# On the necessary philosophical premises of the Gödelian arguments[1]

VINCENZO FANO†
PIERLUIGI GRAZIANI‡

†University of Urbino, Italy
‡University of Urbino, Italy

Lucas-Penrose type arguments have been the focus of many papers in the literature. In the present paper we attempt to evaluate the consequences of Gödel's incompleteness theorems for the philosophy of the mind. We argue that the best answer to this question was given by Gödel already in 1951 when he realized that either our intellectual capability is not representable by a Turing Machine, or we can never know with mathematical certainty what such a machine is. But his considerations became known only in recent times when many scholars were already aware of Benacerraf's and Chihara's analyses on the consequences of Gödel's incompleteness theorems for the philosophy of the mind. Benacerraf and Chihara, in fact, discussing Lucas' paper, arrived at the same conclusions as Gödel in the sixties, but in a more formal way. After Penrose's provocative arguments, Shapiro again shed light on the question. In our paper, after a broad and simple presentation of the contributions to the debate made by different authors, we show how to present Gödel's argument in a rigorous way, highlighting the necessary philosophical premises of Gödel's argument and more in general of Gödelian arguments.

## 1. Introduction

Many scholars have tried to prove the thesis of the irreducibility of human intelligence to a calculator machine using Gödel's famous Incompleteness Theorems. They did so by creating the so called Gödelian arguments. The debate over the centrality of such arguments in the crisis of the so called Strong Artificial Intelligence is well known. But its complexity has often tended to make the debate very difficult to understand, generating widespread misunderstanding. As the logician and philosopher of mathematics Steward Shapiro (*1998*, p. 277) has pointed out: "[…] many philosophers dismiss the whole Lucas-Penrose controversy, often by rolling their eyes": this attitude seems to be a consequence of the complexity that was mentioned above, and of the misunderstandings that some authors have created through their ways of interpreting the issue in question. These misunderstandings have often spread into the popular literature, creating erroneous paradigms.

This article will give a clear and homogeneous reconstruction[2] of both Gödelian arguments and the major literature of reference. It will do so by highlighting the strong and weak points of the reasoning, which will enable us to understand, easily and in analytical depth, the relation between Gödel's results and general considerations about human intelligence. As will be seen, and contrary to a widespread idea, Gödel's theorems do not say anything about our superiority over computers. On the contrary, they tell us something important about our intelligence, and about what we can understand of ourselves, in principle, by means of computational models. Gödel himself, as we shall see, already had a very clear understanding of the real implication of his theorems for the philosophy of the mind, that is: either human intelligence has a non-computational nature, or, even though human intellectual activity can be reproduced by a Turing machine, it cannot fully understand its own working. To put it in an evocative way, we could say, following Paul Benacerraf (*1967*, p. 30): 'if I am a Turing machine, then I am barred by my very nature from obeying Socrates' profound philosophic injunction: KNOW THYSELF'. As will be shown, Gödel expounded his

[1] "Our brain is sufficiently complex to understand that it is too complex to be understood by itself" [Murphy's Law].
[2] An interesting reformulation is offered by *Antonelli 1997*. A more concise presentation is that of *Odifreddi 1992*.

position in a dilemmatic informal way, but it is possible to build a more precise argumentation in favor of that position. In this paper we shall describe a possible argument which can lead to Gödel's conclusion. Not only shall we do so by making use of Gödel's own indications, but also by using the works of other thinkers such as P. Benacerraf and C. Chihara who, independently of Gödel, and in a more formal way, took similar directions, at least partly, to those of Gödel. These analyses will also reveal the necessary philosophical premises of the Gödelian arguments and, in particular, of Gödel's argument.

## 2. Sketching Gödel's theorems[3]

Let's start with an intuitive description of the proofs of Gödel's theorems[4]. This will enable us to bring all fundamental concepts (and related symbols) to mind and to use them as necessary tools for analyzing the Gödelian arguments.

As it is known, it is possible to formulate the *Intuitive Arithmetic* (*N*) as a formalized system, which we call *Formal Arithmetic* (*FA*). In *FA* the natural numbers are formulated by means of closed terms. For example: 0, 1, 2, 3, etc. are in *FA* **0**, **s(0)**, **s(s(0))**, **s(s(s(0)))** etc. These are called *numerals* and we will indicate them with **bold** letters: given the natural numbers *n*, we use **n** to indicate the corresponding numeral.

We can *represent* in *FA* the elements of *Intuitive Arithmetic*, for example the relation ≤. We can show that there exists a formula $\alpha(x, y)$, where x and y are free variables, that represents in *FA* the relation ≤. Such formula is $\exists z(x + z = y)$.

If *n* and *m* are two natural number whatever, then:

$$\text{If } n \leq m \text{ holds in } N, \text{ then } \vdash_{FA} \exists z\ (\mathbf{n} + z = \mathbf{m})$$

$$\text{If } n \leq m \text{ does not holds in } N, \text{ then } \vdash_{FA} \neg\exists z\ (\mathbf{n} + z = \mathbf{m})$$

and if it is defined in *FA* the symbol ≤ by means of the following formula:

$$\forall x\ \forall y\ (x \leq y \leftrightarrow \exists z(x + z = y))$$

then we can prove in *FA* the well note properties of ≤.

Define the following three argument predicate $T(m,k,y)$, where *m*, *k*, *y* are three natural number, and say that it is true, if and only if the Turing Machine of index *m* ($MT_m$), applied to an input *k*, stops after *y* calculation steps. We can prove that *T* is *recursive*. Given that it is possible to formulate the Intuitive Arithmetic (*N*) as a formalized system and that in *FA* are representable all and only the recursive functions and predicate, and that *T* is recursive, then it is possible to *represent T* in *FA*; that is, we can show that there exists a formula **t**(x,y,z) in *FA*, with exactly three free variables, which represents *T*. Therefore:

$$\text{If } T(m,k,h) \text{ holds in } N, \text{ then } \vdash_{FA} \mathbf{t(m, k, h)}.$$

$$\text{If } T(m,k,h) \text{ does not hold in } N, \text{ then } \vdash_{FA} \neg\mathbf{t(m, k, h)}.$$

Let us now define the sentence $\exists y T(m,k,y)$: this sentence is true if and only if there exists a finite number of steps *y* in which a $TM_m$ machine stops, given as input the number *k*; otherwise it does not hold.

---

[3] The present exposition only takes into account a criterion which will be instrumental for the analyses which will appear in later chapters.

[4] There are many wonderful treatments of Gödel's theorems in the literature, at various levels of sophistications, for examples: *Franzen 2005*; *Smullyan 1992*. We are using here the highly intuitive exposition given by *Frixione* and *Palladino 2004*, and by *Kleene 1967*.

We define, furthermore, the particular case of $\exists y T(m,k,y)$, i.e. $\exists y T(x,x,y)$: it holds if and only if there exists a finite number of steps $y$ in which a $TM_x$ machine stops, given as input the number $x$; otherwise it does not hold. It is possible to prove that $T$ is a *general recursive* predicate; and that $\exists y T(m,k,y)$ and $\exists y T(x,x,y)$ are *recursively enumerable*, but not *general recursive*.[5]

But, if $\exists y T(m,k,y)$ is not general recursive, then this is equivalent, for Church's Thesis, to the undecidability of the halting problem for Turing machines. Therefore relation $\exists y T(m,k,y)$ and property $\exists y T(x,x,y)$ are semi-decidable but not decidable.

Now, in *FA*, $T(m,m,y)$ corresponds to the closed formula $\exists y\ \mathbf{t}(\mathbf{m},\mathbf{m},y)$ which we shall shorten to $\mathbf{t_m}$. If $\exists y T(m,m,y)$ is true in $N$, then, for some $h$, it is true in $N$ $T(m,m,h)$, and therefore $\vdash_{FA} \mathbf{t(m, m, h)}$. But then, introducing the existential quantifier we have $\vdash_{FA} \exists y\ \mathbf{t(m, m,}\ y)$, i.e. $\vdash_{FA} \mathbf{t_m}$. So:

(a) If $\exists y T(m,m,y)$ holds in $N$, then $\vdash_{FA} \mathbf{t_m}$.

In addition, the converse holds true, because we know that within the standard model of *FA* $<N,+,\times>$ all *FA* theorems hold true. Hence:

(b) If $\vdash_{FA} \mathbf{t_m}$ then $\exists y\ T(m,m,y)$ holds in $N$.

It is easy to see that if *FA* was *decidable*, that is, if, for any given *numeral* **m,** it was always possible to determine whether, in *FA*, $\mathbf{t_m}$ or $\neg\mathbf{t_m}$ are derivable or not; then $\exists y\ T(m,m,y)$ would be decidable, in contrast to Turing's Halting Theorem[6]. Therefore starting from the demonstration of Turing's Halting Theorem, we can prove that *FA* is *undecidable*.

Given all this we can now describe Gödel's First Theorem proof in a simple way.

**Gödel's First Theorem:**

*There exists a closed formula G of FA so that, in FA the G formula is not provable, and neither it is ¬G. Yet G is true in N.*

Let us consider the formula of *FA* $\neg\ \mathbf{t_m}$.
By definition:

If $\vdash_{FA} \neg\mathbf{t_m}$ then in $N$ it holds that 'there does not exist a $y$ such that $T(m,m,y)$'.

It also possible to prove that the converse:

If 'does not exist a $y$ such that $T(m,m,y)$' is true in $N$, then $\vdash_{FA} \neg\mathbf{t_m}$

does not hold.
Therefore, there exists a natural number $p$ such that 'there does not exist a $y$ such that $T(p,p,y)$' holds true within $N$ and not $\vdash_{FA} \neg\mathbf{t_p}$. Let's call $G$ the formula $\neg\mathbf{t_p}$, then $G$ is not provable in *FA*.[7] Furthermore, $\vdash_{FA} \neg G$ does not hold, because, to the contrary, if $\neg\mathbf{G}$ was provable in *FA*, then $\neg\neg\mathbf{t_p}$,

---

[5] *First Church's Theorem* indeed says that the binary relation $\exists y T(m,k,y)$ is not a general recursive relation.
[6] See *Turing 1936*.
[7] Later, and only when misunderstandings may arise, we will specify the formulae belonging to the system by adding symbols such as, for instance, $G_{FA}$.

that is $\mathbf{t_p}$, would also be provable. Hence, by (b) there would exist a *y* such that *T*(*p,p,y*) would be true in *N*, in opposition to the definition of *G*. In the end *G* holds in *N* because we know that, in the standard interpretation, $\neg\mathbf{t_p}$ is associated to proposition 'there does not exist a *y* such that *T*(*p,p,y*)' which is true in *N*.

