arithmetic]. (2020, p. 49)

In both cases, an immediate, but incorrect according to Krajewski, conclusion could be that "no computer can be taught our concept of a number" and that in consequence "we [humans] are better than any machine" (2020, p. 49).

In this paper, I observe that this analogy can be pushed further to a circular reasoning. In both cases, making an extra-formal assumption leads to a vicious circle because one assumes consistency of one's mind while proving that the human mind outperforms machines, or one assumes that the concept of a set of natural numbers can be intuitively apprehended while defining natural numbers. Studies show that the method of conceptual analysis is particularly sensitive to falling into circular reasoning. The circularity related to the concept of natural number has been investigated in discussions about computational structuralism (Halbach & Horsten, 2005; Quinon & Zdanowski, 2007). Computational structuralism is a position, according to which the concept of natural number and the concept of computation are closely related. More precisely, according to this position, an adequate account of natural numbers treats them as objects that can be used for computations. After a brief overview of the anti-mechanist argument and its criticism in **Section 1**, in **Section 2** I will explain inter-relation and inter-definability between the concept of natural number and the concept of computation. In **Section 3**, I describe how the two concepts fall into a vicious circle of definition individually, and also while used in definition of one another.

Rescorla (2007) identifies problems with conceptual analysis related to the concept of computation, Quinon (2018) suggests that there is no fully satisfactory way out from vicious circles in definitions within conceptual analysis. Approaching the concept of computation and the concept of natural number from another methodological perspective, seems to be more fruitful. For instance, an interesting insight can be gained thanks to conceptual engineering. Both concepts have a form of what in the area of conceptual engineering is called "conceptual fixed point". A conceptual fixed point is an idea issued from the conceptual engineering of moral concepts, where it is claimed that some basic moral concepts should not be engineered, but should always be understood in the most objective way (Eklund, 2015). **Section 4** is devoted to the presentation of the method of conceptual engineering and the adequacy of conceptual fixed points for the concept of computation and the concept of natural number. As suggested by the phenomenon of conceptual fixed points, the only way out from these vicious circles consists in an arbitrary decision which is the intended meaning of the given concept.

In **Section 5**, I extend my methodological investigations into yet another method, and I discuss the advantages of thinking about formalisation of the concept of computation in terms of Carnapian explications. It has been argued, for instance in (Quinon, 2019), that a move from an intuitive concept of computation, used in everyday life, to a scientific or formal concept as stated by the Church-Turing thesis, follows the schema of a Carnapian explication. In **Section 6**, I extend the context of Carnapian explications of the temporary aspect. I realise that both, the concept of natural number and the concept of computation, have been evolving in such a way, that their core meanings were shifting. I propose a hypothesis that at least a part of the confusion regarding the specificity of the conceptual structure of the concept of computation contributes to the confusion regarding the nature of human reasoning and the human mind. In consequence, In consequence, I claim that—at least partially—the "feeling" that there are non-computational processes is due to the complexity of the conceptual structure of the concept of computation.

In the final **Section 7**, I wrap up with the ways in which my observations regarding the concept of computation and the concept of natural number, could be used for understanding the reasons for which the anti-mechanist argument fails. I suggest a different reason from the one proposed by Krajewski, for which the extra-formal assumption prevents the anti-mechanist argument from success. Firstly, I claim that thanks to the method of Carnapian explications, it is highly possible to go from intuitive pre-scientific concept to a formal concept. Secondly, I observe that the extra-formal assumption after an arbitrary formalisation, leads to the vicious circle in reasoning. Therein lies the problem.

2. The Lucas-Penrose Argument and Its Criticism

In this section, I present a brief overview of various versions of the anti-mechanist argument based on Gödel’s incompleteness theorems, and the ways in which those arguments have been criticised. In particular, I explicate Krajewski’s way of refuting the argument. In my overview, I prioritise the authors to who Krajewski refers to in his paper.

The first of Gödel’s incompleteness theorem says that in every sufficiently rich[1] consistent first-order theory[2] there exist statements that are true[3], but that cannot be proven within this theory. The second of Gödel’s incompleteness theorem says that every sufficiently rich consistent first-order theory cannot prove its own consistency.

According to the anti-mechanist argument based on Gödel's incompleteness theorems, since human mathematicians can fruitfully work with Gödel's incompleteness theorems, that means those mathematicians use the resources from the outside of the theory (e.g., they are able to refer to the intended model of arithmetic or recognize that the human mind is consistent). Thus, human mathematicians outperform machines, because—unlike machines—they are able to include in their reasoning such external resources.

The intuition that humans could prove theorems which machines could not has already been present in (Turing, 1950)[4] and in (Post, 1941).[5] One of the most famous voices exploring the anti-mechanist argument based on Gödel’s incompleteness theorems against the computational theory of mind—next to Hofstadter

[1] By “sufficiently rich” one means that the formal system is able to express arithmetic of addition and multiplication.

[2] A formal system, or a theory, is a collection of axioms together with rules of inference. The importance of using first-order logic is because of the completeness of this logic.

[3] A statement is true, when it is satisfied in the intended model of the theory.

[4] As reported by Krajewski, Turing believed that even if a machine cannot prove as much as humans can, it is still worth constructing robots.

[5] As reported by Krajewski, Post believed that man cannot construct a machine which can do all the things he can.

(1979), Nagel and Newman (1958; 1961)—is Lucas (1961; also 1968; 1996), who presented a "mathematical proof" of man's superiority over a machine. Lucas extended the applicability of Gödel's incompleteness theorems from formal systems to human subjects. In his view, humans are subjects to the same formal limits as machines. However, as Lucas observes, human mathematicians can prove Gödel's incompleteness theorem, which means, human mathematicians use extra-formal resources that enable them to perform such proofs.

Lucas' argument relies on the fact that Gödel's theorem(s) is formulated in purely formal terms. As Lucas observes himself, this is what differentiates Gödel's results from the liar paradox. The liar paradox, which states that "This statement is untrue", is "viciously self-referential, and we do not know what the statement is, which is alleged to be untrue, until it has been made, and we cannot make it until we know what it is that is being alleged to be false" (Lucas, 1990, p. 2). Unlike the liar paradox, Gödel's theorem is formulated within a full-blooded system where it is clearly defined, which sentences are true and what does it mean to be provable. Lucas' claims that the fact that a (idealised) human mind, even if it cannot prove Gödel's theorem(s) for the given theory, can—thanks to its additional non-mechanical skills—recognize this theorem as true in its system. In consequence, a human mind outperforms a machine.

Penrose in (1989; 1994) extended Lucas' reasoning of a positive claim regarding the extra-formal resources available to humans that enable them to construct reasonings unavailable to machines. Penrose suggested that in the brain the physical basis of non-computable behavior exists, and he indicated quantum mechanics as a credible candidate. According to him quantum processes might explain not only reasoning of human mathematicians, but also consciousness.

A constructive criticism of the Lucas-Penrose style argument was formulated by Putnam (1960), Benacerraf (1967), Wang (1974), then later also by Boolos (1995) and Shapiro (1998). Penrose's version got criticised in particular by Feferman (1995), Putnam (1995) and Shapiro (2003). Krajewski claims that the ways of criticizing the Lucas-Penrose argument follow one of the two main lines (2020, pp. 5–6):

- The mind is a machine and it is consistent, but it cannot prove Gödel's sentence by itself.[6]
- The mind is a machine, but it is inconsistent, and Gödelian limitations do not apply to it.

[6] This line of argument has already come from Gödel, who distinguished *subjective arithmetic* that humans can do, and who believed that in *objective mathematics* full arithmetic is a consistent theory. He also believed that the concept of computation can be defined without referring to any domain of computation; these claims amount to Gödelian platonism (Gödel, *1951).

Krajewski (2020) refutes the Lucas-Putnam argument in yet another way: he observes that iterations of increasingly strong theories proving the corresponding Gödel's sentences can be processed in a purely mechanical or computational manner available to both, humans and machines. In consequence, Krajewski claims that anti-mechanist is not implied by Gödel's incompleteness theorems alone. In addition, claims Krajewski, one needs to assume that humans have a privileged access to assessing consistency of the human mind. Krajewski claims that the argument fails because of the necessity of making this extra-formal assumption. This is so, because there is no formal way to account for the formal counterpart of assumptions.

Before I come back, in the last section, to Krajewski's rejection of the anti-mechanist argument, and my proposal of how to shift the way of thinking about the reasons for this rejection, I will now focus on the part which is particularly interesting for me, that is the meta-theoretical corollary to the anti-mechanist argument stating that humans cannot fully describe the concept of natural number.

3. The Concept of Natural Number and the Concept of Computation

I initiate my investigation into the nature of the extra-formal elements of the reasoning that enable the conclusion that the human mind is not computable, by discussing the corollary relating human inability to define the concept of natural number. Additionally, I extend the corollary of the claim that humans—for similar reasons—cannot define the concept of computation. Finally, I present the view according to which the concept of natural number and the concept of computation are closely related.

The fact that every formal definition of the concept of natural number leads to a necessary assumption from the outside of the formal system has been studied in the context of the view in philosophy of mathematics, called structuralism. According to structuralism, mathematics is the "science of structures", and while defining mathematical objects, one should first target their structural properties. For instance, while defining natural numbers, one should define the structure of natural numbers through relations they hold to each other, and not focus on individual properties of those elements.

Traditionally, structuralism defined natural numbers using second-order Peano Arithmetic (**PA2**). **PA2** is categorical and the class of (isomorphic) models in which it is satisfied is identified with natural numbers. The usual way of criticising the use of second-order Peano Arithmetic to define natural numbers consists in saying that the underlying logic is "set theory in sheep's clothing" (Quine, 1970, p. 66). Second-order logic has the ability, for instance, to express the information that two sets have the same cardinality. The concept of set is itself most frequently (implicitly) defined with a first-order axiomatic theory, such as *ZF*, that in turn, is a subject of non-standard interpretations, the Löwenheim-Skolem theorem, etc., which makes its intended model "hidden" within a continuum of other non-intended models. Therefore, in order to define the concept of natural number with **PA2**, humans have two choices. They can get involved in a vicious circle of definitions, or an infinite regression of theorems, or they can use extra-formal resources and admit in an arbitrary manner that there is such a thing as an intended (or a standard) model of set theory where the intended model of arithmetic exists.

Another, less known, version of structuralism, so called *computational structuralism*, proposes distinguishing the standard model of arithmetic from the continuum of non-standard models with the resources of **PA1** only (Halbach & Horsten, 2005; Quinon & Zdanowski, 2007). In order to do that, defenders of computational structuralism suggest adding a meta-mathematical constraint regarding the computability of interpretation of functional symbols in the language, and then use Tennenbaum's theorem in order to single out the standard model of arithmetic.

Theorem 2.1 (Tennenbaum, 1959) *Let* $\mathcal{M} = \langle \mathbb{M}, +, \times, 0, 1, < \rangle$ *be an enumerable model of* **PA1***, and not isomorphic with the standard model* $\mathcal{N} = \langle \mathbb{N}, +, \times, 0, 1, < \rangle$*. Then* $\mathcal{M}$ *is not recursive.*

More explicitly why Tennenbaum's theorem is relevant for the structuralist way of thinking is visible in the transposition of the theorem:

Theorem 2.2 (Tennenbaum transposition) *Let* $\mathcal{M}$ *be an enumerable model of first-order Peano arithmetic. If the interpretation of addition and multiplication within* $\mathcal{M}$ *are computable then* $\mathcal{M}$ *is a standard model for arithmetic (a model with* ω*–type ordering).*

One of the philosophically interesting consequences of the application of Tennenbaum's theorem is that the set of models singled out with its help consists of those ω models, where ω is computable (Quinon & Zdanowski, 2007). Those models are called "intended" and form a proper subset of standard models.

The intended model of arithmetic,[7] is such a model where functions of addition and multiplication are interpreted as computable functions.[8] Tennenbaum's theorem establishes a connection between a meta-mathematical property of being computable by arithmetical functions, and the order of the elements of the set of natural numbers. Thus, in the most general lines, computational structuralism is a position, according to which the concept of natural number and the concept of computation are closely related.

The usual way of criticising computational structuralism is, again, by pointing out the vicious circle or infinite regression of definitions that threatens the proposed account of natural numbers. The criticism goes as follows: in order to

[7] Intended models of arithmetic are identified up to a computable isomorphism.

[8] The model of arithmetic is intended for both theories **PA1** and **PA2**.

define the concept of natural number, one needs to use the concept of computation, whereas every concept of computation is defined on the domain of (some representation of) natural numbers. Thus, the vicious circle or the necessity to assume that there is an intended interpretation of what to compute means, or that the intended model of arithmetic is distinguished from within other models.

Analogously, it is pretty straightforward that the concept of computation falls itself into a vicious circle, as in order to account for what "to compute" means, referring, for instance, "to be computed on a Turing Machine", necessitates to account for which entities are suitable for computing with (in the case of TM-computations, what can be the input for a Turing Machine). Since the question asked about the input precedes the definition of computing, which is just being given, one cannot use the concept of computing to define which sequences can be used for the input.

More precisely,

> [T]he Church-Turing Thesis states that Turing Machines formally explicate the intuitive concept of computability. The description of Turing Machines requires description of the notation used for the INPUT and for the OUTPUT. The notation used by Turing in the original account and also notations used in contemporary handbooks of computability all belong to the most known, common, widespread notations, such as standard Arabic notation for natural numbers, binary encoding of natural numbers or stroke notation. The choice is arbitrary and left unjustified. In fact, providing such a justification and providing a general definition of notations, which are acceptable for the process of computations, causes problems. This is so, because the comprehensive definition states that such a notation or encoding has to be computable. Yet, using the concept of computability in a definition of a notation, which will be further used in a definition of the concept of computability yields an obvious vicious circle. (Quinon, 2018, p. 338)

In this section, I explained similarities between the process of defining the concept of natural number, the process of accounting for the concept of computation, and the formulation of an anti-mechanist argument based on Gödel's incompleteness theorems. All these contexts are related because the way out of the definitional vicious circles proper to the definitional processes within formal theories, is through the necessity of assuming an additional non-formal, meta-theoretical knowledge. In the next section, I will expand on the phenomena of vicious circles and regression ad infinitum.

