L&PS – Logic and Philosophy of Science
Vol. IX, No. 1, 2011, pp. 263-274

Gödel and the fundamental incompleteness of human self-knowledge

Vincenzo Fano
Università degli Studi di Urbino Carlo Bo
Dipartimento di Scienze di Base e Fondamenti
e-mail: vincenzo.fano@uniurb.it

Pierluigi Graziani
Università degli Studi di Urbino Carlo Bo
Dipartimento di Scienze della Comunicazione
e-mail: pierluigi.graziani@uniurb.it

ABSTRACT. In our paper, we show how to present Gödel's analysis on the consequences of his incompleteness theorems for the philosophy of the mind in a rigorous way. We also highlight the necessary philosophical premises of Gödel's argument and more in general of Gödelian arguments.

1. 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.[1]
This was one of the few official occasions in which Gödel expounded his opinion on the philosophical implications of his theorems. Without going into

[1] Gödel 1995. A very accurate analysis of this paper is proposed by: Feferman 2006; Tieszen 2006; van Atten 2006.

Published by L&PS – Logic and Philosophy of Science
http://www2.units.it/ ~episteme — ISSN: 1826-1043

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. Gödel's idea is that if someone perceives with absolute certainty that a certain formal system[2] 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"[3]. 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:

> 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.[4]

[2] It is understood that, in this paper, the expression "formal system" indicates a formal system which is adequate to derive incompleteness theorems.

[3] Gödel 1995, p. 309.

[4] 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 Turing Machine (*TM*) connected to it (the *TM* which enumerates the set of theorems of the corresponding formal system).

The question, then, requires due consideration precisely of 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[5] 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".[6]

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".[7] 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-

[5] 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".

[6] Feferman 2006, pp. 135-136.

[7] Feferman 2006, pp. 136-137.

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 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[8] 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)".[9]

So, considering the translatability between the concept of a well defined formal system and that of a *TM*, we can say that Gödel's theorems leave open the three following possibilities:[10]

(I) human intelligence infinitely surpasses the powers of the finite machine (TM), and there are no absolutely irresolvable Diophantine problems.[11]

[8] 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 $\forall x_1, \ldots, \forall x_n \exists y_1, \ldots, \exists y_m$ [p($x_1, \ldots, x_n, y_1, \ldots, y_m$) = 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*.

[9] Gödel 1995, p. 310.

[10] Tieszen 2006.

[11] See note 8.

(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 presented by Benacerraf in 1967 and Chihara in 1972. Let us try to analyse it further by means of a formulation partly provided by Shapiro[12].

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.

[12] Shapiro 1998.

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:[13]

(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[14] 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 $\vdash f \leftrightarrow F(\mathbf{f})$ [15].

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:

Let *f* be any sentence in the language of first-order arithmetic and **B**[16] 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'.[17]

[13] See Löb 1955; Boolos 1993; Detlefsen 2002.

[14] See Smullyan 1992, VIII and IX.

[15] We will indicate numeral with **bold** letters.

[16] 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**'.

G6. Let Con_e be the statement generated by TM_e 'there does not exist a **y** such that **B(y,m)**' where **m** 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 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.

2. A new formulation of the argument

As has rightly been pointed out by Shapiro,[18] a fundamental issue of the debate[19] concerning the consequences of Gödel's incompleteness theorems for

[17] 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.

[18] Shapiro 1998. See also Tamburrini 2002, pp. 130-133.

[19] A very accurate analysis of debate is proposed by: Fano and Graziani 2011.

the philosophy of the mind is that it is not quite clear what the exact content of the mechanistic view should be. Indeed all authors which have dealt with the issue of defining this content, either refuting or valorizing it. Despite this, from their 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[20]: "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 ("every statement that I can derive and that I know to be true")[21] and Penrose's *A* set (all procedures which are followed by the mathematical community to prove theorems)[22] 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,[23] 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, there existed entities like "2" and "4", which were already in that relation. This is obviously an unjustified and groundless metaphysical hypothesis. However, as Quine and Putnam have pointed out, introducing

[20] Shapiro 1998, p. 275.
[21] Benacerraf 1967.
[22] Penrose 1989; 1994; 1996.
[23] See Kripke 1982.

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. 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 develop (following Benacerraf, Chihara, Penrose, Shapiro, and obviously, Gödel's suggestions) a new Gödelian disjunctive argument without referring to the real *performances* (see Chihara 1972) of mathematicians, but rather to their *ideal arithmetical competence* (obtaining an improvement in Chihara's argument).
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 Formal Arithmetic (*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. In Chihara's interpretation, that is completely internal to Formal Arithmetic, this means that one *could* build $s(\mathbf{n})$ which is true in *N* if and only if *n* belongs to *S'*; $s(\mathbf{n})$ is a formula in *FA*. 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*(**m**).

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

CFG6. In *FR'*, it is possible to build the Gödel formula. That is, **m** is the Gödel number of the formula *G*, which states that $\neg s(\mathbf{m}_G)$.

CFG7. By applying CFG4 we obtain that:
$\Box_{FR}\, s(\mathbf{m}_G) \rightarrow \neg s(\mathbf{m}_G)$

CFG8. Hence, we have that:
$\Box_{FR}\, \neg s(\mathbf{m}_G)$

CFG9. Hence, for some numeral **n**, in *FR* '$\mathbf{B(n, m}_G)$' is derivable.

So, by using CFG5 we have $\Box_{FR'}\, s(\mathbf{m}_G)$.
But *FR'* is an extension of *FR*, in which, as we have seen in CFG7, we can derive $\neg s(\mathbf{m}_G)$. We thus have a contradiction in *FR'*.

CFG10. Hence, if *FR'* is contradictory, 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.

References

BENACERRAF P. (1967): "God, the devil and Gödel", *The Monist*, 51, 9-32.

BOOLOS G. (1993): *The Logic of Provability*, Cambridge: Cambridge University Press.

CHIHARA C. S. (1972): "On alleged refutations of mechanism using Gödel's incompleteness results", *The Journal of Philosophy*, 69, 507-526.

DETLEFSEN M. (2002): "Löb's theorem as a limitation on mechanism", *Minds and Machines*, 12, 353–381.

FANO V. AND GRAZIANI P. (2011): "On the necessary philosophical premises of Gödelian arguments", available at http://philsci-archive.pitt.edu/8844/

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

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.

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.

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.

SHAPIRO S. (1998): "Incompleteness, mechanism, and optimism", *The Bulletin of Symbolic Logic*, 4, 273-302.

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.
VAN ATTEN M. (2006): "Two draft letters from Gödel on self-knowledge of reason", *Philosophia Mathematica*, 14, 255-261.