We have to note that if *FA* was *inconsistent*, it would be impossible to prove Gödel's First Theorem and that the *incompleteness* we refer to *Gödel's First Theorem* here is a syntactic one, and not semantic: it concerns the provability of the formulae in *FA*, and not their truth.[8]
Therefore, the *Gödel's First Theorem* can be also set out as follows:

*If FA is consistent, then FA is syntactically incomplete*

Intuitively:
Let us take a formal system (which will be named *U*). Let it be able to express the Robinson's portion of Arithmetic[9], soundness property included. Furthermore, let us consider a logical language *L*. Let us assume by hypothesis that statement *G*, which says about itself that it is unprovable in *U*, can be expressed in *L*:

(*G*) *G* is not provable in *U*.

Let us now ask ourselves whether *G* is provable or not in *U*. Let us suppose that *G* is provable; then, for what it says, it would be a false statement. This would mean that the formal system *U* would be not sound inasmuch it allows a false statement to be proved. Hence, if *U* is sound, *G* is unprovable in it. On the other hand, if *G* is unprovable in *U*, then it is a true statement. Hence *U* is semantically incomplete: there exists a true statement, *G*, which *U* cannot prove. Furthermore, since *G* is true, the formal negation of *G* will be false, because the negation of a statement is false if and only if this statement is true, and vice versa. Therefore, neither *G* nor its negation ¬*G* are provable in *U*, hence the formal system *U* is also syntactically incomplete, and *G* is formally undecidable in *U*.[10]

More precisely Gödel's first theorem therefore states that:

*If U is a sound formal system which is able to express Robinson's arithmetic, then there exists a statement G, formulated in language L of the system, such that G is undecidable in U, that is it is neither provable nor refutable*.[11]

**Gödel's Second Theorem:**

We can formulate *FA* consistency by means of a *FA* closed formula. Let us consider the algorithm that, after taking as input any natural number *m*, scans all the theorems in *FA* looking for a contradiction. If the algorithm finds a contradiction, it returns 0, else it keeps running forever. Let

---

[8] Let us briefly recall that: a formal system *U* is *consistent* when, for each formula α in the formal language *L*, *U* does not allow both the formula itself and its negation to be proved. A formal system *U* is *syntactically complete* when, for each formula α, either *U* proves this formula or *U* proves its negation. When a formula is provable or refutable in *U* we say that it is *formally decidable* in *U*. A formal system *U* is *sound* if it is never the case that if α is a false formula, *U* proves α, that is *U* is a system which only proves things that are true. A formal system *U* is *semantically complete* when it proves all its true sentences; that is when it is not the case that a formula is true and it is not a theorem of the system.
[9] Robinson R. M. showed that, in order to derive Gödel's incompleteness theorems, one only has to assume a fragment of arithmetic that today is currently called *Q*, or *Robinson's Arithmetic*. See *Robinson 1950*.
[10] In the first paragraph of his famous article, see *Gödel 1931*, Gödel expounded his theorems using this informal explanation, that is an argumentation of a semantic kind, which, unfortunately, has usually misguided studies on incompleteness theorems.
[11] As it is known if we add *G* to *FA* as a further axiom, Gödel's Theorem is still provable in the new formal system (*FA*+*G*), and the same holds for any addition of a decidable set of axioms.

$MT_c$ be the Turing Machine which executes such algorithm. The consistency of *FA* is then equivalent, for each *m*, to "there does not exist *y* such that $T(c,m,y)$" and in particular to "there does not exist *y* such that $T(c,c,y)$" which is expressed in *FA* by the closed formula $\neg\mathbf{t_c}$. So $\neg\mathbf{t_c}$ expresses in *FA* the consistency of *FA* which we denote with Con(*FA*). The condition of soundness can be weakened down to consistency[12], and we can prove that

*if FA is consistent, then G is unprovable in FA*

or

*if FA is consistent, then "there does not exist a y so that T(p,p,y)" holds true in N.*

Then this implication can be formulated in *FA* as follows:

$$\text{Con}(FA) \rightarrow \neg\mathbf{t_p}$$

and it can be proved in *FA*:

$$\vdash_{FA} \text{Con}(FA) \rightarrow \neg\mathbf{t_p}$$

From here Gödel's Second Theorem can be easily proved:

*If FA is consistent, it is impossible to prove that within FA*

or, equivalently

*If FA is consistent, then not* $\vdash_{FA}$ Con(*FA*)

If it was $\vdash_{FA}$ Con(*FA*), then, by $\vdash_{FA}$ Con(*FA*) $\rightarrow \neg\mathbf{t_p}$ and by *modus ponens*, follow $\vdash_{FA} \neg\mathbf{t_p}$, hence $\vdash_{FA}$ G. But this is excluded by Gödel's First Theorem.[13]

## *3. Gödelian Arguments*

### *3.1 Lucas' Argument*

Gödel's results mentioned above have had wide application, even beyond the field of logic and mathematics; and this has contributed to their popularity. As we have mentioned, this article aims to investigate their applications which may be considered as the most controversial: that is the implications of the mentioned theorems in the philosophy of mind. Today it is common to think that it would be possible to represent the whole human subjectivity through algorithms (for example: D. Dennett, J. Fodor, P. Churchland). We will not concern ourselves with this point of view, as it encounters great difficulties in relation to the renowned mental experiment proposed by T. Nagel, J. Searle and F. Jackson.[14] On the contrary, we will concern ourselves with a more limited project, which was meant to reproduce, or mechanically simulate, the intelligent behaviour of human beings. This project was launched by Turing in 1950,[15] and was improperly called *mechanism*. We'll adopt this term too, since it has come into use. This project has been fully discussed and if, on the one hand, it has provided its advocates with theoretical tools, on the other hand, it has led anti-mechanists to build its refutations. Aside from a complete understanding of Turing's thought,[16] what we are concerned with here, is how Gödel's theorems were almost immediately seen as tools for refuting the mechanistic thesis; whether we consider it in an *extensional* way (mind's procedures and results are mechanisable), or in an *intensional* one (human intelligence is a particular machine).

---

[12] If a system is sound, it is also consistent.

[13] This, obviously, does not mean that no demonstrations exist which prove the consistency of *FA*, but that these must necessarily use more complex systems, in which, however, Gödel's Second Theorem is provable. Hence, if mathematics is globally consistent, no demonstration of its consistency exists.

[14] For information on these issues, see the work: *Hofstadter and Dennett 1981.*

[15] See *Turing 1950.*

[16] A careful analysis of Turing's remarks on mechanism and Gödel's theorems has been given by *Bruni 2004.*

Turing himself understood such implications of the theorem;[17] beside him, P. Rosenbloom (*1950*), G. Kemeny (*1959*), and E. Nagel and J.R. Newman (*1958*), in the 1950's, developed argumentations hinged upon the idea that Godel's Theorems could provide a logical tool to refute the philosophical thesis of mechanism.[18] Despite this tradition, Gödelian anti-mechanists argument is linked to the name of the English philosopher John Randolf Lucas. In 1961, he developed an argumentation aimed at demonstrating, on the basis of Godel's theorems, that it is not possible to represent human intelligence with a Turing machine. Lucas' argument can be presented schematically as follows.

**L1.** Suppose that there exists a Turing Machine, *TM*, which has exactly the same intellectual ability as human beings.
**L2.** *TM* should be able to produce all theorems of some formal system *U*, which contains arithmetic (*FA*).
**L3.** However, *TM* is not able to produce, as true, (Gödel's) *G* formula of *U*.
**L4.** On the contrary, the human being has the intellectual ability to see that *G* is true.
**L5.** Hence *TM* does not reproduce all intellectual abilities of human beings, against **L1**.

As we have seen, *G* undecidable statement of *FA* is decided through a semantical argumentation: *G* says of itself, by gödelisation, that it is not provable; if it was provable, it would be false, but since *FA* does not prove falsity, then *G* should not be provable in it; hence, *G* is true. This argument cannot be formalized in *FA*, because it would require the notion of truth which, by a famous theorem proved by Alfred Tarski, cannot be formalized in *FA*.
Lucas' argument, although apparently highly persuasive, contains some problems, which provoked intense debate in the literature. Despite these problems, the appeal for this argumentation, as we shall see, pushed several scholars into trying to make improvements and revisions of Lucas' reasoning. An argument similar to that of Lucas had been already put forward, as we said, by J. R. Newman and E. Nagel (*1958*). Nonetheless, it had been criticized by H. Putnam (*1961*) who remarked that step L3 is problematic. This is because Gödel's theorem for *U* formal system states that, *only if U* is consistent, *G* is not decidable. Suppose we have a Turing machine which should reproduce all human intellectual activity, that is the set of sentences of *U*. It is now possible to find an undecidable sentence *G* such that we can prove in *U* that, if *U* is consistent, then *G* is true, but not decidable by *TM*. But, in order to show that the mechanistic thesis leads to a contradiction, we need to prove the consistency of *U*. Yet, it is not easy to know whether or not *TM*, the machine supposed to simulate all human intellectual abilities, will produce a consistent set of theorems. If, however, by the first theorem, Godel's statement is undecidable and true, only if the system is consistent; and if, by the second theorem, it is not possible to give an absolute demonstration of consistency of the system in it; it only remains for us to give a demonstration of relative consistency, that is a demonstration that the machine which represents us is consistent provided that we are. Lucas, aware of the problem raised by Putnam, therefore develops a series of arguments in favour of the consistency of human intelligence. But giving such a relative demonstration of consistency would mean that a human being would be able to do what a machine or formal system would not: that is to assert their own consistency in an absolute way. For this reason, Lucas' development of argumentations in favour of the consistency of the human mind does no more than put forward a Quine-Davidson-style generic principle of indulgence or charity towards human beings; and in this sense his Gödelian argumentation preserves the highlighted weakness.
Apart from the problem raised by Putnam, Charles Chihara (*1971*) emphasized how L4 of Lucas' argument was not clear, because one cannot understand what it means to say that a human being is able to see the "truth" of *G*. What the first part of the Incompleteness Theorem proof states is, as we know, that if Formal Arithmetic is consistent, then *G* is unprovable in it. If *G* is unprovable, then *G*

[17] See *Turing 1992*. For an interesting analysis of this work, see: *Bruni 2004* (chapter 3).
[18] See *Smart 1961*.