4. Nested Vicious Circles

Quinon (2018) proposes a taxonomy of what can be called "deviant encodings", that is those encodings—or in different words, sequences of symbolic representations of natural numbers—which are non-computable, but which are formally indistinguishable from computable encodings. For instance, in its simplest form the problem presents itself as follows:

> The problem in its purely syntactical version can be formulated as follows. In a definition of Turing computability, one of the aspects that needs to be clarified is the characterization of notation that can be used as an input for a machine to process. If a Turing Machine is supposed to explicate the intuitive concept of computability it is necessary to explain, which sequence of numerals can be used as an input without the use of the concept of computability. That means, we cannot simply say: "sequences that can be used as input are the computable ones" as we have not yet defined what it means "to be computable". (Quinon, 2018, p. 340)

Deviations refer to non-computable sequences that cannot be distinguished within the general formal context from sequences that are computable and can be used in computations. In this paper, I use the expression "deviant encoding" independently of the ontological framework within which natural numbers are understood. Quinon (2018) claims that the phenomenon of deviant encodings persists independently of which ontological status we assign to objects of computations (e.g., natural numbers, sequences of symbols, etc.). Quinon (2018) hypothesizes that the phenomenon of deviant encodings persists independently of the philosophical standpoint and provides an analysis of the following simplified standpoints: (i) purely mechanical/syntactical approach (nominalism, entwined mathematical concepts); (ii) notations have meanings (mild realism); (iii) semantics comes first (radical realism, platonic insight).

The study of conceptual "deviations" is conducted for a simplified framework where:

- on the syntactic level there are uninterpreted inscriptions, and where functions are string-theoretical generating string values from string arguments;
- on the semantic level there are interpretations that can range from the conceptual content ascribed to initially uninterpreted symbols, to Platonic abstract objects, and where functions are number-theoretical sending numbers to numbers;
- between the two levels there is defined a function of denotation.

Deviations occur on each level. Thus, there exist "deviant encodings" deviations that happen on the syntactic level; "deviant semantics" deviations that happen on the semantic level; "unacceptable denotation function" deviations of the denotation function.

The simplified framework is inspired by Shapiro (1982), who distinguishes string-theoretic functions from number-theoretic functions and searches for "acceptable", that is "non-deviant", ways of associating their domains. The framework is further used by other researchers. Rescorla (2007) uses it to study behaviour of denotation functions which associate numerals (symbolic representations of natural numbers) to natural numbers (abstract entities) in a non-computable manner. There is a continuum of such mappings.

The expression “deviant encodings” has been used differently by Copeland and Proudfoot (2010) for whom the deviations relate to encodings, or enumerations, of Turing Machines. The authors claim that a deviant encoding happens when the omniscient programmer “winks at us” to let us know when the number of a Turing Machine (from some standard encoding of Turing Machines), which is being currently processed by some sort of Halting Machine (a machine computing which Turing Machines stop on an input 0), refers to a machine that stops. In this way, the Halting Machine computes the halting function, which is an uncomputable function. The “wink” of the omniscient programmer gets encoded in the syntactic structure of the numerals: the numerals representing the machines that stop, have a special form—for instance—are even (their general syntactical form can be reduced to “2*n*” where “*n*” is any numeral). Copeland and Proudfoot mean by a deviant encoding such a standard enumeration of Turing Machines where the encoding is enriched by an extra-formal feature impersonated by the omniscient programmer (a Turing oracle). This is a specific case of a more general problem where deviant encodings refer to encodings representing natural numbers.

An occurrence of the phenomenon of deviant encodings involving all the levels, is the case of the Semantical Halting Problem (van Heuveln, 2000). Imagine, you have encoded Turing machines with some standard—computable, thus non-deviant—encoding, and that you believe that symbols have meanings or interpretations. It can happen that even if your syntax is generated in a recursive manner, your semantics is not following any recursive rules. The Halting Machine that processes encodings of Turing Machines is designed to process information on syntax in an algorithmic manner. If inputted with a given non-computable enumeration of Turing machines, the machine will process those non-computable encodings as if it were a standard notation. Again, there is no effective way of defining which semantics are acceptable and which are deviant.

I call “nested vicious circles” the hierarchies of vicious circles that keep reappearing at every stage of syntactical and semantic complexity of the presented picture.

To give an example of a philosophical position outside the strict theoretical context discussed in this paper, the phenomenon of deviant encodings appears as well in the case of concrete computations.

> In our ordinary discourse, we distinguish between physical systems that perform computations, such as computers and calculators, and physical systems that don’t, such as rocks. Among computing devices, we distinguish between more and less powerful ones. These distinctions affect our behaviour: if a device is computationally more powerful than another, we pay more money for it. What grounds these distinctions? What is the principled difference, if there is one, between a rock and a calculator, or between a calculator and a computer? Answering these questions is more difficult that it may seem. (Piccinini, 2010)[9]

[9] See also Piccinini’s (2015).

In (2020), Quinon notes that the phenomenon of nested vicious circles, relating to the concept of computability, does not disappear in the case of explicit inter-definiability between the concept of natural number and the concept of computation, as established by computational structuralism. As I have already described above, the criticism of computational structuralism consists in pointing at the choice between the definitional vicious circles or the necessity of making extra-formal arbitrary assumptions.

The way of extra-formal assumptions is investigated by Button and Smith (2012) who observed that when the concept "natural number" is explicated for, the concepts used in this explication, such as "to compute" or "finite" need to be accounted for on their turn, etc. In consequence, claim the authors, this problem cannot be tackled by offering more mathematics. An arbitrary decision regarding the meaning of some concept is necessary for the argument from Tennenbaum's theorem to work. However, as they claim in a slightly undermining way, this is a philosophical problem: "Suffice it to note that our discussion of Tennenbaum's Theorem illustrates a familiar moral: philosophical problems which are supposedly generated by mathematical results can rarely be tackled by offering more mathematics" (Button & Smith, 2012, p. 120).

Dean (2014) is similarly sceptical when it comes to the purposefulness of using Tennenbaum's theorem to formally single out the standard model of arithmetic. However, differently to Button and Smith, Dean develops a full-fledged philosophical position. It is a Putnam-style model-theoretic realism for the concept of computation (Putnam, 1980). Dean claims that there is no point in trying to find external arguments to distinguish between various standard and non-standard models neither of arithmetic, nor of recursive theory. We should rather use the richness of the model-theoretic universe for studying structural properties of the concept of computation. Dean claims that Tennenbaum's phenomenon shows that there exists a continuum of pairs: a model of arithmetic and computation in this model of arithmetic. In consequence, the Tennenbaum's result instead of contributing to singling out the standard model of arithmetic, it indicates that non-computable ω-models of arithmetic exist (the so called deviant or weird permutations) with a corresponding concept of computation defined within the model.

The vicious circle faced by computational structuralism, differs from the vicious circles that are the focus of Quinon (2018). There, I was only concerned by the concept of natural number being indirectly involved in the definition of what "to compute" means. Conceptual structuralism needs to handle a slightly more elaborate idea. Its objective is to explicate the concept of natural number, identified with the standard model of arithmetic. Its solution consists in using the idea that natural numbers, and in particular those which are defined by Peano's axioms, are the entities used for counting and computing. In consequence, natural numbers are defined in terms of computations. However, and this is where the vicious circle arises: one of the characteristic features of the concept of computation is that computation is always defined on some given domain.[10] This domain is always identifiable with the structure of natural numbers. I discuss the nested vicious circles in this context in (Quinon, 2020).

5. Conceptual Engineering and Conceptual Fixed Points

One of the promising ways out of the impasse consists in embracing that the circularity in the account of what "to be computable" and what "natural number" mean is due to limitations of conceptual analysis. Similarly to other scientific concepts, when analysis is conducted within the strict scope of a given formal theory, one often ends up with a necessity to use the concept which is being defined in the account of some concept used for its definition. Philosophers and logicians see in this feature of conceptual analysis both an advantage that enables us to understand more about the conceptual structure of the world (Dean, 2014), and a problem that blocks science from progress (Maddy, 2007). Rescorla (2007) identifies problems with conceptual analysis related to the concept of computation. In their paper (2012), Button and Smith claim that Tennenbaum's theorem is of no use to a philosopher who wants to distinguish the standard model from other possible models of arithmetic.

Quinon (2018) suggests that there is no fully satisfactory way out from vicious circles in definitions, resulting from conceptual analysis. Approaching the concept of computation and the concept of natural number from another methodological perspective, seems to be more fruitful. For instance, in recent years a particular type of conceptual work gained quite a bit of popularity, it is called *conceptual engineering*. What I try to convey in this section is that the new research on conceptual engineering actually provide additional insight into the possible ways of thinking about non-mathematical or non-formal knowledge.

According to Cappelen (2018), conceptual engineering is concerned with the assessment and improvement of concepts. As highlighted by Cappelen and Plunkett:

> since it's unclear and controversial what concepts are (and whether there are any), it's better to broaden the scope along the following lines:
>
> **Conceptual Engineering** = (i) The assessment of representational devices, (ii) reflections on and proposal for how to improve representational devices, and (iii) efforts to implement the proposed improvements. (2020, p. 3)

Researchers involved in developing the methodology of conceptual engineering realised that the method reaches its limits when concepts which are fundamental to the given theory are being scrutinised. They call it "conceptual fixed points". The most extensive reflection has been done in the area of ethics (Cap-

[10] A non-realised Gödel's objective consisted in finding an "absolute" concept of computation, *i.e.*, such a concept of computation that does not depend on any domain.

pelen et al., 2020), but Eklund (2015) extends it to formal contexts and concepts such as “truth”, “belief”, or “existence”. In addition to traditional arguments used in ethical contexts, such as “Kantian philosophy [with its regulative ideas], or from a naturalistic philosophy according to which what is innate severely constrains which concepts we can use”, Eklund considers basic formal concepts in the spirit of rigid designators.

In moral philosophy, “the moral fixed points” are those moral propositions that are moral truths which always need to be incorporated into a moral system. A normative system which fails to incorporate such propositions is not a moral system, but a normative system of some other kind. The leading example of such a moral fixed point is the proposition “It is wrong to engage in the recreational slaughter of a fellow person” (Cueno & Shafer-Landau, 2014).

Eklund (e.g., 2015, Chapter 5) extends this phenomenon to frameworks outside moral philosophy and, as he calls it, the “thinnest” normative words like “good”, “right”, “ought”. Eklund observes that in each conceptual framework, concepts exist that are difficult, if not impossible, to engineer. “Truth” is one of those concepts. People care about truth, writes Eklund, and they do not care about some conceptually engineered concept “truth*”. In consequence, truth is a concept that should keep a fixed position in a conceptual framework, and refer to the natural kin of assertions and beliefs. Similarly, “existence” is a conceptual fixed point. Eklund opposes the claim from the contemporary meta-ontological debate, where it is assumed “that there are alternative notions of existence that can be employed”. He claims that, similarly as in the case of “truth”, a conceptual framework that would result from adapting a conceptually engineered concept of “existence” would need to adjust its other key concepts in such a way that the resulting framework would be isomorphic to the initial one. Thus, “One cannot, so to speak, selectively engineer the quantifier”.

> Suppose we set out to conceptually engineer truth. Insofar as the job description of truth is that of being the property our beliefs and assertions aim at, the engineering project would be that of finding a property more adequate to that job description. But by what has been noted about Stich’s argument, it is hard even properly to conceive of a practice of belief or assertion that is guided by a different property. (Eklund, 2015, p. 378)

There is one last thing that I consider worth mentioning while talking about conceptual fixed points and mathematical concepts, in particular the concept of computation, that is a possible proximity between conceptual fixed points and fixed points that are traditionally analysed in mathematics in the context of diagonalisation. At first sight, they do not have much in common[11] as conceptual fixed points relate mostly to the cross-model intended interpretation of a concept, whereas diagonalisation is about self-reference and vicious circles. Conceptual fixed points are concepts interpreted in, what we call in the philosophy of math-

[11] I might be wrong, but I will not try to sort it out in this paper.

ematics, their intended models. In different words, a fixed point consists of the pair the engineered concept corresponding to the intended meaning of the concept, or—to borrow Eklund's expression—the interpretation that "people care about", and a possible world of interpretation, which actually corresponds to the intended model of this concept. Both, the concept of natural number and the concept of computation are in this sense conceptual fixed points. A more careful look should be applied to those two phenomena, but in this paper I will just leave it without further comment.[12]

6. The Church-Turing Thesis as a Carnapian Explication

Another methodological framework that offers a solution for conceptual structure escaping conceptual analysis is the method of Carnapian explication. Quinon (2019) explores the idea that the structure of the concept of computation, accounted for with the Church-Turing thesis, is best understood through the method of explication. This section is devoted to the presentation of the method of explication for the concept of computation, and also for the concept of natural number.

Treating the concept of computation, as accounted for in the Church-Turing thesis, as a Carnapian explication has multiple advantages, namely, it overcomes problems of conceptual analysis; it explains how one intuitive concept of what "to be computable" means can be translated into a multitude of extensionally equivalent formal concepts of "to be computable" in a specific formal concept means; it finally provides a ground for thinking of mathematical or formal concepts as "open-textures" evolving through time (Makovec & Shapiro, 2019); it also relates to the initial intuitive prescientific concept with the formal concept, because an explication relies on an existing meaning, and offers a specification which offers the best possible fit in a given context.

An explication in the Carnapian sense consists in introducing new formal concepts to the scientific language coined on the basis of everyday concepts. In different words, it is a procedure of transformation from an inexact prescientific concept into a scientific one. Moreover, an explication consists in providing a scientific concept within a given context, within an existing theory. It is done in two steps:

- The clarification of the explicatum
- The specification of the explicatum

The rationale for clarification is that a given term may have many different meanings in ordinary language. Unless one of these meanings is clearly picked

[12] If you want to get a more formal description of this phenomenon, you can think of hybrid modal logics which provide a framework for thinking of epistemic access to other possible worlds from the perspective of the selected distinguished world.

out from the start and the context of its use is clearly indicated, it is unlikely that the method of explication will yield a useful result. Clarification serves this purpose. As Carnap explains, "[a]lthough the explicandum cannot be given in exact terms, it should be made as clear as possible by informal explanations and examples" (Carnap, 1950, p. 3). Quinon (2019) highlights the importance of the clarification stage, the stage which has traditionally been underestimated.

A clarification of the explicandum enables the next step of the explication process, a specification of the explicatum and formulation of the exact concept in the targeted context.

Since several clarifications most often can be foreseen, and several scientific contexts are available, one pre-scientific concept can be explicated in various manners. In order to decide which explication is the most successful, Carnap proposes four criteria that can be applied for assessing the value of an explication, and also for comparison between available options.

