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 $f: \mathcal{N} \to \mathcal{N}$ 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 $f: \mathcal{N} \to \mathcal{N}$ (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^0_1-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^0_1-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 prop-

ositions 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 g r o u p s o f l e v e l s 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 p r o v a b l e w i t h u n a s s a i l a b l e m a t h e m a t i c a l c e r t a i n t y...

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** + $Con_T \vdash \psi$, 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.