is what it says it is, and in this sense it is a true statement. But in order to "see" that the Gödelian statement for *FA* is true, we have to "see" before that *FA* is consistent (or sound). Hence, Lucas' Gödelian argument against strong artificial intelligence requires, for any machine (or *U* formal system) that satisfies the well-known hypothesis of Gödel's first theorem, that human beings are always able to "see" the truth of its Gödelian statement. But exactly this "always able to see" could not be taken for granted. The weakness of this thesis appears in all its strength in the development of Lucas' argument. The argument imagines in fact that, the different Gödelian statements being of the same form, it would be possible to augment the considered formal system (or machine) with an axiom scheme which would generate the infinite set of Gödelian axioms. As already noticed, we can add *G* to *FA* thus getting $FA_1$ in which *G*, being an axiom, is provable by definition. But $FA_1$ is in turn an incomplete system because it contains an undecidable statement, *G'*. It is possible to keep adding ad infinitum the Gödelian statement to the initial system. Lucas thinks that this process of adding Gödelian statements could be incorporated in the system precisely by using an axiomatic scheme for Gödelian statements. Lucas' thesis is, at this point, that if we added to a formal system the infinite set of axioms included in the following Gödelian formulae, the resulting system would still be incomplete and it would contain an unproved formula within the system; a formula, nonetheless, that a human being could keep seeing as true. But such human ability is exactly what appears to be doubtful, as has been masterfully stressed by Douglas Hofstadter (*1979*) and by Stewart Shapiro (*1998*, p. 285 ff.). Imagine we add the Gödelian statements $G_1, G_2, G_3, \ldots, G_n$ . At some point, as these statements have the same form, we would find ourselves adding the axiomatic scheme $G_\omega$ where $\omega$ is the first transfinite ordinal set[19]. Obviously we could keep adding to our system $FA_\omega$ the statement $G_{\omega+1}, G_{\omega+2}, G_{\omega+3}, \ldots, G_{\omega+n}$ , and this, *mutatis mutandis* would lead us to add to the original system a new schema which can be denoted by the limit ordinal $2\omega$. Obviously we could consider the successors of $2\omega$ thus coming to limit ordinals $3\omega$, $4\omega$, …, $\omega x\omega = \omega^2$, and then $\omega^3$, $\omega^4$, ..., $\omega^\omega$ etcetera up to $\varepsilon_0$ which is the first ordinal that cannot be obtained by $\omega$ with a finite number of additions, multiplications, and exponentiations. Hence, as Douglas Hofstadter (*1979*, p. 475) points out: by a theorem due to A. Church and S Kleene, there does not exist a recursive system of notations, which is capable of assigning a name to all recursive ordinals, and for this reason it seems highly arguable and certainly strange that human mind itself could go beyond recursive ordinals. Paraphrasing Hofstadter: at a certain point the human being will reach the limits of his ability to gödelize, and henceforth the formal systems (machines) of that complexity will have the same power as this human being.[20] In this sense, the fact that, as Lucas writes, human beings can always "see" the truth of Gödelian statement, could not be taken for granted.[21]

### *3.2 Benacerraf's Argument*

As we observed, Lucas' argument, although not new, provoked a large debate on the issue of whether or not it would be possible to find a refutation of the mechanistic thesis based on Gödel's theorems. Such a debate involved not only philosophers, but also logicians, mathematicians, computer scientists, etc. who dealt with Lucas' argumentation and emphasized its weakness and strengths.[22] Among such contributions are some that command our attention for their analytical depth, thus becoming a benchmark for understanding and development of pro and cons arguments concerning mechanism. One of these is certainly the article 'God, the Devil, and Gödel' by the philosopher of mathematics Paul Benacerraf. Driven by the conviction that Lucas' argument was not capable of proving what it claimed (that is, that mechanism is an indefensible position), he

---

[19] $\omega = \{0,1,2, 3...\}$ where 0,1,2,3 … are ordinals.

[20] A similar thing can be said about Gödel's second theorem.

[21] On transfinite recursive progressions see, aside from the quoted texts by Hofstadter and Shapiro, especially: *Feferman 1962*; a highly intuitive exposition is given by: *Berto 2008* (p. 214 f.).

[22] A bibliography is available in Lucas' home page: http://users.ox.ac.uk/~jrlucas/ .

presents an argument which, starting from the assumption that the human mind is at most a Turing Machine, and that we know this machine leads to a contradiction using Gödel's theorems. Benacerraf arrives at a different conclusion from that of Lucas, which is however very interesting. He tries to solve the various open problems within Lucas' argument, which we outlined in the previous chapter, by building a new argumentation. *In primis* he stresses that it is necessary to limit the notion of "man's intellectual abilities" by introducing a set, *S*, defined as "every statement that I can derive and that I know to be true". In this way Benacerraf gets around the problem, also raised by Chihara, of using the unclear and ambiguous concept of "seeing the truth of *G*". In fact, while Lucas' argument was limited to claiming that *TM*, which was supposed to represent human intellectual abilities, only operated syntactically; and could therefore generate an inconsistent set of statements, that is, a set containing contradictions, *S*, on the contrary, is necessarily consistent, as its statements are not only derivable, but also true. As is well-known, in fact, it is possible to prove that if the statements of some formal system are true within a certain model, then that formal system is consistent. In this way Benacerraf also overcomes Putnam's objection. In addition, Benacerraf identifies a further weakness of Lucas' argument (which will be outlined again independently, and in a slightly different but more effective way, by Daniel Dennett (*1972*)[23]: he asks what it is that, according to Lucas, a *TM* cannot do and I can do. I can find a semantic proof of *G*, that is, I can get out of the system *U* and identify a model of *U*, which makes *G* true. But are we sure that *TM* could not do the same thing? As Dennett will rightly point out in his paper of 1972, *TM* is a physical system, which, therefore, in order to represent a *TM*, should be adequately interpreted; that is, we need to establish what the input and output are, how to code data, etc.. Moreover, the physical process which realizes *TM* is certainly very complex, thus if we interpret it differently, it might generate another *TM* capable of deriving *G*. Hence step L3 step of Lucas' argument is problematic, even from this point of view. In the end, Benacerraf, unlike Lucas, clearly distinguishes the purely mathematical side of the limitative theorems, which as such does not have any philosophical meaning, from the real philosophical argument, which also needs, apart from Gödel's theorem, what he calls a *philosophical* premise; which nevertheless also has, as we'll point out, an *empirical* extent. On the basis of Lucas' argument, Benacerraf builds a new and more precise argument. Let's first try and clarify its points in an informal way, and subsequently in a more formal one.

**B1.** Remember that *S* is the set "every statement that I can derive and that I know to be true". *S** is the logical closure of *S* within a formal system[24] which is sufficiently large and sound; that is, *S** contains every statement which is derivable from *S* within a reasonable formal system. Notice the modal character of the expression "can" . It is clear that, if we were only referring to the statements that I effectively derive, which, however large a set it may be, is of finite cardinality; there is no doubt that it would be representable by a *TM*. That is the reason why we need to introduce the expression "can", which nevertheless prevents a very precise characterization of the set *S*.

**B2.** Assume that there exists an effectively enumerable set $W_j$ such that:

**(a)** The statement that $W_j$ includes all theorems of arithmetic (*FA*) belongs to *S**. In symbols

$$'AF \subseteq W_j' \in S^*$$

Notice that $W_j$ is effectively enumerable if and only if there exists a calculable and total function $f_j$, which associates to each natural number an element of $W_j$ with possible repetitions. Moreover $f_j$ is calculable if and only if there exists an algorithm which, given as input a possible argument of $f_j$, gives as output the respective value of the function. Yet, if we accept the Church-Turing thesis, then the set $W_j$ is effectively enumerable if and only if there exists a *TM* of number *j* which calculates the

[23] Similar objections can be found in *Boyer 1983*.

[24] Benacerraf refers to the first-order predicate calculus, but afterwards, as Chihara points out, he introduces other resources in *S*, so that the first-order predicate calculus is not sufficient.

function $f_j$. Intuitively (a) claims that I can build a *TM* capable of enumerating every theorem of arithmetic.

**(b)** The statement according to which $W_j$ is included in $S^*$ is part of $S^*$. In symbols:

$$'W_j \subseteq S^*' \in S^*$$

Intuitively (b) claims that I am able to derive that $W_j$ is a subset of $S^*$.

**(c)** $S^*$ is a subset of $W_j$. In symbols:

$$S^* \subseteq W_j$$

Intuitively (c) claims that there exists a *TM* capable of generating every true statement that I can derive.
It is clear that if I can derive that $W_j$ is a subset of $S^*$ and I know that all statements that belong to $S^*$ are true, and moreover $S^*$is a subset of $W_j$, then $W_j$ and $S^*$ coincide.

**B3**. From (a)-(c), through gödelisation procedure, it is possible to derive a contradiction in $S^*$, which indeed we know to be consistent. Hence, at least one of the hypotheses (a)-(c) must be false. Consider that the hypotheses (a)-(c) have an empirical-philosophical nature (because although they refer to the contents of *S*, that is to what I can prove, and is true, *S* is nevertheless defined in modal terms). It is difficult to maintain that (a) is false, that is that I am not able to develop an algorithm which generates all the theorems of arithmetic. So there are two possibilities: *either my deductive abilities are not representable by a TM – that is (c) is not true – or I am not able to derive the TM that represents me – that is (b) is not true*.
In conclusion, either my deductive abilities are not representable by a Turing Machine, or I do not know which Turing machine represents myself.[25]
Let's follow in formal details the various steps of Benacerraf argument.[26]

**Bf1.** Let $S = \{x \mid \text{I can prove x and x is true}\}$
*S* represents my deductive output of true statements.
**Bf2.** Let $S^* = \{x \mid S \vdash x\}$
$S^*$ is the formal closure of *S*.
**Bf3.** $S^*$ is consistent.
Since each member of *S* is true – Benacerraf says "I can't prove what is false" – and the system in which *S* is defined preserve the truth, then $S^*$ is consistent.
**Bf4.** $'Con(S^*)' \in S$
Let us use *Con* to indicate the consistency predicate. The previous steps 1-3 constitute a demonstration of the consistency of $S^*$. But since I have proved it, this counts as one of my output.
**Bf5.** $'Con(S)' \in S$
This derives from $S \subseteq S^*$ and Bf4. This corresponds to Lucas' assertion that he knows he is consistent, and so Putnam's objection disappears.
**Bf6.** $(x)\ (W_x \subseteq S^* \supset Con\ (W_x))$
by *W* we denote any effectively enumerable set. Since $S^*$ is consistent, then all enumerable subsets that it contains are also consistent.
**Bf7.** $'(x)\ (W_x \subseteq S^* \supset Con\ (W_x))' \in S$
This results from the fact that Bf1-Bf6 is a proof produced by me. Indeed all the proofs that I'm producing enrich *S*.
**Bf8.** $'(x)\ (W_x \subseteq S^* \supset Con\ (W_x))' \in S^*$
from $S \subseteq S^*$ and Bf7.

[25] Notice that Lucas answers Benacerraf in *Lucas 1968*.
[26] This exposition follows closely the already clear one given by Benacerraf himself, but simplifies it in some parts.