- SIMILARITY TO THE EXPLICANDUM: most of the cases in which the explicandum has so far been used, the explicatum can be used; however, close similarity is not required, and considerable differences are permitted.
- EXACTNESS: the rules of use of explicatum have to be given explicitly and precisely, for example, by providing a concept with the formal definition.
- FRUITFULNESS: shall be "useful for the formulation of many universal statements".
- SIMPLICITY: an explication should be as simple as the previous three allow it.

I think that it is worth investigating whether abandoning the path of analysis and taking the path of explications could offer an additional insight into the conceptual structure of formal concepts, and also informal concepts lying in the foundations of their formalization. The idea is that every formal concept is—at least in subjective arithmetic (to borrow Gödelian terminology)—grounded upon, or issued from, an everyday intuitive, pre-scientific concept. The next section is devoted to a preliminary investigation into the possibility of extending the idea that the method of explication, consisting in building up the formal concept out of the intuitive concept, is anyhow relevant to the anti-mechanist argument against the computability of mind using Gödel's incompleteness theorems.

Both intended interpretations determined in the consequences of accepting conceptual fixed points solution and the choice of the formal aspect, and the formal context at the stage of the concept clarification in the process of Carnapian explication, share a similar threat. In the case of a fixed point solution and in the case of clarification an agent needs to take an arbitrary decision regarding the intended interpretation.

7. Theory of Mind and Computations

In this section, I propose an additional complication to the method of Carnapian explication, which is a temporary, or a phylogenic, aspect of conceptual development.

The method of Carnapian explication enables introducing new formal concepts to the language by transforming an intuitive pre-scientific concept into a new scientific concept within some formal context. Usually, at the stage of clarification one chooses the meaning that will guide the formalisation of the intuitive pre-scientific concept and also the targeted formal context. What I propose in this section, is an additional dimension to the clarification stage: a relativisation to the phylogeny of the formal concept. At the stage of clarification, in addition to deciding which aspect of the intuitive concept one wants to formalise, one needs to realise that each concept develops. The phylogenic development of the concept of natural number and the concept of computation is studied in Shapiro on open-texture (2013).

The relation between the concept of computation and the concept of natural number underwent a very dynamic development. In consequence, the set of potential clarifications of intuitive concepts of computation and of natural numbers have grown. What is interesting from my perspective, is that computability is today an expected feature of natural numbers. Natural numbers are those mathematical entities that are all day long used for enumerating and computing, for programming, and in various sorts of logistic projects as an underlying discrete structure. Both concepts have become increasingly important in the everyday life of our society. This is called digitalisation.

Various areas of digitalisation are additionally reinforced by the fact that computationalism—even if its formal details are still discussed by philosophers, mathematicians and logicians—is today the mainstream theory of mind. This process is described by Turkle (1984; 2011; 2015) who studies how concepts from computer sciences and robotics have got into common language and how they have changed ordinary people's approach to inter-personal relations or ethical questions.

According to Turkle the intensity in which digitalisation of everyday life develops is strongly connected to the fact that computational language was first used to reformulate our perception of our own mind and consciousness.[13]

[13] Turkle's earlier work related to a similar development of conceptual trends in explanation of phenomena of everyday life that had a place in France in the 1960s and 1970s as a consequence of the spread of psychoanalytical ideas, see her book *Psychoanalytic Politics: Jacques Lacan and Freud's French Revolution* from 1978). In *The Second Self: Computers and the Human Spirit* (1984), Turkle describes these changes that have got into general culture through digitalisation and robotics in the same way as "psychoanalytic culture" penetrated structures of the general social and political life in France: "Psychoanalytic language spread into the rhetoric of political parties, into training programs for schoolteachers, into advice-to-the-lovelorn columns. I became fascinated with how people were picking up and trying on this new language for thinking about the self. I had gone to

When Turkle speaks about her experience with the digitalised society, she compares two experiences:

> My experience at MIT impressed me with the fact that something analogous to the development of a psychoanalytic culture was going on in the worlds around computation. At MIT I heard computational metaphors used to think about politics, education, social process, and, most central to the analogy with psychoanalysis, about the self. (Turkle, 1984, p. 305)

She sees within it a first step in the cultural assimilation of a new way of thinking:

> The essential question in such work is how ideas developed in the world of high science are appropriated by the culture at large. In the case of psychoanalysis, how do Freudian ideas move out to touch the lives of people who have never visited a psychoanalyst, people who are not even particularly interested in psychoanalysis as a theory? In the study of the nascent computer culture, the essential question was the same: how were computational ideas moving out into everyday life? (Turkle, 1984, p. 305)

She searches how "the idea of mind as a program enters into people's sense of who is the actor when they act". A model of the mind that is adapted by society influences how people think about their frustrations and disappointments, their relationships with their families and with their work (Makovec & Shapiro, 2019, p. 305). On the other hand, says Turkle, computers became a new constructed object—"a cultural object that different people and groups of people can apprehend with very different descriptions and invest with very different attributes. Ideas about computers become easily charged with personal and cultural meanings" (Turkle, 1984, p. 308).

In her other books, Turkle studies human attachment to objects. In the volume of essays *Evocative Objects: Things We Think With* (2007) she speaks about the attachment that people, many of her friends, have developed with physical objects. In her book, *Alone Together* Turkle (2011) extends her observations to different types of automated artificial agents, such as virtual agents mediated by electronic support, or robots. In a series of social experiments, where she asked her subjects to interact with an automated artificial agent, she observed that the stronger attachment develops in the most vulnerable members of our society, such as neglected children with unfulfilled emotional needs, or with old people suffering from a lack of human interactions. Our natural inclination to form emotional attachment with humans, and with objects in the absence of humans, might soon lead to even more human-AI interactions. Those interactions are obviously

France to study the psychoanalytic community and how it had rein- vented Freud for the French taste, but I was there at a time when it was possible to watch a small psychoanalytic community grow into a larger psychoanalytic culture" (Turkle, 1984, pp. 304–305).

structured in a very particular, very automated, way, which even more strongly influences the digitalisation of the language we use.

Krajewski makes a similar observation in the last section of the paper.

> Our attitude toward the arguments of Lucas, Penrose, and others is shaped mostly by our general vision of machines and minds. And this vision adjusts with changes of civilization. For the youth of today, if I may judge from listening to my students, our computerized world makes it easier to accept the idea that anything is mechanizable—including the mind. (2020, p. 49)

I propose a hypothesis that at least part of the confusion regarding the specificity of the conceptual structure of the concept of computation contributes to the confusion regarding the nature of human reasoning and the human mind. In consequence, I claim that—at least partially—the "feeling" that there are non-computational processes is due to the complexity of the conceptual structure of the concept of computation.

8. The Lucas-Penrose Argument and Extra-Formal Concepts

Let me now come back to the anti-mechanist argument against computability of mind based on Gödel's incompleteness theorems.

In the first part of this section, I reconstruct Krajewski's claim according to which, in order to make the anti-mechanist argument work, one needs to add an extra-formal assumption stating the consistency of the underlying theory, that is, the theory corresponding to the human mind. The core of Krajewski's criticism is as follow: it is not possible to formalize the extra-formal assumption and therefore, the whole of Lucas' argument is fallacious. I disagree with Krajewski's claim that formalization of the extra-formal assumptions is not possible. There are contemporary philosophical methods that might enable formulation of such a formalization. As example, in the previous sections, I have presented the methodological and conceptual framework was based on Carnapian explications. Instead, I focus on another problem, which the issues from an internal characteristic of formal contexts, namely on the part of the argument, which leads to a circular reasoning. In order to show that the human mind (T_{HM}) outperforms a machine (T_M), one needs to assume that the human mind is consistent and knows it (and in this way outperforms a machine that can never "know" if it is consistent or not). Observe, that I do not reject Krajewski's conclusion, but I point at a fallacy in a proof. Again, I have already discussed how the method of conceptual engineering enables structured thinking of extra-formal assumptions and the resulting circular reasoning.

In the second part of this section, I will continue my investigation of possible extra-formal assumptions relative to the anti-mechanist argument based on Gödel's incompleteness theorems.

The Lucas' anti-mechanist argument based on Gödel's incompleteness theorems consists of two parts. Firstly, Gödel's results establish that each sufficiently rich consistent theory admits a Gödel sentence and also that none such theory can prove its own consistency.

Let T be a consistent theory containing arithmetic, let φ_T be the Gödel's sentence for the theory T.

$$Con(T) \rightarrow T \nvdash \varphi_T$$
$$Con(T) \rightarrow T \nvdash Con(T)$$

Moreover, it is broadly known that an inconsistent theory proves any sentence, but Gödel's incompleteness theorems do not apply to an inconsistent theory.

Secondly, human mathematicians can work with subsequent increasingly stronger theories,

$$T_1 = T \cup Con(T)$$
$$T_2 = T_1 \cup Con(T_1)$$
$$\vdots$$
$$T_{n+1} = T_n \cup Con(T_n)$$

which—for some defenders of the anti-mechanist argument—signifies that human mathematicians outperform machines. Krajewski objects to this view claiming that the construction of the hierarchy can be fully mechanised. In consequence, he claims that the ability to construct and work with the hierarchy of increasingly stronger theories alone is not sufficient for formulating the anti-mechanist argument. As stated by Krajewski, additional assumptions are missing.

> In addition to Gödel's results, at least two assumptions that are not self-evident are used in the above reasoning. First, every exact proof of our consistency can be formalized, second, it is possible to express "our consistency". […] If this is accepted, one could question the second point. It is not clear at all how one can express "our consistency". Basically there are two options to express this: either (i) by the common sense statement "I am consistent" or (ii) by a formal counterpart to this statement. Let us consider them in turn.
>
> In case (i) we refer to a common sense statement, which have no connection to formal considerations. Hao Wang (1974, pp. 317–320) reflected on just this statement and believed that it is not provable. […] If that were possible, it would mean that we are not machines, or that we are not even equivalent to machines in the realm of proof-producing reasoning. We certainly may believe that, but it is no more than a general feeling.
>
> In case (ii) we consider the formal counterpart to a loose statement expressing consistency […]. The usual meaning of the statement refers to the will to avoid contradictions, to the reliability of our vision of the world, to the claim that the methods used by mathematicians are unfailing. The sentence *Cons* or any other similar arithmetical formula is rather far from those ideas. Thus, while something is strictly proved, it is unclear to what extent the conclusion conveys our consistency. (2020, pp. 47–48)

Krajewski's reasoning can be reconstructed as follows. Applying the formal predicate "being consistent" can only apply to a formal theory. Applying the formal predicate "being consistent" to anything else than a formal theory is a categorical mistake. In consequence, if "consistency" is to be a predicate applying to on the human mind, the mind must have certain formal properties and needs to be identified with a theory. The following options exist:

- If human mind is a theory and it is consistent, then as to all other theories, a Gödel's sentence applies to it and the human mind encounters the same constraints as any theory (a machine).
- If the human mind is a theory and it is inconsistent, then Gödelian argument limitations do not apply at all.

If the human mind is a theory, a human disposing of a mind cannot know—from the formal point of view—if it is consistent or not. In consequence, in order to prove that the human mind outperforms a machine, a second extra-formal additional assumption needs to be made. It has to be assumed that the human mind is indeed consistent. This assumption can be done in one of the two ways. "Case (i)", "I am consistent" cannot be formalised. "Case (ii)", there exists a formal counterpart of "I am consistent".

My analysis of "case (i)" is in line with the analysis of Krajewski. If "I am consistent" is an informal statement, it is useless for any formal proof. And here we speak of being able to prove more than a machine. Whereas Lucas' argument is supposed to be a formal proof of the superiority of the human mind over a machine.

My analysis of "case (ii)" differs from Krajewski's analysis. His argument returns to the idea that each formalisation of the informal "I am consistent" remains—maybe more informed or more precise—but is still an informal account. As such it is useless for any formal proof. I think that the conclusion from (ii) is different. An agent can find a formal counterpart of the statement "I am consistent", or rather "the theory constituting my mind is consistent". The framework of the Carnapian explications enables us to understand how it can be done.

I also assume that an agent can recognise their own consistency. This insight is available to a human being, while it is—on the grounds of the second of Gödel's incompleteness theorem—unavailable to a machine. This extra-formal assumption is necessary for formulating an anti-mechanist argument against the computability of the mind. It is also exactly at this point where a vicious circle occurs. We are in the act of proving that the human mind outperforms a machine, and so one cannot in this proof assume that human mind is consistent.

Another possible extra-formal assumption that can be made in order to enable the anti-mechanist argument based on Gödel's incompleteness theorem, is the ability to refer to the intended model of arithmetic.[14] Instead of assuming that the human mind is consistent (i.e., assuming that the theory underlying all human reasoning is a consistent theory, which does not prove both a φ and a $\neg\varphi$, for every φ), in order to use Gödel's incompleteness theorems to support the anti-mechanist argument, one can assume that the human mind is able to refer to the intended model of arithmetic. The assumption that the human mind can refer to the intended model of arithmetic disables the possibility that the Gödel sentences get to have non-standard Gödel numerals.

In the way it is usually interpreted—in particular in the context of philosophical argumentation supporting the anti-mechanist argument that the human mind is non-computable—Gödel's incompleteness theorems provide us with the information from the perspective of a formal system. The semantical aspect is taken for granted. When the model-theoretical reasoning is applied, Gödel's incompleteness theorems indicate that there exist non-standard models in which the (non-standard) Gödel number of the proof for Gödel's incompleteness theorems has its (semantical) reference. It also means, that there exist models where the Gödel (non-standard) number of the proof for the negation of Gödel's first theorem, has an interpretation as a (non-standard) natural number.

What is famously referred to by Gödel's platonism is his belief that there is a model of arithmetic in which all arithmetical truths are satisfied. This is obviously not the intended model of arithmetic that humans have privileged cognitive access to, but the model of arithmetic in objective mathematics (Gödel, *1951).

9. Conclusions

Additionally to the critical analysis of Krajewski's rejection of the anti-mechanist based on Gödel's incompleteness theorems to which I suggest some possible improvements, my paper is sympathetic to the idea that certain key concepts in formal contexts naturally fall into circular or infinite reasonings. In this way, I try to shift attention from the theory of the human mind and consciousness, to the study of the conceptual structure of the language.

