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 $\varphi(S^{(n)}0)$ for all $n = 0, 1, 2, \ldots$; here, "$S^{(n)}0$" 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] m a c h i n e 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 ma-

chine 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 i s equivalent to mathematical intuition, but cannot be p r o v e d to be so, nor even be proved to yield only c o r r e c t 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 r e a s o n i n g 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, Tu-

ring'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 pro-

cesses 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 $Cons_{NF}$ 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 $Cons_{NF}$ 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: M_n 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: $n \in C$} 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 $S(M_n)$, 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: $C = \{n: S(M_n)$ 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: $C \subseteq \mathrm{dom}(F)$.

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

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

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.

P r o o f : Assume that the set of F's values, $A = \{F(n): n \in \mathrm{dom}(F)\}$, 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, $A = S(M_k)$. By assumption, A is consistent, so $k \in C$, 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 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 $S(M_n)$ 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 $n \notin C$, 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 $T(M_n)$. If k codes this proof, then the arithmetical statement "the number k is the proof of G_n in $T(M_n)$" has only restricted quantifiers and is true. It is provable in basic arithmetic, so $T(M_n) \vdash Prf(S^{(k)}, \ulcorner \varphi \urcorner \ulcorner G_n \urcorner)$, and this contradicts the provability of G_n, as on account of the definition of G_n, $T(M_n) \vdash \neg(\exists x)\, Prf(x, \ulcorner G_n \urcorner)$.

b) The assumption (ii) does not exclude *a priori* the equality of C and $\mathrm{dom}(F)$, or that F is defined just for $n \in C$. 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 $m(n) = \min \{k: S(M_k) = S(M_n)\}$ 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 re-

sponse 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, $S(M_n)$ 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 para-

doxical 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*.

P r o o f: 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 h i d d e n 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 $\equiv T$"—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 $\equiv T$,

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

(3) so T is sound;

(4) hence $T' = (T +$ "the mind $\equiv T$") 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 $\equiv T$ 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 ≡ *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 ≡ *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 ≡ *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 ¬(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') ∧ (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') ∧ (2). This analysis fits Putnam's criticism. Assuming that (1) ∧ (2) corresponds to Case I (presented above, in Sections 6.1–6.3), the assumption (1') ∧ (2') corresponds to Case II as it was understood by Penrose. And further, the apparently safer assumption (1') ∧ (2) corresponds to Case IV; it does not involve (2'), our understanding of the algorithm *T*.

While (1') ∧ (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 $\vdash$ (*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 con-

sistency 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 philo-

sophical 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 Princi-

pia 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.