**Bf9.** Let us assume that there exists a recursively enumerable set $W_j$ such that
**(a)** '$Q \subseteq W_j$'$\in S^*$
by the symbol $Q$ we denote the formal closure of the axioms of *FA* which are necessary to prove Gödel's theorems. (a) denotes that this formal closure is representable by an enumerable set (Turing Machine) $W_j$.
**(b)** '$Wj \subseteq S^*$ '$\in S^*$
I can prove that $W_j$ is a subset of my output.
**(c)** $S^* \subseteq W_j$
What I can prove is a subset of $W_j$. As we have seen in the informal argumentation, the fact that $W_j$ is enumerable is the condition for it to be an output of a theorem proving Turing Machine. By gödelisation we can interpret $W_i$ (for any integer $i$) as the Gödel number of a set of recursive enumerable theorems, that is the theorems that the $i$-th machine can generate.
**Bf10.** $Q \subseteq W_j$
This results from Bf9a and that everything I can prove has to be true.
**Bf11.** There is a formula $H$ (having the property defined by Gödel,[27] that no number is the Gödelian of a demonstration of the formula whose Gödelian is $H$'s Gödelian) such that precisely if $H \in W_j$, then $\neg H \in W_j$, and $W_j$ is inconsistent. This is the version of the Gödel first theorem applied to $W_j$. $W_j$ is in fact adequate for arithmetic by Bf10, and representable as a formal system by Bf9.
**Bf12.** '$Con(W_j) \supset H$ '$\in W_j$
This is the step immediately preceding Godel's second theorem,[28] applied to $W_j$ (also by Bf9 and Bf10).
**Bf13.** '$Wj \subseteq S^* \supset Con(W_j)$'$\in S^*$
This results from Bf8 and that $S^*$ is formally closed.
**Bf14.** '$Con(W_j)$'$\in S^*$
This results from Bf9b, from Bf13, and from the fact that $S^*$ is closed under *modus ponens.*[29]
**Bf15.** '$Con(W_j)$'$\in W_j$
In this step Bf9c plays a fundamental role. It claims that $S^*$ is part of the output of a Turing Machine, but if this is the case, then the Turing Machine can prove its own consistency and is thereby consistent. So we have:
**Bf16.** $H$ and $\neg H$ are in $W_j$, and $W_j$ is inconsistent.
It follows from Bf9b that $W_j \subset S^*$, hence:
**Bf17.** $H$ and $\neg H$ belong to $S^*$, and $S^*$ is inconsistent, thus contradicting Bf3.
$S$ is inconsistent too.

As we said before, the contradiction derives from the step Bf9 and from the definitions given in Bf1 and Bf2. If we accept the definitions, then it is necessary to reject Bf9. What Benacerraf argues, unlike Lucas, is that from Gödel's theorems a confutation of Bf9c does not derive, but at the most a negation of the conjunction of Bf9 (a)-(b)-(c). Using Benacerraf's words (*1967*, p. 29):

"They [Gödel's Theorems] imply that given any Turing Machine $W_j$, either I cannot prove that $W_j$ is adequate for arithmetic, or if I am a subset of $W_j$, then I cannot prove that I can prove everything $W_j$ can. It seems to be consistent with all this that I am indeed a Turing machine, but one with such a complex machine table (program) that I cannot ascertain what it is. In a relevant sense, if I am a Turing machine, then perhaps I cannot ascertain which one. In the absence of such knowledge, I can

[27] See §2.
[28] See note 10.
[29] Benacerraf points out in his demonstration that it would have been possible to use Bf14 as an assumption instead of obtaining it from Bf9b, but this was not done in order to remain loyal to Lucas's argument steps. The use of Bf9b illustrates the fact that it is necessary for me to know how to prove that I can prove everything the $W_j$ machine can, including how to obtain $W_j$ consistency.

cheerfully go around 'proving' my own consistency, but not in an arithmetic way – not using my own proof predicate. Ignorance is bliss. Of course, I might be an inconsistent Turing Machine. Lucas's protestations to the contrary are not very convincing".

### *3.3 Chihara's Criticism*

Benacerraf's argument, therefore, solves the problem of the consistency of the system of statements produced by *TM*, that is *S**. He attains this result by limiting the discussion to those theorems which are not only derivable, but also true, that is which are *provable* by me in an *absolute* sense. However, Benacerraf himself showed that this notion leads to a contradiction without requiring the introduction of **(a)**-**(c)**: if the fact that any statement in *S* is true holds, then we arrive, still through gödelisation, at statements such as *H* and ¬*H*. This point, often moved into the background, is instead very important for the argument we want to build, and this was rightly emphasized by Charles Chihara.
Let us see, then, in a more detailed way, how this argument works in the reconstruction by Chihara (*1972*):

**BC1.** First of all, let us add to language *FA* the symbol "*S*", and the binary predicate "∈" (which in the privileged interpretation of set theory is read "belongs") and all statements of the form:

'If the numeral of Gödel's number of a formula *f* belongs to *S*, then *f*'.[30]

The intuitive sense of the latter statement is "what I can prove is true", which as we know falls within Benacerraf's definition of *S*. Let us call the new system *FA'*.

**BC2.** Having outlined the new system *FA'*, we can define its derivation predicate **B(n,m)**, which means '**n** is the Gödel's number of a derivation of the sentence whose Gödel's number is **m**'.

**BC3.** Let us then add to *FA'* the following rule, obtaining *FA''*:

'if **B(n,m)** is provable in *FA'* then it is provable that **m** belongs to *S*'[31].

That is, as Chihara writes: "what is derivable in *FA'* I can prove".[32] It is perfectly reasonable to suppose that I am able to prove everything that is derivable in *FA'*.

**BC4.** In *FA''* it is possible to build Gödel's formula. That is **m** is the Gödel's number of the formula $G_{FA''}$, which states '**m** does not belong to *S*'.

**BC5.** From BC1 and BC4 we have that: it is derivable in *FA'* that 'if **m** belongs to *S* then $G_{FA''}$'

**BC6.** By applying BC3 we have that: it is derivable in *FA'* that 'if **m** belongs to *S*, then **not**-**m** belongs to *S*'.[33]
Hence, in *FA',* '**not**-**m** belongs to *S*' is derivable.

---

[30] In symbols: $\vdash f \in S \supset f$.
[31] *Hanson 1971* criticizes this formulation since it presupposes that I have an indefinite available quantity of time to perform my demonstrations in *FA'*. However *Chihara 1971* rightly answers that in order to perform this kind of argumentations one has to presuppose a minimal idealization, that is the fact that I have all necessary available time.
[32] See *Chihara 1971* (p. 515).
[33] In symbols: $\vdash_{FA'} \mathbf{m} \in S \supset \neg(\mathbf{m} \in S)$.

**BC7.** Hence for some numeral **n**, it is derivable in *FA'* **B**(**n,m**). Therefore, using BC3 in *FA''* '**m** belongs to *S*' is derivable. But *FA''* is an extension of *FA'*, in which, as we saw in BC6, '**not**-**m** belongs to *S*' is derivable. We thus have a contradiction in *FA''*.

We can see, therefore, that without using the hypothesis that *S* is effectively enumerable we arrive at a contradiction, simply by formulating the principle that *S* uniquely contains true statements. As we shall see, the same thing will be discovered again by D. T. Chalmers (*1995*) using Gödel's second incompleteness theorem instead of the first. That is, if a formal system is sound, then it is consistent. We know that *S* is sound, so it must be consistent. If we can also derive this, then by Gödel's second theorem, *S* is inconsistent.
Such reasons lead Chihara to propose the following reformulation of Benacerraf's argument:

**C1.** Let *S'* be the set of Gödel's numbers of the statements of *FA* that I can *prove in an absolute sense*, that is that I can prove and are true. The difference from Benacerraf is that we limit *S* to *FA* statements, thus obtaining *S'*. This way, the previous demonstration BC cannot be performed. In fact, since *FA'* also contains "∈", the formula of *FA'*, whose Gödel's number is **m** need not belong to *S'*, since by hypothesis *S'* only contains Gödel's numbers of formulae belonging to *FA*.

**C2.** Let us make the hypothesis that *S'* is effectively enumerable by $TM_{S'}$.

**C3.** Let us also hypothesize that I know what $TM_{S'}$ is like. Then I will be able to build a formula *s*(**n**), which is true within *Formal Arithmetic* if and only if *n* belongs to *S'*. Remember that by **n**, we refer to the numeral of *n* in *FA*.

**C4.** Let us extend *FA* by adding all formulae of the kind:

'If $\mathbf{n}_f$ is Gödel's number of a statement *f* such that $s(\mathbf{n}_f)$ then *f*'

That is, what I can prove is true. Let us call this new formal system *FR*. In *FR* we can define the two-place derivation predicate **B**(**n,m**), which means 'the sentence which has Gödel's number **m** is derivable in *FR* by means of the proof which has Gödel's number **n**'.

**C5.** Let us then add to *FR* the rule of inference:

If for some **n** we can derive in *FR* the statement **B**(**n,m**), then in *RF* we can also derive *s*(**m**).

That is, what is derivable in *FR* can be proved. Let us call the new formal system we obtained, *FR'*.

**C6.** Within it, it is possible to build Gödel's formula. That is, **m** is the Gödel's number of the formula $G_{FR'}$, which asserts '**m** does not belong to $S'_{FR'}$'.

**C7.** By applying C4, we have that:

in *FR* we can derive, 'if **m** belongs to $S'_{FR'}$ then $G_{FR'}$'.

**C8.** By applying C6 we have that: in *FR* we can derive, 'if **m** belongs to $S'_{FR'}$, then **not**-**m** belongs to $S'_{FR'}$'. So, in *FR* we can derive, '**not**-**m** belongs to $S'_{FR'}$'.

**C9.** Hence, for some numeral **n**, we can derive in *FR* **B**(**n,m**). So, using C5 in *FR'* we can derive '**m** belongs to $S'_{FR'}$'. But *FR'* is an extension of *FR*, in which, as we have seen in C7, we can derive '**not**-**m** belongs to $S'_{FR'}$'. We thus have a contradiction in *FR'*.

**C10.** If $\mathbf{n}_f$ is the Gödel number of a statement *f* such that $s(\mathbf{n}_f)$, then *f*' means that 'if *f* is provable according to me, then *f* is true'. The fact that the Gödel number for statements of such a form belongs to *S'* follows from the fact that every statement in *S'* is true.

**C11.** 'if for some *n* the statement **B(n,m)** is derivable in *FR'*, then $s(\mathbf{m})$ is also derivable in *FR"* means 'what is derivable in *FR'* is also derivable by me'. This rule does not generate statements whose Gödel number does not belong to *S'*. This is because if there exists a derivation, I can find it.