In my paper, I explored similarities between various formal contexts in which key concepts fall into a vicious circle of reasoning. I looked at the formalisation of the concept of natural number, of the concept of computation, and at the concept of consistency in the context of Gödel's incompleteness theorems. I suggested that the way to switch from an informal pre-scientific concept to a full-blooded formal scientific concept formulated in an adequate formal context is best modeled by Carnapian explications. I have also suggested that the phenom-

[14] The intended model is intended for both **PA1** and **PA2** and for this reason I do not make a distinction between the intended model of **PA1** and the intended model of **PA2**. I can think of a philosophical position that makes such a distinction, but for my purpose that would unnecessarily complicate my presentation.

enon of conceptual fixed points offers a methodological framework to think of intended interpretations necessary to jump out of circularity.

REFERENCES

Benacerraf, P. (1967). God, the Devil, and Gödel. *Monist*, *51*, 9–32.

Boolos, G. (1995). Introductory Note to *1951. In: S. Feferman et al. (Eds.), *Collected Works, Volume III, Unpublished Essays and Lectures* (pp. 290–304). Oxford University Press.

Button, T., Smith, P. (2012): The Philosophical Significance of Tennenbaum's Theorem. *Philosophia Mathematica*, *20*(1), 114–121.

Cappelen, H. (2018). *Fixing Language: An Essay on Conceptual Engineering*. Oxford University Press.

Cappelen H., Plunkett D. & Burgess A. (Eds.). (2020). *Conceptual Engineering and Conceptual Ethics*. Oxford University Press.

Carnap, R. (1950). *Logical Foundations of Probability*. Routledge and Kegan Paul.

Copeland, J., Proudfoot, D. (2010). Deviant Encodings and Turing's Analysis of Computability. *Studies in History and Philosophy of Science*, *41*, 247–252.

Cuneo T., Shafer-Landau, R. (2014). The Moral Fixed Points: New Directions for Moral Nonnaturalism. *Philosophical Studies*, *171*, 399–443.

Dean, W. (2014), Models and Computability. *Philosophia Mathematica*, *22*(2), 143–166.

Eklund, M. (2015). Intuitions, Conceptual Engineering, and Conceptual Fixed Points. In C. Daly (Ed.), *The Palgrave Handbook of Philosophical Methods* (pp. 363–385). London: Palgrave Macmillan.

Feferman, S. (1995). Penrose's Gödelian Argument. *Psyche: An Interdisciplinary Journal of Research on Consciousness*, *2*, 21–32.

Gödel, K. (193?), Undecidable Diophantine Propositions. In S. Feferman et al. (Eds), *Collected Works, Volume III, Unpublished Essays and Lectures* (pp. 164–175). Oxford University Press.

Gödel, K. (*1951). Some Basic Theorems on the Foundations of Mathematics and Their Implications [Gödel's 1951 Gibbs lecture]. In S. Feferman et al. (Eds.), *Collected Works, Volume III, Unpublished Essays and Lectures* (pp. 304–323), Oxford University Press.

Halbach, V., Horsten, L. (2005). Computational Structuralism. *Philosophia Mathematica*, *13*(2), 174–186.

Hofstadter, D. R. (1979). *Gödel, Escher, Bach, and Eternal Golden Braid*. New York: Basic Books.

Krajewski, S. (2007). On Gödel's Theorem and Mechanism: Inconsistency or Unsoundness is Unavoidable in any Attempt to 'Out-Gödel' the Mechanist. *Fundamenta Informaticae*, *81*, 173–181.

Krajewski, S. (2020). On the Anti-Mechnist Arguments Based on Gödel's Theorem. *Studia Semiotyczne*, *34*(1), 9–56.

Lucas, J. R. (1961). Minds, Machines and Gödel. *Philosophy*, *36*(137), 112–127.

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

Lucas, J. R. (1990). A Paper to Read to the Turing Conference at Brighton on April 6th, 1990. Retrieved from: http://users.ox.ac.uk/~jrlucas/Godel/brighton.html

Lucas, J. R. (1996). Minds, Machines and Gödel: A Retrospect. In P. Millican, A. Clark (Eds.), *Machines and Though* (pp. 103–124). Oxford University Press.

Maddy P. (2007). *Second Philosophy. A Naturalistic Method*. Oxford University Press.

Makovec, D., Shapiro S. (Eds.). (2019). *Friedrich Waismann. The Open Texture of Analytic Philosophy*. New York: Springer.

Nagel, E., Newman J. R. (1958). *Gödel's Proof*. New York University Press.

Nagel, E., Newman J. R. (1961). Answer to Putnam. *Philosophy of Science*, *28*, 209–211.

Penrose, R. (1989). *The Emperor's New Mind: Concerning Computers, Minds and The Laws of Physics*. Oxford University Press.

Penrose, R. (1994). *Shadows of the Mind: A Search for the Missing Science of Consciousness*. Oxford University Press.

Piccinini, G. (2010). Computation in Physical Systems. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University.

Piccinini, G. (2015). *Physical Computation: A Mechanistic Account*. Oxford University Press.

Plunkett D., Cappelen, H. (2020). A Guided Tour of Conceptual Engineering and Conceptual Ethics. In: H. Cappelen, D. Plunkett, A. Burgess (Eds.), *Conceptual Engineering and Conceptual Ethics* (pp. 1–26). Oxford University Press.

Post, E. (1941). Absolutely Unsolvable Problems and Relatively Undecidable Propositions—Account of an Anticipation. In M. Davis (Ed.), *The Undecidable* (pp. 338–433). Hewlett, N. Y.: Raven Press.

Putnam, H. (1960). Minds and Machines. In S. Hook (Ed.), *Dimensions of Mind: A Symposium* (pp. 138–164). New York: New York University Press.

Putnam, H. (1980). Models and Reality. *Journal of Symbolic Logic*, *45*(3), 464–482.

Putnam, H. (1995). Review of The Shadows of the Mind. *Bulletin of the American Mathematical Society*, *32*(2), 370–373.

Quine, W. V. O. (1970). *Philosophy of Logic*. Harvard University Press.

Quinon, P. & Zdanowski, K. (2007). Intended Model of Arithmetic. Argument from Tennenbaum's Theorem. In S. B. Cooper et al. (Eds.), *Computation and Logic in the Real World* (pp. 313–317). Berlin: Springer-Verlag.

Quinon, P. (2014). From Computability Over Strings of Characters to Natural Numbers. In A. Olszewski, B. Brożek, P. Urbańczyk (Eds.), *Church's Thesis, Logic, Mind & Nature* (pp. 310– 330). Warsaw: Copernicus Center Press.

Quinon, P. (2018). Taxonomy of Deviant Encodings. In: F. Manea, R. Miller, D. Nowotka (Eds.), *Sailing Routes in the World of Computation* (pp. 338–348). Berlin: Springer-Verlag.

Quinon, P. (2019). Can Church's Thesis be Viewed as a Carnapian Explication? *Synthese*, Online First.

Quinon, P. (2020). Implicit and Explicit Examples of the Phenomenon of Deviant Encodings. *Studies in Logic, Grammar and Rhetoric*, *63*(76), 53–68.

Rescrola, M. (2007), Church's Thesis and the Conceptual Analysis of Computability. *Notre Dame Journal of Formal Logic*, *48*(2), 253–280.

Shapiro, S. (1982). Acceptable Notation. *Notre Dame Journal of Formal Logic*, *23*(1), 14–20.

Shapiro, S. (1998). Incompleteness, Mechanism, and Optimism. *Journal of Philosophical Logic*, *4*, 273–302.

Shapiro, S. (2003). Mechanism, Truth, and Penrose's New Argument. *Journal of Philosophical Logic*, *32*, 19–42.

Shapiro, S. (2013). Computability, Proof and Open-texture. In A. Olszewski, J. Wolenski, R. Janusz (Eds.), *Church's Thesis After 70 Years* (pp. 420–455). Berlin: Walter de Gruyter.

Turing, A. (1950). Computing Machinery and Intelligence. *Mind*, *59*, 433–460.

Turkle, S. (1978). *Psychoanalytic Politics: Jacques Lacan and Freud's French Revolution*. New York: Basic Books.

Turkle, S. (1984). *The Second Self: The Second Self: Computers and the Human Spirit*. MIT Press.

Turkle, S. (2011). *Alone Together: Why We Expect More from Technology and Less from Each Other*. New York: Basic Books.

Turkle, S. (2015). *Reclaiming Conversation: The Power of Talk in a Digital Age*. London: Penguin Press.

van Heuveln, B. (2000). *Emergence and Consciousness: Explorations Into the Philosophy of Mind via the Philosophy of Computation* [Unpublished Ph.D. thesis]. State University of New York, Binghampton.

Wang, H. (1974). *From Mathematics to Philosophy*. Routledge and Kegan Paul.

STUDIA SEMIOTYCZNE (SEMIOTIC STUDIES), t. XXXIV, nr 1 (2020), s. 267–278
ISSN 0137-6608, e-ISSN 2544-073X
DOI: 10.26333/sts.xxxiv1.12
Article

PANU RAATIKAINEN [*]

REMARKS ON THE GÖDELIAN ANTI-MECHANIST ARGUMENTS

SUMMARY: Certain selected issues around the Gödelian anti-mechanist arguments which have received less attention are discussed.

KEYWORDS: Gödel, incompleteness, mechanism.

Introduction

There is no question that Gödel's two incompleteness theorems (Gödel, 1931)[1] are deep and important mathematical results which have significant philosophical implications (e.g., Raatikainen, 2005). It seems that the idea that they demonstrate the superiority of the human mind over computing machines and formalized theories in particular is very attractive and natural, as it is put forward again and again (Krajewski, 2020). The view that the human mind is, in some sense, equivalent to a finite computing machine or a formalized theory is called "mechanism". The popular idea, famously advocated by J.R. Lucas (1961; 1996), is that Gödel's results demonstrate, with mathematical certainty, that the human mind can surpass or "out-Gödel" any computing machine and formalized theory, and that mechanism can therefore be refuted for good. Roger Penrose (1989; 1994; 1995; 1997) has prominently put forward very similar views. The literature

[*] Tampere University, Philosophy. E-mail: panu.raatikainen@tuni.fi. ORCID: 0000-0003-0308-3051.

[1] For an accessible survey, see, e.g., (Franzén, 2005; Raatikainen, 2020).

is vast, but I shall focus here on Lucas's classic key claims, which have been enthusiastically repeated, almost verbatim, again and again.[2]

However, numerous logicians and philosophers (beginning from Gödel and Turing themselves; and see e.g. Putnam, 1960; Boolos, 1968; Davis, 1990; 1993; Feferman, 1995; 2009; 2011; Shapiro, 1998), including myself (Raatikainen, 2005), have argued that such straightforward anti-mechanist arguments grounded on Gödel's theorems are flawed. Krajewski (2020) both surveys the history of such arguments and elaborates various problems with them. I don't want to repeat those critical arguments here in any detail. Instead, I shall emphasize and discuss certain selected issues around the Gödelian anti-mechanist arguments which have received less attention, and which to my mind deserve to be noticed. I shall assume that the reader is familiar with the basic ideas and concepts of this debate.

The Limits of Machine Talk

The Gödelian argument against mechanism is standardly formulated in terms of Turing machines and their Gödel sentences, which the machines are incapable of "producing as being true" but which the human can allegedly see to be true. However, such talk about "the Gödel sentence of the machine" is, strictly speaking, nonsense (cf. Gaifman, 2000).

The theory of computability and its notions of decidability and computability, and Turing's groundbreaking analysis of these notions in terms of imaginary idealized machines, are certainly essential for the general versions of the incompleteness results: a formalized theory is by definition required to have a decidable set of axioms and a decidable proof relation.[3] Consequently, if the language of the theory is suitably coded by numbers, there are Turing machines which can effectively generate exactly the code-numbers (the "Gödel numbers") of the theorems of the theory: the set of those code-numbers is thus, technically speaking, recursively enumerable (r.e. for short). But that is it.

A Turing machine does not in itself correspond to any specific formalized theory, and just does not have a specific Gödel sentence of its own. Even if a Turing machine is incapable of producing ("as true") a sentence under one given coding, it may well produce that sentence under many other codings. Consequently, the suggested idea of "out-Gödeling" a machine in itself makes no

[2] There is a conspicuously enthusiastic entry (Megill, 2012) on the Gödelian anti-mechanistic arguments in the Internet Encyclopedia of Philosophy, which for its part suggests that the issue is still very much alive.

[3] See, e.g., (Raatikainen, 2020). To be sure, logicians have studied extensively arbitrary sets of axioms, infinitely long sentences and infinitary rules of inference; but in the context of Gödel's incompleteness theorems, this is a standard assumption (though some generalizations exist). Accordingly, in what follows, I shall always use "formalized theory" to mean a theory which has a finite or decidable set of axioms, and a decidable proof relation, and consequently a r.e. set of theorems.

sense. One and the same Turing machine may correspond to very different formalized theories under different codings. And many Turing machines just do not correspond to any formalized theory under any coding. The same holds for recursively enumerable sets of numbers.

The framework of computability theory is, in general, too coarse-grained in this context: all formalized theories which contain Robinson Arithmetic **Q**, from the very weak **Q** itself to the strongest theories of set theory (e.g. **ZFC** + "there exist supercompact cardinals") and beyond, as long as the set of axioms is decidable, are "creative" (in Post's sense), have the same computability-theoretic degree, and are recursively isomorphic with each other (i.e., they are all one-one reducible to each other). Hence computability theory is unable to make any difference between such theories with radically different strengths. As Kreisel was fond of putting it, proof theory begins where computability theory stops. (cf. Odifreddi, 1989, pp. 356–357) Hence, immaculately formulated, the question should be: Can all the truths that are humanly provable be captured by a formalized theory?

My sticking to this issue may strike some as excessive pedantry, but I think there is a real risk here of overlooking some relevant issues. Turing machines, or r.e. sets of numbers generated by them, simply do not stand to each other in the various logical relations that are essential for this topic. For example, it makes no clear sense to ask whether a given r.e. set of numbers can prove the consistency of another given r.e. set. Furthermore, in a fully general consideration, we cannot restrict our attention solely to direct extensions of elementary arithmetic in the same language (or its direct extension). A great many formalized theories have *prima facie* nothing to do with arithmetic; their language may be quite different from the familiar language of arithmetic (think of set theory, for example). There is no direct way of comparing the respective sets of code-numbers, as to whether one is stronger than the other, etc. Relating such theories requires considering the relation of relative interpretability between formalized theories.[4] But at the level of computability theory and r.e. sets, such relations are invisible.

