Do the Self-knowing Machines Dream of Knowing their Factivity?

Alessandro Aldini[1], Vincenzo Fano[1], and Pierluigi Graziani[2]

[1] Università di Urbino "Carlo Bo", Italy
[2] Università di Chieti-Pescara, Italy

Abstract. The *Gödelian Arguments* represent the effort done to interpret Gödel's Incompleteness Theorems in order to show that minds cannot be explained in purely mechanist terms. With the purpose of proving the limits of mechanistic theses and investigate aspects of the Church-Turing Thesis, several results obtained in the formal setting of Epistemic Arithmetic (EA) reveal the relation among different properties of knowledge of machines, including self-awareness of knowledge and factivity of knowledge. We discuss the main principles behind the *Gödelian Arguments* and extend the results obtained in EA. In particular, we define a machine that, in a specific case, knows its own code and the factivity of its own knowledge, thus providing new insights for the analysis of the *Gödelian Arguments*.

1 Introduction

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* [7]. The theorems in question were precisely those of Incompleteness and the philosophical implications were concerned with the nature of mathematics and the abilities of the human mind [3]. This was one of the few official occasions in which Gödel expounded his opinion on the philosophical implications of his theorems. Without going into details about Gödel's paper, what is interesting here is the first part, where he derives the thesis of essential incompleteness of mathematics from his famous theorems. Such a thesis was sanctioned by the second theorem. Gödel's idea is that if one perceives with absolute certainty that a certain formal system [4] is correct (sound), s/he will also know the consistency of the system, that is, s/he will know the truth of the statement establishing 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 one makes such a statement he contradicts himself" [7, p. 309].

[3] A very accurate analysis of this work is proposed by Feferman [6], Tieszen [16], and van Atten [17].

[4] In this paper, the expression "formal system" indicates a system that is adequate to derive Incompleteness Theorems.

But what does all of this mean? Does it mean perhaps that a well defined system of correct (sound) axioms cannot contain everything that is strictly mathematical?

In the following, we first recall and discuss Gödel believes about the possible answers to such a question and then analyze the so called *Gödelian Arguments*. In the last decades many scholars dealt with these arguments, which represent the effort done to interpret Gödel's Incompleteness Theorems with the purpose of showing that minds cannot be explained in purely mechanist terms. Among them, we concentrate on the approach followed by Reinhardt [13], Carlson [3], and Alexander [1], who demonstrated a series of results in the formal setting of *Epistemic Arithmetic*, which encompasses some typically informal aspects of the Gödelian Arguments about the knowledge that can be acquired by (knowing) machines. These results emphasize several relations among different properties characterizing the expressiveness of machines, including self-awareness of knowledge and factivity of knowledge. As a contribution of this paper, we integrate these results with novel insights, thus providing the formal base for additional elements supporting the Gödelian Arguments [5].

2 Gödel Perspective

With reference to the previous question, Gödel believes that it has two possible answers:

> It does, if by mathematics proper is understood the system of all true mathematical propositions; it does not, however if someone 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 minds)] boundlessness or inexhaustibility [7, pp. 309-310].

Therefore, not only does the previous question pose the problem of the inexhaustibility or incompleteness of mathematics considered as the totality of all

[5] Although there are some very interesting connections between Gödel's Theorems and contemporary research on deep learning, we do not analyze them in this contribute. On this issue you can see [15].

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 minds demonstrative abilities are extensionally equivalent to a certain formal system, or to the Turing Machine (TM) connected to it (the TM that 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 first-order arithmetic, and call this *objective arithmetic*, or, following Gödel, "objective mathematics", that is "the body of those mathematical propositions which hold in an absolute sense, without any further hypothesis" [7, p. 305]. 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 that a human being can know and prove absolutely and with mathematical certainty, that is what one can derive [6] 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, p. 135-136]. What is then the relation between K and T? Quoting Feferman, we could synthesize Gödel's answer by saying that if K was equal to T:

> then demonstrations in subjective mathematics [would not be] 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 demonstrated, and those constitute absolutely unsolvable problems [6, p. 136-137].