**C12.** *S'* is, against the hypothesis, a contradictory set of Gödel numbers.

Chihara, as we can notice, avoids the contradiction in *S* by limiting the discussion to arithmetic sentences alone. In this way, to obtain the contradiction again, we must introduce hypothesis C2 that I am a Turing machine and I know which one. Chihara asserts that his argument, while having a different demonstrative procedure with respect to that of Benacerraf, it is nevertheless similar to the latter. This is because it starts from the same premises and leads to the same conclusions. In fact, the only ways to remove the contradiction are: (1) to eliminate the premise C2, i.e. that *S'* is representable through a *TM*; (2) to eliminate C3, i.e. that I know $TM_{S'}$.
On the other hand, as Chihara notes, there is a problem in his reformulation of Benacerraf's argument: step C1, which in any case was in the original argument, is not rigorous. This is because it contains the not further explicable expression "sentences of *FA* that I can prove in an absolute sense". This expression implies an involvement of my knowledge, as the notion of "absolute proof" does not simply refer to the realization of an automatic algorithm, but also to the fact that I know that the axioms of *FA* are true and that its inferential rules are valid, that is that they maintain the axioms' truth. If this is the case, what I can prove at the time $t_1$, when I start an argument, can be different from what I can prove at the time $t_2$, when I am at a subsequent step. This means that the step C5 is dubious, in that it claims that I can prove in an absolute sense everything that is derivable from *FR*. But what I can prove depends upon what I know and so, even if C1 holds, C5 does not need to hold as well.
We can then ask what would happen if we replaced C1 and C3 (as we shall see Penrose will) with:

**C1'.** Let *S'* be the set of Gödel numbers for the statements of *FA* which a *human being can in general* prove in an absolute sense; that is which are derivable by a human being and are true.

**C3'.** Let us assume by hypothesis that in general a *human being* knows what a $TM_{S'}$ is like. Then *we could* build a formula $s(\mathbf{n})$ which is true within the Elementary Arithmetic if and only if *n* belongs to *S'*. By **n** we refer to the numeral of *n* in *FA*.

Thus, any doubts over the validity of C5 are removed, even if the set *S'* defined in the step C1' is substantially that of *FA* theorems. As we shall see in the last section, Shapiro will make important objections against this reformulation too.
The conclusion of Chihara's reformulation of Lucas' argument, which had already been elaborated by Benacerraf, is not as dramatic as it was claimed by Benacerraf's friend, who argued that in that case psychology would be impossible. A position that which was put forward again by Chihara himself. This conclusion, as Benacerraf rightly observed, is indeed a "poor result". For the fact that a human being, if his intellectual abilities are representable by means of a Turing machine, will never be able to discover it, does not entail that we will not be able to represent increasingly important parts of our intelligence in computational terms.

Scientific psychology is possible; a complete scientific psychology of human intellect is impossible; but there was no doubt about this, even before the discovery of Gödel's Theorem![34].

### *3.4 Penrose's Arguments*

Despite many errors and obscurities, Lucas' argument had a profound impact on advocates of both mechanistic and anti-mechanistic positions, often contributing to the construction of reasoning, which has created more problems than they have solved. An emblematic case is the reflection of the English mathematical physicist Roger Penrose (*1989*). As is well-known, he is convinced that human intelligence, which characterizes the activity of the human mind, is not representable in terms of a Turing machine, even if it is not an activity which transcends physics, but to explain it we need a new and non-computational theory of matter. To prove this, Penrose makes use of Lucas-like, but in some respects more complex and sophisticated, arguments. In particular, as is by now accepted in the literature,[35] Penrose provides two arguments: one in *The emperor's new mind* (1989) and in the second chapter of *Shadows of the Mind* (1994), and the other in the third chapter of *Shadows of the mind*. Surprisingly such arguments, although taking into account Lucas' lesson, do not make any reference to Benacerraf and Chihara's sophisticated works.
Let us start from Penrose's first argument.[36]

**PI1.** Let us suppose that there exists a $TM_A$ capable of simulating all procedures, call them $A$, which are followed by the mathematical community to prove theorems. In addition, and this is similar to what is assumed by Benacerraf, $A$ must be *sound*, that is it must only produce true results.[37]

**PI2.** In particular $TM_A$ will be able to prove that a certain Turing machine $n$ with a certain input $m$ does not halt, that is that for certain pairs of natural numbers $n$ and $m$: for all $y$ not $T(n,m,y)$. For instance, we can assume that $TM_A$ has the following characteristic: if $TM_A$ with input $C_n$ (the encoding of $TM_n$) and $m$ halts, then $TM_n$ with input $m$ does not halt.

**PI3.** Let us apply Cantor's diagonal method. Give to $TM_n$ its own number as input; then, since $A$ is sound: if $TM_A$ with input $C_n$ and $n$ halts, then $TM_n$ with input $n$ does not halt.

**PI4.** Let us enumerate all computations, which do not halt for at least one possible input:
$C_1, C_2, C_3, \ldots C_n$,
which is an effectively enumerable set.
$TM_n$ with input $n$ will be one of them, say $C_k$. It follows that:
($TM_A$ with input $C_n$ and with $n$) = $C_n(n)$

**PI5.** Let $n=k$. We have that:
($TM_A$ with input $C_k$ and with $k$) = $C_k(k)$
and by PI3:
if $TM_A$ with input $C_k$ and $k$ stops, then $TM_k$ with input $k$ does not halt, that is $C_k(k)$ does not halt.
Hence:
If $C_k(k)$ halts, then $C_k(k)$ does not halt.

---

[34] *Gaifman 2000* reaches similar conclusions independently and remarks on the affinity between this result and the thesis of substantial inaccuracy of psychology advocated by Davidson. *Chihara 1971* (p. 518), on the other hand, reports that Benacerraf became sceptical about the soundness of his argument.
[35] *Chalmers 1995* and *McCullough 1995* first argued the presence in Penrose of two arguments.
[36] See *Penrose 1994* (p.73 ff.).
[37] If Penrose had taken into account Benacerraf and Chihara's works, which we discussed before, he would have understood that PI1 leads to a contradiction without the hypothesis that $A$ is computable.

This means that:
$C_k(k)$ does not halt.

**PI6.** But $C_k(k)$ is equal to $TM_A$ with input $C_k$ and $k$, which does not halt, that is $TM_A$ is not able to "realize" that $C_k(k)$ does not halt, against the hypothesis. We know, however, that $C_k(k)$ does not halt, because $A$ is sound; hence we know something that $TM_A$ is not able to calculate. Penrose (*1994*, p. 76) concludes that: 'human mathematicians are not using knowably sound algorithms in order to ascertain mathematical truths'.

This first argument, as well as Lucas' paper, revitalized the debate on mechanism and Gödel's theorems, and so provoked several reactions, many of which concerning the procedure of Penrose's argument: in particular the works by George Boolos (*1990*), Martin Davis (*1990*; *1993*), Hilary Putnam (*1994*) and Solomon Feferman (1996)[38] helped in clarifying many incorrect, or at least doubtable, aspects of the above argument.[39] Beyond criticisms answered by Penrose with his work *Beyond the doubting of a shadow* in 1996, it is to be noted that the above argument improves the previous ones of Lucas, Benacerraf and Chihara in at least one aspect. This is because when it defines $A$, which replaces Benacerraf' $S$, it does not refer to just one person, but to all human mathematicians, thus bypassing Chihara's criticisms, which incidentally were probably unknown to him. In his argument Penrose arrives at a fork similar to that of Benacerraf.

A *sceptical hypothesis*: mathematical methods of proof are not all contained in one algorithm;
An *agnostic hypothesis*:[40] mathematical methods of proof are all contained in a sound algorithm, which however human beings will never know with absolute certainty. In either alternatives it must be stressed that the problem is not hardware, but software. In practice, the fork opened by the previous variants is equivalent to the premise (b) of Benacerraf's B2 point and to the first part of Chihara's C3 premise.
The presence of such alternatives explains those parts of *Shadows of the mind* aimed at unravelling the fork in favour of the sceptical position, and sets out the reasons for developing a second Penrosian argument.

In *Shadows of the mind*, in fact, as we have already emphasized, Penrose returns to his Gödelian argument refining its previous version. Penrose's new argument presents the same problems as the first one, but it seems to advance some interesting ideas. We will critically discuss it using the penetrating analysis of the philosopher of the mind David Chalmers (*1995*), who however, not knowing Benacerraf's and Chihara's papers, cannot see the possibility, highlighted by the latter, of removing the contradiction.
Chalmers reconstructs Penrose's first argument as follow:

**PI1'.** We know that we are a sound formal system, i.e. that our reasoning abilities can be simulated by a *TM* which only produces true statements.

---

[38] In various papers Feferman intervenes on the issue. In particular in *Feferman 2007* he describes in detail Godel's reflection on the significance of incompleteness theorems for the philosophy of mind. He also talks about the difficult relationship between Nagel and Gödel concerning the intention of publishing a popular booklet written by the former in the late 1950's. Finally he authoritatively takes part in the controversy over Penrose's argument. In his first argument, Feferman notices a series of logical inaccuracies and so argues that he is not convinced of Penrose's conclusion, for such arguments raise more problems than they solve. However he is persuaded that it is not possible to find a possible algorithm that reproduces mathematicians' methods of proof; even if it is still possible to re-represent the proof in mechanical terms. In the end it is the same conviction as Benacerraf's, which however Feferman does not justify on the basis of the incompleteness theorems.
[39] For a general examination on criticisms on Penrose's arguments see *Antonelli 1997*.
[40] For a wider reflection on concepts such as "sceptical hypothesis" and "agnostic hypothesis" see *Bruni 2004*.

**PI2'.** We also know *a priori* that *F* represents our reasoning abilities, that is we know a priori which system represents us.

**PI3'.** Thus, we know that *F* is sound. Hence we know that it is part of *F* that *F* is consistent.

**PI4'.** *F* cannot derive its Gödel statement $G_F$. *F*, however, can derive the conditional 'if *F* is consistent, then $G_F$ is true'.

**PI5'.** Thus, through *modus ponens F* knows that its Gödel's statement is true, hence *F* is not representable by means of a *TM*, or rather, is not a formal system.

As we know, the whole weight of the argument lies on assumption PI2', which claims that we know a priori *F* represents our reasoning abilities. Let us remember that this is exactly what Benacerraf denied. It is thus not possible to prove that we are not a *TM*, unless we show that PI2' holds. Nor can Penrose claim that we can empirically ascertain that we are *F*, in that, despite this being entirely legitimate (as Gödel already observed), it would not lead to the desired conclusion.
To arrive at the conclusion that we are not a Turing Machine, Penrose cannot assume the very weak premise PI2'. Nevertheless he could reason in the following way:

**PI1''.** If we are a sound formal system *F*, then we are able to establish the soundness of *F*.

**PI2''.** We are a sound formal system *F*. This is the crucial premise, whose negation Penrose would show.

**PI3''.** We are able to establish the soundness of *F*. Through *modus ponens* from PI1'' and PI2''.

**PI4''.** We are able to establish the consistency of *F*. It follows from PI3''.

**PI5''.** PI4'' contradicts PI2'', since, according to second Gödel's theorem, a sound formal system is not able to prove its consistency.