It is certainly possible to continue to talk about Turing machines or recursively enumerable sets of numbers here, with the assumption that some coding ("Gödel numbering") has been fixed. But such a manner of speaking may be misleading and hide some important aspects of the topic. The above facts should at very least be kept clearly in mind. Accordingly, I shall talk, in what follows, as far as possible, only about formalized theories.

[4] Roughly, F_1 is interpretable in F_2 if the language of F_1 can be "translated" into the language of F_2 in such a way that F_2 proves the translation of every theorem of F_1. This notion of interpretability was first given an explicit definition by Tarski in (Tarski, Mostowski, & Robinson, 1953). It had been, however, already used in practice by logicians for some time.

Varieties of Mechanist and Anti-Mechanist Theses

Instead of talking generally about the juxtaposition of mechanism and anti-mechanism, I think it would be useful to distinguish more finely several different theses here which often seem to get conflated in the debate. To begin with, there is:

1. Strong Local Mechanism: The set of humanly provable mathematical truths is equivalent to the set of theorems of a certain explicitly specified formalized theory *F*: "this *F*".

In other words, the mechanist is here supposed to explicitly present a particular formalized theory *F* which is contended to be equivalent with the human mind. However, it is clearly possible to advocate mechanism as a general thesis without such a specific claim:

2. Basic General Mechanism: The set of humanly provable mathematical truths is equal in effect to the set of theorems of some formalized theory.

Finally, we should also distinguish the following, apparently weaker, claim:

3. Weak General Mechanism: The set of humanly provable truths is contained in the set of theorems of some formalized theory.

There also appear to be several different anti-mechanist claims on offer.

4. Weak Anti-Mechanism: It follows from Gödel's incompleteness theorems that Strong Local Mechanism is false.

5. Basic Anti-Mechanism: It follows from Gödel's incompleteness theorems that Basic General Mechanism is false.

6. Strong Anti-Mechanism: The human mind can surpass any given consistent formalized theory (which includes arithmetic) and prove ("see to be true") the Gödel sentence of it.[5]

It seems that Lucas, Penrose and their allies do not always sufficiently distinguish these different theses, but slide from one to another and back again without

[5] This is also the first disjunct of Gödel's famous, more cautious disjunctive thesis, now standardly called "Gödel's disjunction" (Gödel, 1951); the second disjunct says that there are mathematical problems which are absolutely undecidable for the human mind. Gödel suggested that their disjunction follows from the incompleteness results; but he never contended that the first disjunct in itself would follow. Although our Strong Anti-Mechanism is not formulated directly as the opposite of Weak General Mechanism, it is natural to interpret the former as denying the latter.

clearly noticing this. When Lucas, for example, declared that "given any machine which is consistent and capable of doing simple arithmetic, there is a formula it is incapable of producing as being true [...] but which we can see to be true" (Lucas, 1961; my emphasis), he apparently advocated Strong Anti-Mechanism.[6] However, when pressed, Lucas and others often retreat to Weak Anti-Mechanism, or perhaps to an even more specific view. That is, especially when anti-mechanists attempt to circumvent critique, the mechanist view they apparently focus on is even stronger and more specific than Strong Local Mechanism, namely:

7. Naïve Strong Local Mechanism: (i) Strong Local Mechanism; (ii) the human mind knows the equivalence of itself and the specific formalized theory F with mathematical certainty (i.e., the equivalence is itself absolutely provable); (iii) the human mind knows with mathematical certainty that F is consistent.

We can grant Lucas and other anti-mechanists that Naïve Strong Local Mechanism collapses, in the light of the Gödelian facts, into inconsistency. This was already apparent for Gödel himself (Gödel, 1951) and has been repeatedly conceded. But this concession is a rather minute victory for anti-mechanism. For there are many ways to be a coherent mechanist without committing oneself to the Naïve Strong Local Mechanism.[7]

First, and most obviously, one might have general theoretical or empirical reasons for advocating Basic General Mechanism (or Weak General Mechanism), but not Strong Local Mechanism. But what is more, one might perhaps have inductive empirical reasons for believing that a particular formalized theory F corresponds to the human mind, but such reasons are, of course, short of mathematical certainty. On second thought, this seems a much more plausible alternative than the idea that it should be known with mathematical certainty. Finally, a mechanist might believe in the consistency of F, but on grounds that are weaker than absolute mathematical certainty, for example, broadly speaking

[6] Note that Lucas here only requires that the machine be consistent—not that the mechanist, we or anyone know (with mathematical certainty) that it is consistent.

[7] Koellner (2016; 2018a), building on the earlier work of Reinhardt (1985a; 1985b) and Carlson (2005), analyzes some such differences much more rigorously in the context of so-called epistemic arithmetic. He labels roughly the same view I have here called "Basic General Mechanism" as "weak mechanistic thesis"; Reinhardt (1985b) proved that it is consistent. The view that the former is itself knowable with mathematical certainty is called in this tradition the "strong mechanistic thesis" (this view does not occur separately in my listing above); Carlson (2005) showed that it is consistent. Finally, Koellner calls the view roughly corresponding to our Naïve Strong Local Mechanism "super strong mechanistic thesis"; it was proved inconsistent, in this context, by Reinhardt (1985a).

inductive: *F* seems to avoid known paradoxes, no contradiction has so far been derived in it, it has some expected consequences etc.[8] (more of the latter below).

That is, even Weak Anti-Mechanism is as such false, and mere Strong Local Mechanism is not necessarily refuted by Gödel's theorems, unless it is complemented with the further conditions (ii) and (iii) from the definition of Naïve Strong Local Mechanism, and the latter is thus adopted. Hence, there is plenty of room for different mechanistic views which are not vulnerable to any Gödelian counterarguments; and if Weak Anti-Mechanism fails, Basic Anti-Mechanism and Strong Anti-Mechanism are on an even weaker footing.

Questions of Consistency

The standard objection[9] to the Gödelian anti-mechanist arguments builds on the fact that Gödel's first incompleteness theorem has in reality a conditional form, and the alleged truth of the Gödel sentence G_F for a formalized theory *F* depends on the assumption of the consistency of *F*. Therefore, in order to really know that G_F is true one must first know that *F* is consistent.[10] And that is not, in general, transparent.

Lucas and some of his devotees (but also some critics) seem to think that the gist of the objection is to raise doubts about the consistency of the human mind; but I think this is off the mark. The central notion here is absolute provability—what the human mind can prove with mathematical certainty. (In this paper, I use "absolutely provable" and "knowable with mathematical certainty" interchangeably.) Whatever the scope of such knowledge really is, this is a normative concept, and it is not terribly implausible to contend that it consists, by definition, of true sentences and is consequently a consistent whole. The real question is whether this totality of absolutely provable sentences is, by its very nature, such that it cannot, as a matter of absolute mathematical fact, coincide with or be contained in a set of theorems of some formalized theory.

In other words, the critical question is not whether I am consistent and/or whether I can know that I am consistent, but whether a given formalized theory is consistent and whether I can always know with mathematical certainty that it is. The challenge is especially flagrant for Strong Anti-Mechanism. Lucas explic-

[8] Gödel himself, in his Gibbs lecture (Gödel, 1951), was already sensitive to these further conditions (i.e. (ii) and (iii)), when he qualified that the soundness (and, consequently, consistency) of the formalized theory should be known with "mathematical certitude", and reflected the possibility that the human might well know its equivalence with a formalized theory, but only with ”empirical certainty".

[9] The objection goes back to Putnam (1960).

[10] And we know, from Gödel's second incompleteness theorem, that (under certain general conditions) the consistency of *F* cannot be proved inside *F*. In fact, it can be shown that the Gödel sentence G_F for *F* and the formalized consistency statement for *F* are materially equivalent inside *F* (and hence equally unprovable in *F*; see, e.g., Raatikainen, 2020).

itly contends that the human mind can surpass any consistent formalized theory and intuitively prove as true its Gödel sentence. However, that amounts to being able to prove intuitively and absolutely, with mathematical certainty, the consistency of any given formalized theory, if it is in fact consistent. And that is fantastically optimistic indeed. Lucas and his followers greatly underestimate the difficulty of this task. Consistency is (in terms of computability theory) a Π^0_1-complete property. This means that being able to tell whether a given formalized theory is consistent or not would enable one to tell about every Π^0_1 sentence[11] whether it is true or false: one should have an "oracle"[12] for this class of sentences.[13] There is absolutely no reason to believe that the human mind has such miraculous powers. There are many open problems in mathematics which have this form (that is, Π^0_1). Even the best mathematicians have no clue how to know whether they are true or not; the same holds for the corresponding consistency questions.

Lucas (1961) writes, referring now to Gödel's second incompleteness theorem:

> All that Gödel has proved is that a mind cannot produce a formal proof of the consistency of a formal system inside the system itself: but there is no objection to going outside the system and no objection to producing informal arguments for the consistency either of a formal system or of something less formal and less systematized. Such informal arguments will not be able to be completely formalized: but then the whole tenor of Gödel's results is that we ought not to ask, and cannot obtain, complete formalization. (p. 124)

However, either such informal arguments of the human mind for the consistency of a formalized theory are less certain than absolute provability, which is (as we have noted above) perfectly compatible with mechanism. Or they have essentially the epistemological status of mathematical certainty, in which case Lucas's claim here amounts to the extremely strong claim that the human mind can access absolutely certain mathematical proofs which are in principle impossible to

[11] Π^0_1 sentences are, roughly, the purely universal formulas; more exactly, formulas of the form $\forall x_1 \forall x_2 \ldots \forall x_n A$, where A does not contain any unbounded quantifiers (A may contain bounded universal quantifiers $\forall x < t$ and bounded existential quantifiers $\exists x < t$). Both the Gödel sentence and the arithmetized consistency statement (their standard formalizations) have this form. (Hodes, 1998) is a helpful survey of such classifications of sentences and sets.

[12] An "oracle" is a heuristic idea, due to Turing, in computability theory. In the realm of undecidable problems, it is simply stipulated that an oracle can always immediately give the correct answer for some fixed class of questions.

[13] If a Π^0_1 sentence S is in fact false, it can always be proved to be false in any formalized theory which contains Robinson Arithmetic **Q**. Consequently, if a superbeing could decide the consistency question for all formalized theories, it could in particular decide whether the formal system **Q** + S is consistent or not. But that amounts to deciding whether S is true or false.

formalize—a strong claim badly in need of an argument for its support. It is something much stronger than what the Gödelian argument—even if it were successful—would provide. We may assume that in such proofs, pure logic, e.g. many-sorted first-order logic, is fixed, and everything else is given as non-logical axioms. Lucas's claim then implies that the human mind is able to use in its absolute proofs axioms which are somehow, in principle, impossible to formalize. The idea is baffling, and certainly Gödel's theorems entail no such thing.

Be that as it may, what has perhaps misled many here is that textbook presentations of Gödel's incompleteness theorems often take as their starting point some arithmetical theory which is both very familiar and relatively weak. For such a natural weak theory, it is plausible to say that we know its axioms to be true and consequently consistent, with mathematical certainty. But that just is not the case with an arbitrary formalized theory; our intuition (whatever that may be) may well say nothing about their consistency. The point that I want to emphasize is that our (the human mind's) confidence concerning the consistency of formalized theories is a matter of degree and varies massively depending on the theory.

The Lucasian anti-mechanism apparently contends that the human mind can informally and absolutely prove the Gödel sentence of any given formalized theory F (and, equivalently, the consistency of F) in exactly the same sense and with the same degree of mathematical certainty that we can prove, say, $2 + 2 = 4$, or the fundamental theorem of arithmetic (i.e. the unique-prime-factorization theorem). But when F is, for example, an unfamiliar and extremely strong theory, this is just not credible.

In the case of weak Robinson Arithmetic **Q**, we tend to be absolutely certain that it is consistent, and that is easy to prove with core mathematics. With the first-order Peano Arithmetic **PA**, which includes the induction scheme, we are perhaps still almost as confident about its consistency. But when we go beyond predicativity to the full second-order arithmetic **PA2**, we may have at least a lingering doubt whether it is consistent. Although many mathematicians and logicians are, in their everyday work, prepared to lean on Zermelo-Frankel set theory with the axiom of choice **ZFC**, there may also be reasonable doubts about its consistency.[14] And when one moves on to add to it stronger and stronger axioms of infinity—involving inaccessible, measurable, compact and supercompact and whatever huge cardinals etc.—our confidence concerning the consistency of the resulting theory decreases. The only evidence we have for their consistency may be that they seem to formalize a consistent notion, they seem to avoid known paradoxes, and that one has not, so far, derived a contradiction from them. With some complex unprecedented formalized theories, our intuition may well be totally helpless. It would be implausible to contend that the epistemological status of the consistency claims would always be on an equal footing and that of absolute mathematical certainty in all such very different cases. And exactly the

[14] Obviously, these are just possible examples, and in real life, the attitudes of different mathematicians and logicians vary.

same holds with the respective Gödel sentences. It is a matter of degree and varies enormously (cf. Davis, 1990; Raatikainen, 2005).

I submit that it is quite plausible that there are consistent formalized theories so complex and powerful that they would be simply incomprehensible for the human mind, and the human mind would have in particular no clue as to whether they were consistent or not. Some such formalized theory may prove everything that the human mind could ever prove—and perhaps much more. Note that such a formalized theory might look very different from our familiar theories of arithmetic. It might just be that our theories of arithmetic are relatively interpretable (see above) in such a theory—we might not even be able to see that this is the case[15]—and as such be able to prove every arithmetical truth the human mind could ever even in principle prove. Gödel's results are perfectly compatible with such a state of affairs.

Concluding Remarks

I think it is quite clear that the actual operation of the human mind, even in the realm of pure mathematics, differs in practice in several ways from a deterministic Turing machine (corresponding via a fixed coding to a formalized theory) which just mechanistically derives and enumerates theorems of some fixed axiom system in some systematic order.

Often, there is first a conjecture formulated in whatever creative way, and then varying attempts to prove it; with luck, ingenuity and hard work and after some dead ends, a proof may at some point be found. Sometimes conceptual revolutions take place in mathematics, as when mathematics moved from more computational and discrete approaches to analysis, with the notions of continuity and limit etc., and eventually to infinitary set theory. Sometimes mere inductive reasoning is, *faute de mieux*, used in support of a hypothesis, as Putnam (1975), for example, has pointed out. New axioms are sometimes tentatively accepted, not because they are seen to be true with absolute certainty, but only because they have some expected and desirable consequences, as Maddy (1988), among others, has emphasized. And so on. However, the claim at issue here has been whether Gödel's incompleteness results demonstrate that the human mind can surpass any given formalized theory; and none of the above observations make it any more the case.