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 characteristics of T, for each formal system there would be a provable statement by the human mind, but not within the formal system. Hence, the mechanistic thesis 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 were equivalent to a given formal system or to the TM related to it, the existence of arithmetical statements humanly undecidable in an absolute sense would follow. In fact, as underlined by Gödel, the second incompleteness theorem does allow this conclusion: the proposition expressing the consistency of K, say ConK, is true but is not provable within the system itself; the negation of ConK is false and is not provable in K. Having established the equivalence between the human mind and a formal system, ConK is not even provable by the human mind. Finally, since ConK can be put in the form of a Diophantine

[6] As Feferman [6, 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".

problem, it is an absolutely undecidable problem. Such a proposition is, thus, an unknowable truth. These arguments lead Gödel to the idea that from the incompleteness results one can at the most derive 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) [7, p. 310].

So, following Tieszen [16], and considering the translatability between the concept of a well defined formal system and that of a TM, we can say that Gödel's Incompleteness Theorems show that it could not be true that: (i) the human mind is a finite machine (a TM) and there are for it no absolutely undecidable Diophantine problems.

> The incompleteness theorems show that if we think of the human mind as a TM then there is for each TM some absolutely undecidable Diophantine problem. The denial of the conjunction (i) is, in so many words, Gödel's disjunction. In formulating the negation of (i) Gödel says that the human mind infinitely surpasses the powers of any finite machine. One reason for using such language, I suppose, is that there are denumerably many different Turing machines and for each of them there is some absolutely diphantine problem of the type Gödel mentions. So Gödel's disjunction, understood in this manner, is presumably a mathematically established fact. It is not possible to reject both disjuncts. [16, pp. 230-231].

The disjunction leaves open the three following possibilities:

(I) human intelligence infinitely surpasses the powers of the finite machine (TM), and there are no absolutely unsolvable Diophantine problems (see [7, p. 310]).
(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 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". 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". In 1972 Gödel expressed further on the matter saying [18]:

> On the other hand, on the basis of what has been proved so far, it remains possible that there may exist (and even be empirically discoverable) a theorem-proving machine which in fact is equivalent to mathematical intuition, but cannot be *proved* to be so, nor even be proved to yield only *correct* theorems of finitary number theory.

This formulation is significantly different from that of 1951, as now Gödel appears to recognize that the mind, at least in his doing mathematics, could be a machine and we could not recognize this fact or not be able to prove it.

3 Knowing Machines

After the speculative ideas formulated by anti-mechanists, like the famous argument by Lucas [8, 9], several authors, like Benacerraf [2], Penrose [10–12], Chihara [4], and Shapiro [14] (see [5] for a comprehensive survey), proposed more formal lines of reasoning on the implications of Gödel's Theorems. Here, we consider the results by Reinhardt [13], Carlson [3], and Alexander [1], who analyzed a formal theory, called *Epistemic Arithmetic* (EA), encompassing some typically informal aspects of the Gödelian Arguments about the knowledge that can be acquired by (knowing) machines. EA is the language of Peano Arithmetic enriched with a modal operator K for *knowledge* (or for *intuitive provability*). The formal interpretation of K passes through the definition of the properties at the base of an epistemic notion of *knowability*:

- Logic Consequence: if ϕ and $\phi \to \psi$ are known, then ψ is known.
- Infallibilism: what is known is also true.
- Introspection: if ϕ is known then such a knowledge is known.