Therefore we can infer PI5'' only if *F* autonomously establishes that it is sound and therefore consistent. Hence, to render PI5'' plausible, Penrose must prove that *F* is sound without us knowing that we are *F*. The attempt to do so explains the section § 3.3 section of *Shadows of the mind*, where Penrose argues that any formal system is representable as a set of axioms and inferential rules. And so if we examine *F*, even if we do not know that *F* represents our argumentative abilities, we see that the axioms are true and that the inferential rules are valid, that is they lead to true theorems. Chalmers sharply criticizes this argument on grounds that the formal system which represents our abilities is so complicated that, in all likelihood, we do not have the chance to examine every single part of it. Hence also this version of Penrose's first argument fails.
Penrose's new argument, as we said, tries to improve on the previous one, and does so by replacing PI2' with the following assertion:

**PII2.** We know *a posteriori* that *F* represents our reasoning abilities.

But, by

**PII1.** We know that we are a sound formal system.

**PII3.** Hence we know *a posteriori* that *F* is sound. Let us, then, build the system *F'* which also includes that 'I am *F*'.

**PII4.** *F'* is certainly sound, hence by Gödel's first theorem, *F'* cannot derive its Gödel statement $G_{F'}$; but I know that $G_{F'}$ is true, hence I am not *F'*. If I am not *F'*, all the more so, I am not *F*, against the hypothesis.

Notice that PII2 is not sufficient to create the contradiction, since we arrive at the conclusion that *F* represents our intellectual abilities only *a posteriori*. But by adding the sentence 'I am *F*' to the system *F* nonetheless we create the contradiction.
But, as Chalmers observes, PII1 is a contradictory premise.[41] We would say: "They realized at last!", since it had already been noticed by Benacerraf in the appendix to his 1967 article, that is thirty years before.
Chalmers then concludes that this first premise is to be removed, that is we cannot know that we are a sound formal system. However, we have already seen the elegant way in which Chihara bypassed the problem in 1972, reducing the whole discussion to arithmetic statements alone.

### *4. Gödel's view*

In 1951 Gödel held one of the prestigious *Gibbs Lectures* for the American Mathematical Society. The title of his lecture was *Some basic theorems on the foundations of mathematics and their implications.* The theorems in question were precisely those of incompleteness, and the philosophical implications concerned the nature of mathematics and the abilities of the human mind.[42]
This was one of the few official occasions in which Gödel expounded his opinion on the philosophical implications of his theorems. Without going into detail about Gödel's paper, what is interesting here is the first part, which is devoted to the derivation of the thesis of essential incompleteness of mathematics from his famous theorems.
Such a thesis was, for Gödel particularly, sanctioned by the second theorem, in fact:

> It [the second theorem] makes it impossible that someone should set up a certain well defined system of axioms and rules and consistently make the following assertion about it: All of these axioms and rules I perceive (with mathematical certitude) to be correct, and moreover I believe that they contain all of mathematics. If someone makes such a statement he contradicts himself. For if he perceives the axioms under consideration to be correct, he also perceives (with the same certainty) that they are consistent. Hence he has a mathematical insight not derivable from his axioms. (*Gödel 1995*, p. 309)

Let us try to better understand Gödel's argument.
Gödel's idea is that if someone perceives with absolute certainty that a certain formal system[43] is correct (sound), he will also know the consistency of the system, that is he will know the truth of the system statement which establishes the consistency of the system itself. But, by Gödel's second theorem, the formal system considered cannot prove its own assertion of consistency, therefore the system does not capture all arithmetical truths, and for this reason "if someone makes such a statement he contradicts himself". But what does all of this mean? Does it mean perhaps that a well defined system of correct (sound) axioms cannot contain all that is strictly mathematical?
Gödel believes that such a question has two possible answers:

---

[41] *Penrose 1996* argues against Chalmers that he does not add to *F* the sentence 'I know that I am *F*', but simply 'I am *F*', without noticing that it is exactly the fact that the sentence 'I am *F*' is added to *F* which makes certain that I know I am *F*.
[42] See *Gödel 1995*. A very accurate analysis of this writing is proposed by: *Feferman 2006*; *Tieszen 2006*; *van Atten 2006*.
[43] It is understood that, in this paper, the expression "formal system" indicates a formal system which is adequate to derive incompleteness theorems.

It does, if by mathematics proper is understood the system of all true mathematical propositions; it does not, however if one understands by it the system of all demonstrable mathematical propositions. […] Evidently no well-defined system of correct axioms can comprise all [of] objective mathematics, since the proposition which states the consistency of the system is true, but not demonstrable in the system. However, as to subjective mathematics it is not precluded that there should exist a finite rule producing all its evident axioms. However, if such a rule exists, we with our human understanding could certainly never know it to be such, that is, we could never know with mathematical certainty that all the propositions it produces are correct; or in other terms, we could perceive to be true only one proposition after the other, for any finite number of them. The assertion, however, that they are all true could at most be known with empirical certainty, on the basis of a sufficient number of instances or by other inductive inferences. If it were so, this would mean that the human mind (in the realm of pure mathematics) is equivalent to a finite machine that, however, is unable to understand completely its own functioning. This inability [of man] to understand himself would then wrongly appear to him as its [(the mind's)] boundlessness or inexhaustibility. (*Gödel 1995*, pp. 309-310)

Not only, then, does the previous question pose the problem of the inexhaustibility or incompleteness of mathematics considered as the totality of all true mathematical propositions; but it also raises the question as to whether mathematics is in principle inexhaustible for the human mind, that is to say, whether the human mind's demonstrative abilities are extensionally equivalent to a certain formal system, or to the *TM* connected to it (the *TM* which enumerates the set of theorems of the corresponding formal system).

The question, then, requires due consideration of precisely the relation between what Gödel calls *objective* and *subjective mathematics*. First let *T* be the set of mathematical truths expressible within the first-order arithmetic, and call this "objective arithmetic", or following Gödel, spell it "objective mathematics", that is "the body of those mathematical propositions which hold in an absolute sense, without any further hypothesis". By Tarski's theorem *T* is not definable within the language of arithmetic, hence *T* is not recursively enumerable. Let us then define *K* as the set of arithmetical statements which a human being can know and prove absolutely and with mathematical certainty, that is what he can derive[44] and know to be true. Let us call it "subjective arithmetic" or following Gödel "subjective mathematics", which "consists of all those theorems whose truth is demonstrable in some well-defined system of axioms all of whose axioms are recognized to be objective truths and whose rules preserve objective truth".[45]

What is then the relation between *K* and *T*?

Quoting Feferman we could synthesize Gödel's answer by saying: if *K* was equal to *T* "then demonstrations in subjective mathematics [were] not confined to any one system of axioms and rules, though each piece of mathematics is justified by some such system. If they do not, then there are objective truths that can never be humanly proved, and those constitute absolutely unsolvable problems".[46] That is, if the equivalence $K=T$ held, the human mind would not be equivalent to any formal system or *TM* connected to it. In fact, having established *T* characteristics, for each formal system there would be a provable statement about the human mind, but not within the formal system. Hence, the mechanism would certainly be false: *T* non-recursive enumerability entails, in fact, the non-existence of any effective deductive system whose theorems are only and all truths of arithmetic.

If, on the contrary, *K* did not coincide with *T*, and thus the human mind was equivalent to a given formal system or to the *TM* related to it, the existence of arithmetic statements humanly undecidable

[44] As *Feferman 2006* (p. 140) emphasizes, Gödel believes that "the human mind, in demonstrating mathematical truths, only makes use of evidently true axioms and evidently truth preserving rules of inference at each stage".

[45] *Feferman 2006* (pp. 135-136).

[46] *Feferman 2006* (pp. 136-137). The expression "absolutely unsolvable problems", or Gödel's expression "Diophantine problems which are undecidable" refers to the following fact: Gödel's unprovable proposition which expresses the consistency of a formal system within the same system (with the formal system satisfying the first incompleteness theorem hypothesis) has the form $\forall(x)R(x)$, where *R* is a primitive recursive predicate and each statement of such a form is equivalent (Gödel proved it) to a statement of the form $\forall x_1, \dots, \forall x_n \exists y_1, \dots, \exists y_m$ [$p(x_1, \dots, x_n, y_1, \dots, y_m) = 0$] where the variables vary on natural numbers, and "p" is a polynomial with integer coefficients, that is it has the form of those *problems* faced by the Greek mathematician Diophantus of Alexandria in his book *Arithmetica*.

in an absolute sense would follow. In fact, as Gödel underlined, the second incompleteness theorem does allow this conclusion: the proposition expressing the consistency of *K*, say $Con_K$, is true but is not provable within the system itself; the negation of $Con_K$ is false and is not provable in *K*. Having established the equivalence between human mind and formal system, $Con_K$ is not even provable by the human mind. Finally, since $Con_K$ can be put in the form of a Diophantine problem[47] it is an absolutely undecidable problem. Such a proposition is, thus, an unknowable truth. Such questions and arguments lead Gödel to the idea that from the incompleteness results can at the most be derived the following disjunction: "Either [subjective] mathematics is incompletable in this sense, that its evident axioms can never be comprised in a finite rule, that is to say, the human mind (even within the realm of pure mathematics) infinitely surpasses the powers of any finite machine, or else there exist absolutely unsolvable diophantine problems of the type specified (where the case that both terms of the disjunction are true is not excluded, so that there are, strictly speaking, three alternatives)" (*Gödel 1995*, p. 310).

So, considering the translatability between the concept of a well defined formal system and that of a Turing Machine, we can say that Gödel's theorems leave open the three following possibilities:[48]

(I) human intelligence infinitely surpasses the powers of the finite machine (TM), and there are no absolutely irresolvable Diophantine problems.[49]
(II) human intelligence infinitely surpasses the powers of the finite machine (TM) and there are absolutely unsolvable Diophantine problems. That is, although human intelligence is not a finite machine, nevertheless there are absolutely irresolvable Diophantine problems for it.
(III) human intelligence is representable through a finite machine (TM) and there are absolutely irresolvable Diophantine problems for it.

Gödel was convinced that (I) held, but he was also aware that his incompleteness theorems did not make the existence of a mechanic procedure equivalent to human mind impossible.
Gödel, however, as we expounded, believed that from his theorems it followed that if a similar procedure existed we "with our human understanding could certainly never know it to be such, that is, we could never know with mathematical certainty that all the propositions it produces are correct". But this, established Godel's idea that "the human mind, in demonstrating mathematical truths, only makes use of evidently true axioms and evidently truth preserving rules of inference at each stage", this exactly means that "the human mind (in the realm of pure mathematics) is equivalent to a finite machine that, however, is unable to understand completely its own functioning".
This argument, as it can be noticed, reminds those already presented by Benacerraf and Chihara. Let us try to analyse it further by means of a formulation partly provided by Shapiro.
First let *K* be the set of all those arithmetic sentences (theorems) whose truth is provable within some well defined axiomatic systems whose totality is recognized as objective truth and whose rules preserve objective truth. Moreover, Let *T* be the set of mathematical truths expressible within first-order arithmetic:

**G1.** Let us hypothesize that *K* is effectively enumerable and that *e* is the *TM* number enumerating it.

**G2.** *K* is equal to the sentences generated by $TM_e$, call this set $W_e$, hence $K = W_e$. Let us hypothesize that '$K = W_e$' is provable in *K,* that is it belongs to $W_e$.