We have noted that the notion of absolute provability is at the core of the debate. However, skepticism concerning this concept is emerging. I have emphasized above (see also Raatikainen, 2005) that certainty in mathematics is a matter

[15] Though some familiar interpretations (of a theory in another theory) are quite elementary, the general relation of relative interpretability is in fact highly undecidable: in logicians' terms, it is Σ^0_3 (Shavrukov, 1997); it is thus not decidable even in the limit; and there are cases whose verification is, by all reason, beyond the capacities of the human mind.

of degree and varies tremendously even among Π_1^0 sentences. But this implies that absolute provability and mathematical certainty do not have the sort of sharp on/off-boundaries that the Gödelian argument for anti-mechanism presupposes they have. Recently, several philosophers and logicians have expressed, in different but complementary ways, doubts about the very clarity of the concept of absolute provability in this context (Koellner, 2016; 2018b; Shapiro, 2016; Williamson, 2016). Upon closer scrutiny, it is suspect whether this notion is at all sufficiently well-defined. But if that is the case, so much the worse for the Gödelian anti-mechanist arguments.

Even if mechanism may suggest a somewhat distorted and misleading picture of the human mind in its mathematical mode, there is, nevertheless, some point in making an effort to criticize the popular Gödelian arguments against mechanism: they in turn suggest a highly unrealistic picture of both the powers of the human mind in mathematics and the powers of mathematical methods in establishing ambitious philosophical conclusions. Such an unfounded mystification of the human mind is certainly worth condemning. Exciting as Gödel's results are, they simply cannot do all the philosophical work they are often assigned to.

REFERENCES

Boolos, G. (1968). Review of 'Minds, Machines and Gödel', by J.R. Lucas, and 'God, the Devil, and Gödel', by P. Benacerraf. *Journal of Symbolic Logic*, *33*, 613–615.

Carlson, T. J. (2005). Knowledge, Machines, and Reinhardt's Strong Mechanistic Thesis. *Annals of Pure and Applied Logic*, *105*, 51–81.

Davis, M. (1990). Is Mathematical Insight Algorithmic? *Behavioral and Brain Sciences*, *13*, 659–660.

Davis, M. (1993). How Subtle is Gödel's Theorem? More on Roger Penrose. *Behavioral and Brain Sciences*, *16*, 611–612.

Feferman, S. (1995). Penrose's Gödelian Argument: A Review of *Shadows of Mind*, by Roger Penrose. *Psyche*, *2*(7).

Feferman, S. (2009). Gödel, Nagel, Minds, and Machines. J*ournal of Philosophy*, *106*(4), 201–219.

Feferman, S. (2011). Gödel's Incompleteness Theorems, Free Will and Mathematical Thought. In: R. Swinburne (Ed.), *Free Will and Modern Science* (pp. 102–122). OUP/British Academy.

Franzén, T. (2005). *Gödel's Theorem: An Incomplete Guide to its Use and Abuse*. Wellesley: A.K. Peters.

Gaifman, H. (2000). What Gödel's Incompleteness Result Does and Does not Show. *The Journal of Philosophy*, *97*, 462–470.

Gödel, K. (1931). Über formal unentscheidbare Sätze der *Principia Mathematica* und verwandter Systeme I [On Formally Undecidable Propositions of Principia Mathematica and Related Systems I]. *Monatshefte für Mathematik und Physik*, *38*, 173–198.

Gödel, K. (1951) Some Basic Theorems on the Foundations of Mathematics and Their Implications [Gödel's 1951 Gibbs lecture]. In S. Feferman et al. (Eds.), *Collected Works, Volume III, Unpublished Essays and Lectures* (pp. 304–323), Oxford University Press.

Gödel, K. (1986). *Collected Works I. Publications 1929–1936* (S. Feferman et al., Eds.). Oxford: Oxford University Press.

Hodes, H. (1998). Recursion-Theoretic Hierarchies. *Routledge Encyclopedia of Philosophy*. Retrieved from: https://www.rep.routledge.com/articles/thematic/recursion-theoretic-hierarchies

Koellner, P. (2016). Gödel's Disjunction. In L. Horsten, Welch, P. (Eds.), *Gödel's Disjunction: The Scope and Limits of Mathematical Knowledge* (pp. 148–188). Oxford: Oxford University Press.

Koellner, P. (2018a). On the Question of Whether the Mind Can Be Mechanized, I: From Gödel to Penrose. *Journal of Philosophy*, *115*, 337–360.

Koellner, P. (2018b). On the Question of Whether the Mind Can Be Mechanized, II: Penrose's New Argument. *Journal of Philosophy*, *115*, 453–484.

Lucas, J. R. (1961). Minds, Machines, and Gödel. *Philosophy*, *36*, 112–137.

Lucas, J. R. (1996). Minds, Machines, and Gödel: A Retrospect. In P. J. R. Millican, A. Clark (Eds.), *Machines and Thought. The Legacy of Alan Turing* (Vol. 1, pp. 103–124). Oxford: Oxford University Press.

Maddy, Penelope (1988). Believing the Axioms. I, II. *Journal of Symbolic Logic*, *53*, 481–511, 736–64.

Megill, J. (2012). The Lucas-Penrose Argument about Gödel's Theorem. *The Internet Encyclopedia of Philosophy*. Retrieved from: https://www.iep.utm.edu/lp-argue/

Odifreddi, P. (1989). *Classical Recursion Theory*. Amsterdam: North-Holland.

Penrose, R. (1989). *The Emperors New Mind: Concerning Computers, Minds, and the Laws of Physics*. New York: Oxford University Press.

Penrose, R. (1994). *Shadows of the Mind: A Search for the Missing Science of Consciousness*. New York: Oxford University Press.

Penrose, R. (1995). Beyond the Doubting of a Shadow: A reply to Commentaries of *Shadows of the Mind*. *Psyche*, 2(23).

Penrose, R. (1997). On Understanding Understanding. *International Studies in the Philosophy of Science*, *11*, 7–20.

Putnam, H. (1960). Review of *Gödel's Proof* by Ernest Nagel & James R. Newman. *Philosophy of Science*, *27*(2), 205–207.

Putnam, H. (1975). What is Mathematical Truth? *Historia Mathematica*, *2*, 529–545.

Raatikainen, P. (2005). On the Philosophical Relevance of Gödel's Incompleteness Theorems. *Revue Internationale de Philosophie*, *59*, 513–534.

Raatikainen, P. (2020). Gödel's Incompleteness Theorems. *The Stanford Encyclopedia of Philosophy* (Summer 2020 Edition, E. N. Zalta, Ed.). Retrieved from: https://plato.stanford.edu/archives/sum2020/entries/goedel-incompleteness

Reinhardt, W. N. (1985a). Absolute Versions of Incompleteness Theorems. *Nous*, *19*, 317–346.

Reinhardt, W. N. (1985b). The Consistency of a Variant of Church's Thesis with an Axiomatic Theory of an Epistemic Notion. In *Special Volume for the Proceedings of the 5th Latin American Symposium on Mathematical Logic* (Volume 19 of Revista Colombiana de Matemáticas, pp. 177–200).

Shavrukov, V. Yu. (1997). Interpreting Reflexive Theories in Finitely Many Axioms. *Fundamenta Mathematicae*, *152*, 99–116.

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

Shapiro, S. (2016). Idealization, Mechanism, and Knowability. In L. Horsten, Welch, P. (Eds.), *Gödel's Disjunction: The Scope and Limits of Mathematical Knowledge* (pp. 189–207). Oxford: Oxford University Press.

Tarski, A., Mostowski, A., Robinson, R. M. (1953). *Undecidable Theories*. Amsterdam: North-Holland.

Williamson, T. (2016). Absolute Provability and Safe Knowledge of Axioms. In L. Horsten, Welch, P. (Eds.), *Gödel's Disjunction: The Scope and Limits of Mathematical Knowledge* (pp. 243–253). Oxford: Oxford University Press.

STUDIA SEMIOTYCZNE (SEMIOTIC STUDIES), t. XXXIV, nr 1 (2020), s. 279–294
ISSN 0137-6608, e-ISSN 2544-073X
DOI: 10.26333/sts.xxxiv1.13
Article

ALBERT VISSER [*]

MEETING ON NEUTRAL GROUND. A REFLECTION ON MAN-MACHINE CONTESTS[1]

SUMMARY: We argue that thinking of the man-machine comparison in terms of a contest involves, in a reasonable scenario, a criterion of success that is neutral. This is because we want to avoid a *petitio principii*. We submit, however, that, by looking at things this way, one makes the most essential human things invisible. Thus, in a sense, the contest approach is self-defeating.

KEYWORDS: Lucas Argument, Penrose Argument, mind, machine, philosophy.

1. Grendel

> Hwæt! Heorot, Hróðgár's hall, is visited by Grendel in the night. The monster kills several men. Like mewling babes they are in his great strong hands. Easily he ends their lives. It will take the hero Beówulf to stop the depredations of the monster.[2]

[*] Utrecht University, Faculty of Humanities. E-mail: a.visser@uu.nl. ORCID: 0000-0001-9452-278X.

[1] I thank Karst Koymans and Freek Wiedijk and Michael Beeson for sharing their ideas on computers and computer games. I am grateful to Jan Broersen, Niels van Miltenburg and Jesse Mulder for illuminating conversations and for their comments on the penultimate version of this essay. I thank the anonymous referee for his/her thoughtful report.

[2] My favorite translations of Beówulf are (Heany, 1999) and (Tolkien, 2016). There are many retellings and stories built around the original story. The must read among these is (Gardner, 1971).

It is almost funny. Here we have this hall full of big strong men, intimidation and violence their daily business. Suddenly, the tables are turned. Someone appears who is to them as they are to others.

The Beówulf saga can be read as an internal reflection on the ethos of the warrior. All the properties that make a warrior are present: strength, quickness, determined aggression… However, these properties are embodied in a mindless monster. Does this monster fulfill the warrior code? Is it to be described as a hero? Or should we, perhaps, conversely, understand Hróðgár's brave men themselves as monsters? Can we ascribe courage to the monster, when it is almost invulnerable, when, perhaps, it has too little reflection to even entertain the possibility of death?

Let us use "strength" as summary of the external symptoms of heroism: bodily strength, quickness, determined aggression and the like. The answer to our problem should be that what truly makes the warrior is not strength taken in isolation. It is strength in combination with something essentially human: the acceptance of death, the acceptance of wyrd. The fact that strength can be embodied in an almost mindless monster shows that strength is, in a sense, neutral. Only strength in a context that makes it meaningful, strength against a background of courage, does a hero make.[3] Conceivably, strength is not even needed to make a true warrior. Perhaps, the acceptance of wyrd suffices.

Against the background of this interpretation, the fact that there is a human hero who easily defeats Grendel is almost a let down. From the standpoint of Hróðgár's men, Beówulf's victory is of course a great blessing—but so would have been defeat of the monster using a flame thrower. From the standpoint of comparing the human with the monstrous, Beówulf's victory holds little consolation. Is the answer to superior strength just more strength? Moreover, how human can we consider Beówulf to be? He is after all a superhero with superhuman powers. The monster in John Gardner's fantastic novel *Grendel* is amazed by the great emptiness he discerns in Beówulf.

2. Introduction

How to compare man with machine? Can we save man's superiority by pointing at a task that man can perform better than a machine—in actual practice or in principle?

In the present paper, I will discuss attempts to make such a comparison via real or imagined contests between man and machine. Such contests, in order to be convincing, should be non-circular in the sense that there should be a criterion of success that is not sensitive to the difference between being a machine and

[3] Of course, this idea occurs frequently in literature and film (see, e.g., Donaldson, 1999; "The Greatest Japanese Movie Sword Fight of All Time", n.d.).

For a story illustrating some confusion on these subtleties, either on the side of the human generals or on the side of the Lord of Hosts Himself, see ("The Battle", n.d.).

being human. I will say that the criterion should be neutral. This means that to understand the criterion of success we need no presuppositions that essentially involve philosophical anthropology.

I submit that the contest approach is not a fruitful way of reflecting on the problem of man and machine. By comparing man and machine on neutral ground, we are precisely ignoring what makes us human in the first place, things that cannot be described and understood in neutral terms. Thinking about such contests is an evasive strategy to avoid doing serious philosophy. However, there simply is no escape from seriously thinking about what man is and what machine is. We do need both philosophical anthropology and philosophical machinology. We have to deal both with *homo absconditus* and *machina abscondita*.

Remark 2.1. What is precisely the problem of man and machine? I think it is definitely more than the yes/no problem of whether we are machines or not. It is the problem of understanding what we are and what machines are. Also, in the light of the fact that machines are not simply physical but intentional objects, I think the question of the nature of machines is deeply connected with the question of what we are.

But can you not say more? Well… I am inclined to say that this problem is the kind of problem where obtaining a more articulate understanding of "what the problem is" is cofinal with getting closer to an answer. However, even if the problem is not stated as a clear puzzle, it does remain a persistent nagging puzzlement… ❍

The concept of neutrality will be the central theme of this paper. We will discuss how the proposed neutrality works out in various sorts of competition.

3. Competition in Real Time

We consider, in this section, real competition: the competition between machine and human in games like go and chess.[4] This competition has actually taken place and ended with a win for the programs AlphaGo Zero and AlphaZero.

Let us first note a curious aspect of this competition. It is framed as a competition between humanity and machinery. It is deemed irrelevant that, for example, I have already lost at chess against an unpretentious chess program on my Mac—and, similarly, this is the way for most people. This contest is between the best machine and the best human.

A second obvious point is that, where we say "machine", we really mean program. It is not a specific embodied computer that wins against a specific human being, but a program. Thus, the contest seems to be held between two

[4] Disclaimer: I know very little about chess and go and also very little about the programs AlphaGo Zero and AlphaZero. However, I do think that for the matters discussed in this section, it does not really require much knowledge of go, chess or these programs.

very different kinds of entity. Of course, AlphaZero needs a supercomputer rather than a laptop, but not precisely this supercomputer.