The basic axioms of knowledge are:

B1. $K\forall x\phi \to \forall x K\phi$
B2. $K(\phi \to \psi) \to K\phi \to K\psi$
B3. $K\phi \to \phi$
B4. $K\phi \to KK\phi$

where $B2$-$B4$ formalize the intuitions above and are stricly related to, e.g., the modal system $S4$, while $B1$ expresses a first-order condition stating that the assertion "ϕ is known to be valid" implies the knowledge of each element that can be assigned to x in ϕ and the truth of the formula under each such assignment. [7] Assumed that the K-closure of ϕ is the universal closure of ϕ possibly prefixed by K, the axioms of EA are the K-closure of $B1$-$B4$ and of the axioms of Peano Arithmetic. The theory of knowledge defined in such a way extends conservatively the classical interpretation of Peano Arithmetic.

[7] We are assuming that ϕ is a formula with one free variable x.

Under this theory of knowledge, variants of Church-Turing Thesis are investigated to analyze the relationship between properties that are *weakly K-decidable* [8] and the TMs that formalize the decision algorithm for these properties. In the following, we assume that W_e is the recursively enumerable set with Gödel number e.

Theorem 1 (Reinhardt's schema [13]). *$\exists eK\forall x(K\phi \leftrightarrow x \in W_e)$ is not consistent in* EA[9].

Informally, Reinhardt's schema states that a TM exists for which *it is known* that it enumerates all (and only) the elements (for which *it is known*) that make ϕ true. More precisely, as the assignments making ϕ true are a known recursively enumerable set, we then derive the computability, through a known TM, of the (weak K-) decision problem for ϕ. Following Carlson, the intuitive interpretation is: *I am a TM and I know which one.* A weaker version of Reinhardt's schema is conjectured by Reinhardt himself and proved by Carlson, in which the outermost K operator prefixes the statement.

Theorem 2 (Carlson's schema [3]). *$K\exists e\forall x(K\phi \leftrightarrow x \in W_e)$ is consistent in* EA.

Quoting Carlson, *I know that the set of x for which I know $\phi(x)$ is recursively enumerable*, or, by rephrasing an analogous hypothesis studied by Benacerraf independently [2], *I know I am a TM but I do not know which one.* Carlson uses the term *knowing machine* to denote any recursively enumerable proof system that represents a model for the theory of knowledge, and shows that, indeed, EA integrated with his schema is a knowing machine. As a corollary of this result, the schema obtained by removing the outermost K operator is still consistent in EA.

The proofs of the results above rely on the validity of $K(K\phi \rightarrow \phi)$, stating that in the formal system the *factivity* of knowledge is known. In between these two limiting results, Alexander has recently proved a dichotomy: a machine can know its own factivity as well as that it has some code (without knowing which, as stated by Carlson's schema), or it can know its own code exactly (proving the consistency of Reinhardt's schema) but cannot know its own factivity (despite actually being factive). Providing that the axioms of EA *mod factivity* consist of the axioms of EA except for the universal closure of $B3$ prefixed by K (that represents knowledge of factivity of knowledge), it is possible to prove that:

Theorem 3 (Alexander [1]). *Reinhardt's schema is consistent in* EA *mod factivity.*

and then to construct the previous dichotomy.

In this setting, we show a result related to a specific case. An interpreter f_u is a function mimicking the behavior of any other function. Formally, $f_u(x, y) = f_x(y)$. For instance, the universal TM is an interpreter. Interpreters represent a

[8] The assignments of x satisfying ϕ are known.

[9] The inconsistency of this schema is proved as a consequence of first Gödel's theorem.

classical tool in computability theory and play a fundamental role for programming languages. Now, let us consider Reinhardt's schema in EA *mod factivity* and $\phi(x) := (f_x(x) = 1)$. Then, from:

$$\exists e K \forall x (K\phi \leftrightarrow x \in W_e)$$

by taking $x = e$ we derive:

$$\exists e K (K\phi(e) \leftrightarrow e \in W_e) \tag{1}$$

and:

$$K(K\phi(e) \rightarrow \phi(e)) \tag{2}$$

which expresses a limited form of knowledge of factivity that is allowed in EA *mod factivity*. More precisely, we have a machine that, for (at least) a specific choice of the function ϕ and of the input x, i.e., the interpreter function and the Gödel number of the machine itself, knows its own code and its own factivity. We have to note that taking $x = e$ roughly speaking means that *if I allow the machine to know its own identity, then of course it will possess this knowledge*. Attributing this capacity to a machine is very natural for us and in our opinion it shows that Alexander's framework is adequate to analyze the machines' knowledge [10]. By virtue of such a choice, the intuition that we stem is that the machine knows its own code and is aware of the factivity of the knowledge resulting by interpreting its own code, while such an awareness, according to the dichotomy above, is lost when interpreting other inputs. As a consequence, by rephrasing Carlson and Benacerraf intuitions, we could say: *If I know which universal TM I am, then I know the factivity of my knowledge.*[11] Hence, to some extent, self-reference increases the expressiveness of knowledge, provided that the machine is an interpreter. In our opinion, this is an interesting enhancement of the tradeoff result provided by Alexander that can represent an additional formal element for the analysis of the Gödelian Arguments.

References

1. S. Alexander. A machine that knows its own code. *Studia Logica*, 102:567–576, 2014.
2. P. Benacerraf. God, the Devil and Gödel. *The Monist, 51*, pages 9–32, 1967.
3. T.J. Carlson. Knowledge, machines, and the consistency of Reinhardt's strong mechanistic thesis. *Annals of Pure and Applied Logic*, 105:51–82, 2000.
4. C.S. Chihara. On alleged refutations of mechanism using Gödel's incompleteness results. *The Journal of Philosophy*, 69:507–526, 1971.
5. V. Fano and P. Graziani. Mechanical Intelligence and Gödelian arguments. In E. Agazzi, editor, *The Legacy of A.M. Turing*, pages 48–71. Franco Angeli, 2013.

[10] Note also that e could have a very high complexity and this fact is compatible with the incapability of humans to understand how the most advanced algorithms produce their results.

[11] Roughly speaking: *If I know which universal TM I am, then I know my factivity.*

6. S. Feferman. Are There Absolutely Unsolvable Problems? Gödel's Dichotomy. *Philosophia Mathematica*, (III) 14:134–152, 2006.
7. K. Gödel. Some basic theorems on the foundations of mathematics and their implications. In K. Gödel, editor, *Collected Works*, volume III, pages 304–335. Oxford University Press, 1995.
8. J.R. Lucas. Minds, Machine and Gödel. *Philosophy*, 36:112–127, 1961.
9. J.R. Lucas. Satan stultified: a rejoinder to Paul Benacerraf. *The Monist*, 52:145–158, 1968.
10. R. Penrose. *The emperor's new mind*. Oxford Univ. Press, Oxford, 1989.
11. R. Penrose. *Shadows of the mind*. Oxford University Press, Oxford, 1994.
12. R. Penrose. Beyond the doubting shadow. *Psyche*, 2-1, 1996.
13. W. Reinhardt. Epistemic theories and the interpretation of Gödel's incompleteness theorems. *Journal of Philosophical Logic*, 15:427–474, 1986.
14. S. Shapiro. Incompleteness, mechanism, and optimism. *The Bulletin of Symbolic Logic*, 4:273–302, 1998.
15. B. R. Steunebrink and J. Schmidhuber. Towards an Actual Gödel Machine Implementation: a Lesson in Self-Reflective Systems. *Theoretical Foundations of Artificial General Intelligence*, 4:173–195, September 2012.
16. R. Tieszen. After Gödel: Mechanism, Reason, and Realism in the Philosophy of Mathematics. *Philosophia Mathematica*, (III) 14:229–254, 2006.
17. M. van Atten. Two Draft Letters from Gödel on Self-knowledge of Reason. *Philosophia Mathematica*, 14:255–261, 2006.
18. H. Wang. *From Mathematics to Philosophy*. Humanities Press, N.Y., 1974.