[47] See note 46.
[48] See *Tieszen 2006.*
[49] See note 46.

**G3.** Let us hypothesize, moreover, that everything within $W_e$ is effectively *provable*, that is all elements of $W_e$ satisfy Hilbert and Bernays' famous derivability conditions, reformulated by M. H. Löb:[50]

**(i)** For each statement *f* in the arithmetical language, if *f* belongs to $W_e$, then the arithmetical statement '*f* belongs to $W_e$' belongs to $W_e$ too. (That is to say $W_e$ knows that it contains *f*, i.e. each theorem has to be provable).

**(ii)** For each statement *f* and *g* in the language of arithmetic, arithmetical statements like 'if *f* entails *g* it belongs to $W_e$, then, if *f* belongs to $W_e$, *g* belongs to $W_e$' belong to $W_e$. (What $W_e$ knows is closed under the *modus ponens*, that is such a rule holds for the provability predicate).

**(iii)** For each statement *f* of arithmetical language, arithmetical sentences like 'if *f* belongs to $W_e$, then it belongs to $W_e$ that *f* belongs to $W_e$', belongs to $W_e$. (That is $W_e$ knows that it knows that it contains *f*, that is that provability is provable).

**G4.** It is possible to prove[51] that for $W_e$ corresponding formal system the next condition holds (*Diagonalization Lemma*):

**(iv)** For each formula $F(x)$ of the formal system language, where $x$ is a free variable, there is a statement *f* of the formal system language such that $\vdash f \leftrightarrow F(\boldsymbol{f})$ .

**G5.** It is possible to prove that if for $W_e$ corresponding formal system the condition (iv) holds; the usual classic inferential forms (that is a $\alpha \supset \beta, \beta \therefore \alpha$ ; $\alpha \supset \beta, \beta \supset \gamma \therefore \alpha \supset \gamma$ ; $\alpha \supset (\beta \supset \gamma)$, $\alpha \supset \beta \therefore \alpha \supset \gamma$) hold, and Löb's conditions (i)-(ii)-(iii) hold, then the following *Löb's Theorem* holds as well:[52]

Let *f* be any sentence in the language of first-order arithmetic and **B**[53] the usual provability predicate for the formal system corresponding to $W_e$,

'**B**(***f***) entails *f*' belongs to $W_e$ if and only if '*f* belongs to $W_e$'.[54]

**G6.** Let $Con_e$ be the statement generated by $TM_e$ 'there does not exist a **y** such that **B**(**y**,**n**)' where **n** is Gödel's number for the statement '*1=0*'. Practically, $Con_e$ expresses the consistency of the set of sentences generated by $TM_e$, which we have called $W_e$. By the assumption G2, $Con_e$ is true, *K* being the system of all arithmetical sentences whose truth is derivable by the human being in some well defined system of axioms and rules. $Con_e$ however cannot be in *K* because of Gödel's second theorem. But neither can the negation of $Con_e$. Hence $Con_e$ is true, but unknowable, that is absolutely undecidable. If hypotheses G1 and G2 hold and so do G3, G4 and G5, nobody can know about *e* that $W_e$ is consistent. It follows that no human being could know that each sentence in $W_e$ is true (that is that $W_e \subseteq T$), since it should know that $W_e$ is consistent. If, in fact, we suppose that $Con_e$

---

[50] See, for instance, Verbrugge. See especially *Löb 1955*; *Boolos 1993*. A very interesting article on these issues is: *Detlefsen 2002*.

[51] For demonstrations and further clarifications see *Smullyan 1992* (VIII and IX).

[52] See note 52.

[53] In particular we can define a derivation predicate **B(n,m)**, which means '**n** is the Gödel's number of a derivation of the sentence whose Gödel's number is **m**'.

[54] Quoting *Shapiro 1998* p. 281: "That is, there is no unknowable sentence Φ [*f*] such that we can know that if Φ [*f*] is in $W_e$ then Φ [*f*] is true. In other words, there is no trivial hypothetical knowledge about the contents of $W_e$. By hypothesis, a sentence Φ [*f*] is knowable if and only if it is in $W_e$. For a particular sencence Φ [*f*], we can *know that* 'Φ is knowable if and only if it is in $W_e$' only if Φ [*f*] is knowable". See also *Detlefsen 2002*.

belongs to $W_e$, then by Löb's first condition '*Con*$_e$ belongs to $W_e$' belongs to $W_e$, but because of the usual definition of negation '*non-Con*$_e$ belongs to $W_e$ entails '*0=1*'' belongs to $W_e$. It follows that, by Löb's theorem, we have '*0=1*' belongs to $W_e$. But this is not possible because *K* is consistent (Gödel's second theorem).

**G7.** So, either we rule out G1, i.e. that human arithmetical abilities are reproducible by a *TM*, and therefore we accept that 'the human mind [...] infinitely surpasses the powers of any finite Machine'; or, if we accept G1, we have to rule out G2, i.e. that we can know which this *TM* is. Paraphrasing Shapiro: even if the mechanist was right and there was a system capable of *K*, nobody could claim to know with mathematical certainty that the system axioms and rules are correct (sound). In other words: there would not be any *TM* such that we could know that *TM* enumerates all and only the knowable statements.

This is a very similar position to that of Benacerraf and Chihara: if there exists the *TM* which represents true and human derivable arithmetic truths, it cannot be known. As one can notice, not only was Gödel very careful to maintain that we cannot know *with mathematical certainty* that a certain *TM* represents *K*; but he was also just as accurate in seeing that, despite of this, we could use other methods, for example of an empirical nature, to pursue such a knowledge: "We could perceive to be true only one proposition after the other, for any finite number of them. The assertion, however, that they are all true could at most be known with empirical certainty, on the basis of a sufficient number of instances or by other inductive inferences".

### *5. Some Final Considerations*

As has rightly been pointed out by Shapiro (*1998*),[55] a fundamental issue of the debate considered here is that it is not quite clear what the exact content of the mechanistic view should be. Indeed all authors analyzed here have dealt with the issue of defining this content, either refuting or valorizing it. Despite this, from these different views it clearly emerges that whatever the content, both the mechanist and anti-mechanist need to set *idealizations* without which it would not be possible to make any analysis and comparison concerning it. Quoting Shapiro (*1998*, p. 275): "The mechanist claims that there can be a machine whose outputs are the same as those of a human or a group of humans. What sort of machine? What outputs? What aspect of what humans? […] Things get interesting only when we idealize, but things also get murky". The same, *mutatis mutandis*, could be said for the anti-mechanist. Without going into details, for which one can refer to Shapiro's work, here we wish to linger over a part of the issue of idealization, noting that on the one hand both Benacerraf's *S* set and Penrose's *A* cannot have a finite cardinality, while on the other hand, human life being finite, the set of procedures and theorems of a group of mathematicians cannot but be finite as well.

Benacerraf's and Penrose's sets, clearly presuppose an idealization, namely the one of the set of theorems which mathematicians *can* prove. If now we consider a finite set of theorems proved by mathematicians, it must be stressed that, however large it might be, it will never determine a univocal set of rules, that is an algorithm, which should produce them. Using now Saul Kripke's wittgensteinian considerations,[56] this is equivalent to saying that no finite set of theorems determines a single algorithm which produces them. But if this is true, what is the point of speaking about the algorithm which produces all arithmetical theorems, which a mathematician community could produce if it had an indefinite amount of available time? One can argue that an assumption of any discussion concerning mechanism is the one that might be called "minimal Platonism". As is well-known, a somewhat caricatured picture of Platonism circulating in the field of mathematics would be like this: long before the first arithmetician realized that '2+2=4', beyond space and time,

[55] Similar views are shared by *Tamburrini 2002* (pp. 130-133).
[56] See *Kripke 1982*; *Wittgenstein 2001*.

there existed entities like "2" and "4", which were already in that relation. This is obviously an unjustified and groundless metaphysical hypothesis.[57] However, as Quine and Putnam have pointed out, introducing abstract entities explains the objectivity of mathematized science. Therefore we need to attribute some reality to such entities, at least within the context of their application, by abduction, that is by a sort of inference to best explanation.[58] Yet, without however introducing a sort of Platonism on entities, in order to answer the previous question, one could argue that mathematics bears a certain *normativity*, which can be expressed by statements like: any thinking being which would be able to perform the abstractions and idealizations necessary to grasp the concepts of "2", "4" and "addition" would realize that '2+2=4'. Thus, here the point is not so much to support a Platonism of entities, as to support a *Platonism of procedures*.
On this basis, beside a *merely descriptive level*, it makes sense with regard to arithmetic to speak of a *normative level*: the set of mathematicians who work for an indefinite time will produce theorems in accordance with a normativity, which, if the mechanism is right, is reproducible by means of an algorithm. By introducing this normativity we can further develop the premise C1' of Chihara's argument without referring to the real *performances* of mathematicians, but rather to their *ideal arithmetical competence*, obtaining an improvement in Chihara's argument. Remember that Chihara criticized the premises C1 and C3 of his reconstruction of Benacerraf's argument; for this reason, we replaced them with C1' and C3' following Penrose's suggestion. We have already seen that Shapiro criticizes these too from a sheer empiricist point of view. This is why in our reformulation we will introduce a minimal Platonism needed to reach the conclusion.

**CFG1.** Let *S'* be the set of Gödel numbers of sentences of *FA* that *a set of mathematicians can prove in an absolute sense and in compliance with a normativity*.

**CFG2.** Let us hypothesize that *S'* is effectively enumerable.

**CFG3.** Let us also hypothesize that the *human being* knows what a $TM_{S'}$ looks like. Then one *could* build $s(\mathbf{n})$ which is true in *N* if and only if *n* belongs to *S'*. By **n** we indicate the *n* numeral in *FA*.

**CFG4.** Let us extend *FA* by adding all formulae such as:

If $\mathbf{n}_f$ is the Gödel number of a sentence *f* such that $s(\mathbf{n}_f)$ then *f*

Let us call this new formal system *FR*. It is clear that in *FR* we can define the two-place derivation predicate **B**(**n**,**m**), which means 'the statement which has Gödel number **m** is derivable in *RF* by means of a proof which has Gödel number **n**'.

**CFG5.** Let us then add to *FR* the inference rule:

if for any ***n*** one can derive in *FR* the sentence **B**(**n**,**m**), *then* in *FR* one can also derive $s(\mathbf{m})$.

Let us call the newly obtained formal system *FR'*.

**CFG6.** In it, it is possible to build the Gödel formula. That is, **m** is the Gödel number of the formula $G_{FR'}$, which states that '**m** does not belong to $S'_{FR'}$'.

**CFG7.** By applying CFG4 we obtain that:

in *FR* it is derivable that, 'if **m** belongs to $S'_{FR'}$ then $G_{FR'}$'.