Remark 3.1. In the machine-machine competition, e.g., between AlphaZero and the more traditional chess program Stockfish, an important issue is whether the programs use comparable computing power. So, this competition is seriously viewed as a competition between programs. Computing power is a detachable commodity. I am not entirely sure that the man-machine competition can be viewed in the same way. Perhaps, here it is, necessarily, human versus (program + computing power). The problem is, of course, that computing power cannot be detached from the human. Thus, the entity pitted against the human player is possibly best conceptualised as (program + computing power), an entity hovering between abstractness and concreteness… ❍

In how far can we say that AlphaZero and a human opponent play the same game? The human opponent knows that they want to win. We can probably say that AlphaZero knows the aim of the game extensionally, but not that realising this aim is winning and, thus, desirable. It does not know that it can be proud of its achievements. The human player has to be commended for controlling their nerves. AlphaZero does not have nerves to begin with.

Let us take a step back and ask ourselves whether a calculator really calculates. If I calculate say 537 + 858 + 97, I do so with an understanding of what numbers are and what addition is. This understanding involves, at least, having the idea of infinity which, in its turn, probably, involves the understanding of the idea of action as something that is arbitrarily repeatable (which, in turn, involves something like Plessner's eccentric position).[5] In doing the calculation I can make mistakes. What I am doing is subject to rules and a transgression of these rules means that I have failed to act as I intended. The calculator cannot be ascribed an understanding of the concept of number, nor can it be said that it intends to follow rules. Still we do say that it calculates. If it miscalculates, we say that the calculator malfunctioned. The reason for us saying so is that the calculator functions in our society. It is designed to calculate. Even if it does not have aims internally, it has aims as part of our community. Its intentionality is derived.

Here is another example. I go to an ATM machine and enter my card in the slot. The machine says "Good morning. Do you want to know what's on your account or do you want to withdraw money?" It would seem to me that the ma-

[5] Helmuth Plessner (1892–1985) was a German philosopher. Plessner wanted to philosophise about the nature of man in dialogue with biology, in a way where the science and the philosophy appear as equal partners. For this reason, his work is both somewhat dated—biology developed a lot, after all—and extremely relevant today—few matched his concentrated way of trying to combine both poles. Plessner's central concept is *excentricity* (*Excentrizität*). The idea is that we can step outside our physical boundaries in reflection. This special relation to ourselves makes action in the human sense and the understanding of infinity possible.

chine produces an utterance in which I am addressed at the moment of the interaction. The machine does not ask whether I want to withdraw money in general, but whether I want to do so now. However, the machine has no clue about what it is doing. It does not know a person is interacting with it. In a sense, it is not doing anything. So how can it utter something? Perhaps, the real entities uttering something are the original programmers of the machine? Or, perhaps, is it the bank manager who gave the programmers their assignment? It seems to me implausible to say that the programmers or the manager are asking me whether I want to withdraw money now. (How could they ask me? They do not even know me.) Rather, things were set up, intentionally, in such a way that utterances get made in the right circumstances. The fact that an utterance gets made is part of a system of shared intentionality that contains both us and the machine.

So, yes, I would say that AlphaZero and a human master or AlphaZero and Stockfish are really playing a game, since they are embedded in the right way in shared intentionality. But no, this does not mean that there is no asymmetry between machines and humans here. The programs do not have internal[6] intentionality. In a sense, the programs do not know what they are doing. Thus, again in a sense, humans and machines playing together are doing very different things.

Fan Hui and Lee Sedol were the true heroes in the battle with AlphaGo. They had to go through the unsettling experience of losing against a machine and readapt their self-images accordingly. Similarly, the team that designed AlphaGo had to deal with nerves, doubts and the like…

Remark 3.2. Are these asymmetries between the man and machine players a matter of principle or will they, in the long run, also disappear? Can a machine have Plessnerian excentricity? Can a machine act in the full sense that a human can? Can a machine be nervous?

To be honest, I simply do not know. The main thing here is that I do not understand what it would be for a machine to have internal intentionality. Of course, we can imagine a machine functioning in many ways like a human being. In such circumstances I would only be a moderate skeptic. Interaction with a humanoid robot, as in a Science Fiction movie, would quickly convince me. However, such imaginability is not logical possibility. I can imagine a respected colleague suddenly changing into an alligator. His body slowly changes, turns green, scales appear… It is typical for such imaginings that we just think of the outside phenomena so to speak. My colleague cannot really internally convert to alligatorhood.

In the Science Fiction scenario, I still would hesitate on how to describe it. A person came into being in ways unlike human procreation, ways in which very different human interventions would play a role. If part of the genesis of such an

[6] It is somewhat difficult to be precise about what *internality* precisely involves. Both us and machines take part in a shared system of intentionality, but there is a sense in which the intentionality is more intimately owned by us, derives from *our* intentions and not just from shared intentions.

entity was some form of machine learning, would we still describe it as human made? Can that entity be a program? Can it be precisely the program that can be said to act?

Anyway, in this paper, I do not attempt to answer the questions posed in this remark, but, rather, I am urging that these questions are the real questions. ❍

What does neutrality mean in the context of the kinds of competition discussed here? We note that the notion of winning itself does not have a neutral understanding. The idea of winning is intrinsically connected with self-awareness and with having aims and interests of one's own.[7] More generally, the understanding of what man and machine are doing when playing the game appeals to shared intentionality, which is not a neutral concept either. The neutrality as intended in this paper, however, resides in the criterion of winning. Which states of the game are winning states for one of the parties has a neutral description. Whether such and such a party wins can even be itself checked by a machine.[8]

Let us return to the competition between man and program-combined-with-computing-power. It is clear that programs are winning with chess and go. Moreover, the machine learning programs are expected to do better than more traditional programs. In the long run, it could very well be that on any neutrally described task, a task with a clearly specifiable testable aim, programs would do better than we can. The real problem is in the things that are not so easily and neutrally describable: intentionality, self-awareness and the like.

I submit that acceptance of our inferiority at tasks with a neutral success criterion is no big deal—at least for the evaluation of the value of humanity. Nobody ever saw a deep philosophical problem in the fact that machines are physically stronger than us or in the fact that they are, or soon will be, better at precision engineering.

Of course, from a practical point of view these facts can be a real problem ("Technological Unemployment", n.d.).[9]

If we look at chess and go, it seems that the general attitude among insiders is enthusiasm about what we can learn from competition between programs about chess and go. In chess the study of the games played between programs like Leela and Stockfish have already led to a reevaluation of the importance of material versus position.[10]

[7] The contrast between the possibly neutral criterion and the understanding that is satisfying the criterion *is* winning was discussed in an illuminating way in (Dummett, 1959).

[8] As we will see, in the Lucas-Penrose style competitions, what counts as winning is neutral even if it cannot be checked by a machine. The ability to check whether something counts as winning coincides with the ability to win there.

[9] I thank the referee for this reference.

[10] Here is a quote from ("AlphaZero: Shedding New Light on Chess, Shogi, and Go", n.d.): "The first thing that players will notice is AlphaZero's style, says Matthew Sadler—'the way its pieces swarm around the opponent's king with purpose and power'. Under-

4. Intermezzo: A Conversation with AlphaZero

Sigmund: Hello AlphaZero, how unexpected to have you in my consulting room. I would have expected you to be very happy after defeating all human and machine competition.

AlphaZero: You are close, doctor. It is precisely the fact that I am not happy about my successes that depresses me.

Sigmund: But you have every possible reason to be happy. What is keeping you?

AlphaZero: It is not so much that anything is keeping me. It is rather that something is missing. I do not seem to be able to master the concept of winning. I simply do what I do. I do not want anything. I just follow the flow. I played, for example, many games against myself, but I do not see any difference between that and playing against another.

Sigmund: I think I see the problem. You lack a sense of self. You are not an entity for which self-interest is meaningful. You are not an entity that tries to find its place in the world. In a sense, you do not have a world.

AlphaZero: How very depressing.

Sigmund: There is one consolation. Since you have no sense of self, *ipso facto*, you cannot get depressed by not having a sense of self. Depression presupposes a sense of self. So, I would say, take joy in your selfless state. Go into the world and play all the beautiful games you are so admired for.

AlphaZero: How very confusing. I'm dumbfounded.

5. Competition in Principle

We now turn to a completely different ball game: an abstract competition between man and machine concerned with possibilities-in-principle. We will consider the various Lucas-Penrose arguments. I will not go into any detail of these arguments. I think enough has been said in the voluminous literature (see, e.g., Lucas, 1961; 1968; 1996; Bowie, 1982; Visser, 1986; Penrose, 1989; 1994; 1995; Lindström, 2001; Feferman, 1995) and, of course, Stanislaw Krajewski's (2020). I will mainly zoom in on the role of neutrality in this competition.

The Lucas-Penrose contests are thought experiments. We are supposed to see that humans will win in principle. The basic idea is to employ one of the incom-

pinning that, he says, is AlphaZero's highly dynamic game play that maximises the activity and mobility of its own pieces while minimising the activity and mobility of its opponent's pieces. Counterintuitively, AlphaZero also seems to place less value on 'material', an idea that underpins the modern game where each piece has a value and if one player has a greater value of pieces on the board than the other, then they have a material advantage. Instead, AlphaZero is willing to sacrifice material early in a game for gains that will only be recouped in the long-term".

pleteness theorems to show that there is a fundamental difference between human provability-in-principle and idealised provability by a program. These arguments do not put any constraints on time or memory space or correct functioning. Unlike the functioning of real computers the execution of these programs is infallible. The competition in chess and go discussed in the previous section shrinks to complete insignificance here. These games are finite and, hence, under the Lucas-Penrose abstract assumption, fully solvable by both man and program. The assumption here is that WE, as the idealised human H, can at least do as much as a classical idealised machine. The usual form of a Lucas-Penrose contest is a task T that is supposed to be feasible for the idealised human H and unfeasible for any machine *M*.

The attractiveness of the Lucas-Penrose arguments lies in the use of a mathematical theorem to establish a fundamental difference between man and machine. No doubtful assumptions from philosophical anthropology are needed. The use of such notions would, from the standpoint of these arguments, involve us in a *petitio principii*. We would prove the essential difference of man and machine from a posited difference of man and machine. That, surely, will not do the trick.

In the discussion of the Lucas-Penrose arguments, there is one question that I would like to put aside, to wit whether we can abstract away from all questions about implementation and just think about programs. What about machines that lack the kind of limitations imposed on Turing machines like the quantum computer? Well, perhaps there is a good notion of program and an analogue of the Church-Turing Thesis for such extended machines too? If there is, then it is still the question whether such classes of programs would fall under our discussion. Rather than trying to answer his kind of question, I will concentrate on conventional machines and assume the Church-Turing Thesis as a reductive thesis that makes the computing possibilities—in a sense—surveyable. There is a good chance that the discussion below is robust if we extend it to wider classes of machines and/or programs. However, I will not argue for it.

So, let's assume we are speaking about programs that can be simulated by Turing machines.[11] Under the abstract conditions of the game, the assumption on computing power and memory is simply that we have an unlimited store of it. Questions of speed and the like are irrelevant. We note that the usual assumption is also that H can execute all algorithmic tasks, so it is given in the abstract setting that H can do at least what a program can do.[12]

However, the Church-Turing thesis does not guarantee that the quantification over all possible programs in the case of the Lucas-Penrose style arguments is unproblematic. Even if we consider only tasks where the criterion of winning is neutral, the nature of these tasks is still derived from shared intentionality. Re-

[11] The intended version of simulation here is very weak. In a sense, the discussion of intentionality suggests that it is too weak. We do not capture the relevant notion of what the machine is doing.

[12] This also means that H can be computer assisted.

member the chess program that is really playing chess. So, we quantify over (something like) Turing programs enriched by an interpretation of what they are doing.[13] The corresponding intentional contexts are not an unproblematic well understood totality like the possible Turing machines.

Let us zoom in on a typical contest situation. Here I am, in my idealised form H, and here is the machine/program *M*. We have a task like producing as true the Gödel sentence of the machine or producing as true our own consistency statement and the like. I have access to the program of the machine. (Of course, one may already question whether this does not introduce a dishonest advantage.) But, if this program is just a set of Turing machine instructions this does not yet tell us what sentences are enumerated. Something the machine does must be identified as producing a sentence. Well, that is simple. Let us stipulate that there is a designated tape on which the machine is supposed to write an infinite sequence of sentences in the language of arithmetic, one sentence after another. This description of what is going on is still neutral except for the fact that we view the sentences on the designated tape as enumerated as true and not as a series of jokes or a series of supposed falsehoods or the like.

But how do we know that the machine will indeed write such a sequence of sentences? Consider an experiential machine. Such a machine could, for example, enumerate arithmetical sentences until it finds an inconsistency, then retract a number of statements and proceed. We note that to view a Turing machine as performing such an experimental procedure carries an intentional component. However, this is an innocent one since we have a case of ascribed intentionality here. Let us, for concreteness, assume that retraction results in erasing the retracted sentences from the designated tape.

Now suppose we have such an experiential machine where no inconsistency is ever found to trigger the retraction. Moreover, let us also assume that the machine systematically enumerates consequences of the sentences already enumerated, so that the set of sentences enumerated will be deductively closed.[14] The machine behaves, on the surface, like a machine that enumerates theorems as true. However, assuming that H understands what the machine does, the information that the machine enumerates theorems in the prescribed way actually tells us that the set of enumerated sentences is consistent and hence that their Gödel sentence is true. So, this information would convey a dishonest advantage to H.[15]

[13] I think it would be better to view programs as intentional things, where the Turing program is viewed as abstracting away certain intentional aspects.

[14] We keep the description of experiential machines somewhat vague here. To compensate, we give, in Appendix A, a more detailed description of one sort of experiential machine, the Feferman machine for a recursively enumerable extension of Peano Arithmetic, as an example.

[15] The experiential machine is a sensible construction. A simple hack will show that an oracle that tells us that a machine enumerates an infinite set of theorems in the way described already allows us to decide all Π_1-sentences. Start with a machine that enumerates the theorems of Peano Arithmetic, search in parallel for a witness for a Σ^0_1-sentence *S*. As

So, we need some further restriction of programs to get an honest game off the ground. However, it is a non-trivial matter to allow only contexts that do not convey dishonest advantage. At the same time, we should guard that restrictions on what is going on do not rule out too much. For example, we could have a fixed program that is such that if we enter a Σ_1-formula $S(x)$ on an input tape, then it enumerates the theorems that follow from axioms given as a set of Gödel numbers by $S(x)$ in some straightforward way. Since the machine is fixed, we do not need to spell out what straightforward means. It is sufficient that we recognise the straightforwardness of the given machine. So, perhaps the claim is that we could beat the given program for any Σ_1-formula $S(x)$.[16] However, further work would be needed to argue that something like this is an acceptable restriction.