---

[57] See, for instance, the classic *Benacerraf 1996*.
[58] This is the famous "indispensability argument", see Colyvan http://plato.stanford.edu/entries/mathphil-indis/.

**CFG8.** By applying CFG6 we have that:
in *FR* it is derivable that 'if **m** belongs to $S'_{FR'}$, then **not**-**m** belongs to $S'_{FR'}$'. Hence in *FR* it is derivable that '**not**-**m** belongs to $S'_{FR'}$.

**CFG9.** Hence, for some numeral **n**, in *FR* '**B**(**n**,**m**)' is derivable. So, by using CFG5 in *FR'* we can derive '**m** belongs to $S'_{FR'}$'. But *FR'* is an extension of *FR*, in which, as we have seen in CFG7, we can derive '**not**-**m** belongs to $S'_{FR'}$'. We thus have a contradiction in *FR'*.

**CFG10.** 'If $\mathbf{n}_f$ is the Gödel number of a sentence *f* such that $s(\mathbf{n}_f)$, then *f*' means that 'if *f* is provable by the group of mathematicians, then *f* is true'. That the Gödel number of sentences of this form belongs to *S* follows from the fact that all sentences comprised in *S'* are true.

**CFG11.** 'If for some *n* the sentence **B(n,m)** is derivable in *FR*, then *s*(**m**) is also derivable in *FR*' means 'what is derivable in *FR* is also derivable by the group of mathematicians'. This rule does not produce statements whose Gödel number does not belong to *S'*, because if there exists a derivation, mathematicians can find it.

**CFG12**. Hence, if *FR'* is contradictory, and $S'_{FR'}$ is a subset of *S'*, then *S'* represents a contradictory set of statement, against *S'* definition. So, the only ways to avoid the contradiction are: (1) by removing premise CFG2, i.e. the sentence that *S'* is representable by a *TM*; (2) by removing CFG3, i.e. the statement according to which the human being knows $TM_{S'}$.

For reasons of principle, therefore, we cannot know with absolute certainty whether or not a formal system representable by *TM* captures our reasoning abilities. This conclusion, already highlighted by Gödel, and proposed again by both Benacerraf and Chihara, does not have any great relevance to the philosophy of psychology. Nothing prevents one from building computational models, which would simulate ever-increasing parts of our intelligent behaviour. One day, we could even build a Turing machine, which will simulate in every way human intelligent behaviour, but we will not know this with absolute certainty! We believe, then, that the significance of this conclusion is more anthropological, than scientific: it simply reasserts the fundamental incompleteness of human self-knowledge.

## Acknowledgements

Part of this work was presented at a seminar held at the 'Scuola Normale Superiore di Pisa', and in lectures given at the 'Università degli Studi di Urbino Carlo Bo'. We are grateful to Gabriele Lolli, Massimo Mugnai, Angelo Vistoli, Mario Piazza, Carlo Cellucci, Stefano D'Egidio, Maurizio Colucci for their suggestions and questions, which helped us to improve the paper.

## References

Antonelli G. A. 1997. 'Gödel, Penrose e i fondamenti dell'intelligenza artificiale', *Sistemi Intelligenti*, 9, 353-376; also available at http://orion.uci.edu/~aldo/papers/penrose.pdf.

Benacerraf P. 1967. 'God, the devil and Gödel', *The Monist*, **51**, 9-32; also available at http://www.univ.trieste.it/~etica/2003_1/3_monographica.htm.

Benacerraf P. 1996. 'Mathematical truth', in W. D. Hart, *The philosophy of mathematics*, Oxford: Oxford University Press, pp. 14-30.

Berto F. 2008. *Tutti pazzi per Gödel*, Bari: Laterza.

Berto F. 2009. *There's Something About Gödel!: The Complete Guide to the Incompleteness*

*Theorem*, Oxford: Wiley-Blackwell.
Boolos G. 1990. 'On seeing the truth of the Gödel sentence', *Behavioral and Brain Sciences*, **13**, 655-656.
Boolos G. 1993. *The Logic of Provability*, Cambridge: Cambridge University Press.

Boyer D. L. 1983 'J. R. Lucas, Kurt Gödel and Fred Astaire', *The Philosophical Quarterly*, **33**, 147-159.
Bruni R. 2004. 'Riflessioni sull'incompletezza. I teoremi di Gödel tra logica e filosofia', Ph.D. Thesis, Firenze: Università degli Studi di Firenze; also available at: http://www.philos.unifi.it/CMpro-v-p-88.html.
Chalmers D.J. 1995. 'Minds, machines, and mathematics', *Psyche*, **2**, 11-20; also available at the: http://psyche.cs.monash.edu.au/v2/psyche-2-09-chalmers.html
Chihara C. S. 1971. 'On alleged refutations of mechanism using Gödel's incompleteness results', *The Journal of Philosophy*, **69**, 507-526.
Colyvan M. 'The indispensability argument in philosophy of mathematics', http://plato.stanford.edu/entries/mathphil-indis/.
Davis M. 1990. 'Is mathematical insight algorithmic?', *Behavioral and Brain Sciences*, **13**, 659-660.
Davis M. 1993. 'How subtle is Gödel's theorem? More on Roger Penrose', *Behavioral and Brain Sciences*, **16**, 611-612.
Dennett D. C. 1972. 'Review of The freedom of the will', *The Journal of Philosophy*, **69**, 527-531.
Detlefsen M. 2002, 'Löb's theorem as a limitation on mechanism', *Minds and Machines*, **12**, 353–381.
Feferman S. 1962. 'Transfinite recursive progressions of axiomatic theories', *Journal of Symbolic Logic*, **27**, 383-390.
Feferman S. 1996. 'Penrose' Gödelian Argument', *Psyche*, **2**, 21-32.
Feferman S. 2006. 'Are There Absolutely Unsolvable Problems? Gödel's Dichotomy', *Philosophia Mathematica*, **14**, 134–152.
Feferman S. 2007. "Gödel, Nagel, minds and machines", http://math.stanford.edu/~feferman/papers/godelnagel.pdf.
Franzen T. 2005. *Gödel's Theorem: an Incomplete Guide to its Use and Abuse*, Wellesley: A K Peters.
Frixione M. and Palladino M. 2004. *Funzioni, Macchine, Algoritmi*, Rome: Carocci.
Gaifman H. 2000. 'What Gödel's incompleteness results does and does not show', *Journal of Philosophy*, vol. **97** (**8**), 462-470.
Gödel K. 1931. 'Über formal unentscheidbare Sätze der Principia mathematica und verwandter Systeme', *Monatshefte für Mathematik und Physik*, **38**, 173-198; also in Gödel K. 1986. *Collected Works*, I, Oxford: Oxford University Press, pp. 144-195.
Gödel K. 1995 'Some basic theorems on the foundations of mathematics and their implications', in *Collected Works,* III, Oxford: Oxford University Press, pp. 304-335
Hanson W. H. 1971. 'Mechanism and Gödel's theorem', *British Journal for the Philosophy of Science*, **22**, 9-16.
Hofstadter D. 1979. *Gödel, Escher, Bach: an eternal golden braid*, New York: Basic Books.
Hofstadter D. R. and Dennett D. 1981. *The mind's I*, New York: Basic Book.
Kemeny G. 1959. *A philosopher's look at science*, Princeton: van Nostrand.
Kleene S. C. 1967. *Mathematical logic*, New York: John Wiley & Sons.
Kripke S. 1982. *Wittgenstein on rules and private language*, Cambridge: Harvard University Press.
Löb M. H. 1955. 'Solution to a problem of Leon Henkin', *Journal of Symbolic Logic*, **20**, 115-118.
Lucas J.R. 1961. 'Minds, machine and Gödel', *Philosophy*, **36**, 112-127; also available at: http://users.ox.ac.uk/~jrlucas/mmg.html.
Lucas J.R. 1968. 'Satan stultified: a rejoinder to Paul Benacerraf', *The Monist*, **52**, 145-158; also

available at: http://www.univ.trieste.it/~etica/2003_1/4_monographica.htm.
McCullough D. 1995. 'Can Humans escape Gödel?', *Psyche*, **2-1**; also available at: http://psyche.cs.monash.edu.au/v2/psyche-2-04-mccullough.html.
Newman J. R. and Nagel E. 1958. *Gödel's proof*, New York: New York University Press.
Odifreddi P. 1992. 'Il teorema di Gödel e l'I.A.', *La Rivista dei Libri*, **II**, June, 37-39.
Odifreddi P. 2003. *Il diavolo in cattedra. La logica da Aristotele a Gödel,*Torino: Einaudi.
Palladino D. 2004. *Logica e Teorie formalizzate. Completezza, Incompletezza, Indecidibilità*, Rome: Carocci.
Penrose R. 1989. *The 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*, **2**, 89-129; also available at: http://journalpsyche.org/ojs-2.2/index.php/psyche/article/viewFile/2409/2338.
Putnam H. 1961. 'Minds and Machines', in S. Hooks, *Dimensions of mind*, New York: Collier, pp. 148-179.
Putnam H. 1994. 'Review by R. Penrose. Shadows of the mind, Oxford University Press, 1994', *Bulletin of the American Mathematical Society*, **32**, 370-373.
Robinson R. M. 1950. 'An Essentially Undecidable Axiom System' in *Proceedings of the International Congress of Mathematics*, 729-730.
Rosenbloom P. 1950. *Elements of Mathematical Logic*, New York: Dover.
Shapiro S. 1998. 'Incompleteness, mechanism, and optimism', *The Bulletin of Symbolic Logic*, **4**, 273-302.
Smart J. J. C. 1961. 'Gödel theorem, Church's theorem and mechanism', *Synthèse,* **13**, 105-110.
Smullyan R. 1992. *Gödel's Incompleteness Theorems*, Oxford: Oxford University Press.
Tamburrini G. 2002. *I matematici e le macchine intelligenti*, Milan: Bruno Mondatori.
Tieszen R. 2006. 'After Gödel: mechanism, reason, and realism in the philosophy of mathematics', *Philosophia Mathematica*, **14**, 229–254.
Turing A. 1936. 'On computable numbers, with an application to the Entscheidungsproblem', *Proceedings of the London Mathematical Society*, **42**, 230-265.
Turing A. M. 1992. 'Lecture to the London Mathematical Society on 20 February 1947' in D. C. Ince, *The collected works of A. M. Turing*, vol II, Amsterdam: North Holland, pp. 87-105.
van Atten M. 2006. 'Two draft letters from Gödel on self-knowledge of reason', *Philosophia Mathematica*, **14**, 255-261.
Verbrugge R. "Provability logic", http://www.seop.leeds.ac.uk/entries/logic-provability/
Wittgenstein L. 2001. *Philosophical investigations,* Oxford: Blackwell.
Webb J. 1980. *Mechanism, Mentalism and Metamathematics*: *an essay on Finitism*, Dordrecht: Reidel.