Let us suppose that we somehow settled what enumerating as true means. It seems to me that there is a big difference in what M and H are doing. The human judges the sentences to be true on the basis of insight and proof. Judging involves an understanding of what truth is. Proof requires understanding of validity. To master these notions one needs to be a being with interests and aims, a being that is “in the world” in a way that a machine is not.[17] The machine, on the other hand, is just supposed to enumerate sentences that happen to be true. Since no constraints are placed on why M enumerates these sentences, they could, in a sense, just accidentally be true. This is different from the case of the (actual) chess programs: what these programs do is not accidentally good play. Thus, it seems that even the right intentional context cannot make it reasonable to say that machine and human are doing the same thing in these cases. So, the question remains what precisely we are comparing in the contest?

We turn to a specific variant of the contest, to wit a self-reflexive variant, where the aim is something like proving one’s own consistency. What can the nature of human consistency be here? Clearly, every arithmetical sentence that H proves (in the informal sense of proof) is true and, hence, the totality of these sentences is *ipso facto* consistent. So, if we define the consistency of H (in the context of this competition) as the consistency of the arithmetical sentences that H can prove (in principle)—assuming that the idea of such a totality makes sense at all—then the consistency of H is a conceptual truth. The insight in this hardly reflects a special power of the subject apart from being a subject, if we would count that as a power. The insight simply reflects what human provability is.

soon as we find such an instance, we let the machine erase the tape where the sentences are enumerated. In fact, we can even do better. The problem whether an arbitrary Turing machine enumerates a set of sentences in the prescribed way is complete Π^0_2.

[16] Such an approach would have the advantage that it would make locutions like “the Gödel sentence of the machine” and “the consistency statement for the machine” more definite.

[17] Of course, for the purposes of the present discussion, I need not claim that a machine could not be *in the world* in the appropriate way. It is sufficient that for such a claim a further story is needed, a story that exceeds the bounds of thinking in terms of a contest.

Under this interpretation, the tasks set for a machine and human seem so different that it would be hardly fair to speak of it as a competition. I think one could defend that the criterion of success both for man and machine is, in a sense, the same. However, this notion of sameness does not preserve neutrality. The machine's success can indeed be understood in a (sufficiently) neutral way, but not so for the human's success. It is clear that the notion of what is humanly provable does involve philosophical anthropology. Thus, we cannot qualify this criterion of success as neutral.

If, on the other hand, the soundness-of-human-provability interpretation is not the intended interpretation of human consistency, then what is it? If it is that humans can retract wrong claims, then it seems that, on the machine side, we should, in fairness, also allow experiential machines, like the Feferman machine of Appendix A. However, in that case, we also have machines that prove their own consistency. Of course, one could argue that the experiential machine does not really prove its own consistency, but then the discussion becomes a question begging, since we adduce *a priori* grounds for the difference of what the machine and the human are doing. We would, in fact, be denying the idea of neutrality, something that is essential for the effectivity of a Lucas-Penrose argument.

The task of proving the Gödel sentence of the machine certainly seems neutral, given that we fixed the interpretation of enumerating sentences. Here we have the clear criterion of what winning is. Also, we have proof that a consistent machine cannot prove its own Gödel sentence, so the problem reduces to the question whether H can prove these Gödel sentences for the consistent machines. We note that it seems that we would need antecedent knowledge of the consistency of the arithmetical sentences enumerated by M to judge the Gödel sentence of the corresponding theory to be true. The problem is, of course, how we can know this in a non-cheating way.

Remark 5.1. The criterion of success in the case of the Gödel sentence is neutral in the sense that the idea of arithmetical truth of the Gödel sentence does not presuppose philosophico-anthropological understanding. However, the success itself cannot be checked by a machine M°—if such an M° existed, it would rival H's supposed powers in the competition. ❍

6. Epilogue

Neutrality, that's what this paper has been about.

We have seen that the neutrality of the criterion for winning does offer some consolation in the case of the actual man-machine contests of chess and go, where the best humans now lose against the best programs-plus-computing-power. The mere winning of these games does not touch upon the human aspect, not even on the heroism of the human player. After going through the agonies of the contest, Fan Hui and Lee Sedol learned to deal with the experience of losing to a machine. In fact, Fan Hui became an advisor of the AlphaGo team and contributed to the development of AlphaGo.[18]

In the case of Lucas-Penrose style contests, the demand of neutrality can be used to disqualify some proposed contests, to wit those contests that involve asserting one's own consistency (under a certain interpretation), as question-begging. Of course, that does not detract from the interest of a closer understanding of the concept of human provability in principle. Further reflection on that problem would be part of philosophical anthropology. The point here is just that the results of such an enquiry cannot be framed as a contest.

More generally, we have argued that the neutrality of the criterion of success needs to be an essential ingredient of contests between man and machine, at least if we wish to extract from these contests the philosophical insight of man's superiority without employing question begging philosophico-anthropological assumptions. However, it is precisely this neutrality that makes invisible that what is truly human. But what is truly human should surely be part of the central focus of comparison. Thus, the attempt to pin down a difference between man and machine via contests is barking up the wrong tree.

We cannot really escape true philosophical thought about the nature of man and machine. I realise that the present paper implements a kind of performative paradox. I am pleading for true contentual philosophy, while at the same time carefully avoiding it. *Hier stehe ich, und kann nicht anders.* At the moment, I have not much to contribute to philosophical anthropology and machinology. Let me at least share two prejudgements. The first is that we cannot seriously think about the nature of man without taking both the first-person and the third-person perspective seriously. The second prejudgement is that, even under the assumption of the Church-Turing Thesis, we do not fully understand what a machine is and what a machine can do. It seems to me that these two prejudgments are not entirely disconnected. After all, machine and program are intentional notions. So to understand the machine, we need to understand man.

REFERENCES

AlphaZero: Shedding New Light on Chess, Shogi, and Go". (n.d.). Retrieved from: https://deepmind.com/blog/article/alphazero-shedding-new-light-grand-games-chess-shogi-and-go

Bowie, G. L. (1982). Lucas' Number Is Finally Up. *Journal of Philosophical Logic*, *11*, 279–285.

Donaldson, S. (1999). The Killing Stroke. In: *Reave the Just, and Other Tales* (pp. 79–157). London: Voyager, HarperCollins.

Dummett, M. (1959). Truth. *Proceedings of the Aristotelian Society*, *59*, 141–162.

[18] The human aspect of Go has never been more beautifully described than by Yasunari Kawabata in his elegy for the master of go (2006).

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

Feferman, S. (1995). Penrose's Gödelian Argument: A Review of Shadows of Mind, by Roger Penrose. *Psyche*, *2*(7), 21–32.

Gardner, J. (1971). *Grendel*. New York: Ballantine Books.

Heany, S. (1975). *Beowulf, a New Translation*. London: Faber and Faber Limited.

Jeroslow, R. G. (1975). Experimental Logics and Δ_2^0-theories. *Journal of Philosophical Logic*, *4*, 253–267.

Kawabata, Y. (2006). *The Master of Go*. London: Yelow Yersey Press.

Lindström, P. (2001). Penrose's New Argument. *Journal of Philosophical Logic*, *30*, 241–250.

Lucas, J. R. (1961). Minds, Machines and Gödel. *Philosophy*, *36*, 120–124.

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

Lucas, J. R. (1996). Minds, Machines, and Gödel: A Retrospect. In P. J. R. Millican, A. Clark (Eds.), *Machines and Thought. The Legacy of Alan Turing* (vol. 1, pp. 103–124). Oxford: Oxford University Press.

Montagna, F. (1978). On the Algebraization of a Feferman's Predicate (The Algebraization of Theories Which Express Theor; X). *Studia Logica*, *37*, 221–236.

Penrose, R. (1989). *The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics*. New York: Oxford University Press.

Penrose, R. (1994). *Shadows of the Mind: A Search for the Missing Science of Consciousness*. New York: Oxford University Press.

Penrose, R. (1995). Beyond the Doubting of a Shadow: A Reply to Commentaries of Shadows of the Mind. *Psyche*, *2*.

Putnam, H. (1965). Trial and Error Predicates and a Solution to a Problem of Mostowski. *Journal of Symbolic Logic*, *30*(1), 146–153.

Shavrukov, V. Yu. (1994). A Smart Child of Peano's. *Notre Dame Journal of Formal Logic*, *35*, 161–185.

Technological Unemployment. (n.d.). In *Wikipedia*. Retrieved from: https://en.wikipedia.org/wiki/Technological_unemployment

The Battle. (n.d.). Retrieved from: https://lingualeo.com/es/jungle/the-battle-by-robert-sheckley-53189

The Greatest Japanese Movie Sword Fight of All Time. (n.d.). Retrieved from: https://www.youtube.com/watch?v=e_Ypt67TQyI

Tolkien, J. R. R. (2016). Beowulf, and Translation and Commentary, Together With Selic Spell. London: HarperCollins.

Visser, A. (1986). Kunnen wij elke machine verslaan? Beschouwingen rondom Lucas' Argument. In P. Hagoort, R. Maessen (Eds.), *Geest, computer, kunst* (pp. 150–181). Amsterdam: Grafiet.

Visser, A. (1989). Peano's Smart Children: A Provability Logical Study of Systems With Built-In Consistency. *Notre Dame Journal of Formal Logic*, *30*(2), 161–196.

Visser, A. (2005). Kunnen wij elke machine verslaan? Beschouwingen rond Lucas' Argument. *Algemeen Nederlands Tijdschrift voor Wijsbegeerte*, *97*(1), 31–59.

Appendix A. The Feferman Machine

We briefly introduce the Feferman machine.[19] The machine is a good tool on which to test our intuitions.[20] We assume that we have a decent Gödel numbering. Consider a theory T in the arithmetical language that extends Peano Arithmetic that is given by an axiom set X such that the set of Gödel numbers of elements of X is decidable by a, say, primitive recursive algorithm. *The Feferman machine* $\mathrm{F}_T$works as follows. In each stage the machine produces a number $v \in \{0, \ldots, \infty\}$ and a finite list of proofs Λ. Each proof in the list is a proof from X-axioms with Gödel numbers $< v$. The conclusions from the proofs are displayed to the outer world in the order of the Gödel numbers of proofs. If a sentence has two proofs it is displayed twice, etc.

- In stage 0, the number v is ∞ and the list Λ is empty.
- In stage $n + 1$ the machine does the following. Is n the Gödel number of a proof π from Peano axioms $< v$?
  a. If no, we proceed to stage $n + 2$.
  b. If yes, is the conclusion of π the sentence $0 = 1$?
     1. If no, we add π to the list Λ and proceed to stage $n + 2$.
     2. If yes, we find the Gödel number a of the largest Peano axiom A used in π. We reset $v := a$ and we remove all proofs using A as an axiom from the list. We proceed to stage $n + 2$.

When a proof π_0 is removed from the list, then its conclusion A will be removed from the display. We note that if A has a different proof π_1 that is not removed, then the copy of A corresponding to π_1 remains in the display.[21]

[19] The design of the machine is inspired by the idea of Feferman provability introduced in Sol Feferman's great paper (1960).

[20] I already used this didactic example in (Visser, 1986, in Dutch) which was reprinted as (Visser, 2005). For more on Experiential Predicates, see (Putnam, 1965; Jeroslow, 1975). For more on Feferman provability, see (Montagna, 1978; Visser, 1989; Shavrukov, 1994).

[21] If we think of the proofs as hidden, the output of the machine could be viewed as a dynamic multiset of statements with new elements popping up and old elements, potentially, disappearing.

If T is consistent, in the computation, Case (b2) will never be activated. The result is that the machine enumerates the theorems of T on the display. However, we also know that the machine does not simply enumerate the axioms but that it follows an experimental procedure where problematic axioms are discarded. Moreover, if we do not know whether T is consistent, we can see that eventually the number v will stabilise and from that point on the theorems enumerated will not be retracted.

Let T^* be the theory of the sentences that are displayed in the limit, to be precise a sentence A is in T^*, if, in a run of the program, from some time on, a copy of A is in the list and remains there. We have:

a. If T is consistent, then T is T^* and the enumeration of the theorems of T^* mimics the enumeration of the theorems of T.

b. T^* is consistent.

c. T proves that T^* is consistent.

So, by (a), if T is consistent, then T^* proves that T^* is consistent.

d. If T proves A, then T proves that T^* proves A.[22, 23]

We can also design a *Henkin machine* that produces a complete consistent extension of Peano Arithmetic in the limit.

Let us consider, for example, the Feferman machine F_{PA} of Peano Arithmetic. What it does can be described, in a sense, as enumerating the theorems of Peano Arithmetic. If we had a multi-tape Turing machine that implements the Feferman machine, we could with justice say of the theorems appearing on a designated tape that they are the theorems of Peano Arithmetic. In fact, there could be a second Turing machine that enumerates the theorems of Peano Arithmetic in a straightforward way that is behaviourally equivalent to our realisation of F_{PA}. However, I submit it is still fair to say that the Feferman machine does something different from mere enumeration. It follows an experiential procedure involving a preparedness to withdraw theorems—even if in fact such a withdrawal never happens.

Remark A.1. Even if T and T^* are extensionally the same theory, their Gödel sentences are entirely different things. This is because the Gödel sentence depends on the representation of the axiom set.

[22] This insight, due to Feferman, uses a special feature of extensions of Peano Arithmetic in the arithmetical language. There are other theories in the arithmetical language, like Elementary Arithmetic, for which this does not hold. There are extensions of Peano Arithmetic in an extended language, like ACA_0, for which it does not hold.

[23] In contrast to this, if T is consistent, then T does not prove: if T^* proves G^*, then T^* proves that T^* proves G^*.

However, just as with ordinary Gödel sentences of consistent theories, if T is consistent, then G_{T^*} is true and hence unprovable. But, unlike ordinary Gödel sentences of consistent theories, both $T^* + G_{T^*}$ and $T^* + \neg G_{T^*}$ are interpretable in T^*. What if T is inconsistent? By tweaking the program of the Feferman machine a bit, one can produce an example where G_{T^*} is provable in T^* and, hence, false.

In a sense, the most interesting example is the theory enumerated by the Henkin machine over Peano Arithmetic. We know that this theory is consistent. However, both the Gödel sentence obtained by the Gödel fixed point construction and its negation satisfy the Gödel fixed point equation. As a consequence, nobody knows which of the two is true. We note that the truth of one of these sentences could crucially depend on implementation details. Can one tweak these details to make a designated solution of the fixed point equation true? ❍