A Note on Knowing Machines

Alessandro Aldini*, Vincenzo Fano*, Pierluigi Graziani**
* University of Urbino "Carlo Bo", Italy
**University of Chieti-Pescara, Italy

The *Gödelian Arguments* represent the effort done to interpret Gödel's Incompleteness Theorems [6] with the purpose of showing that minds cannot be explained in purely mechanist terms. In particular, in response to the mechanist project launched by Turing [13], several speculative ideas, like the famous anti-mechanist argument by Lucas [7, 8], have been proposed to refute it informally. On the other hand, authors like Benacerraf [2], Chihara [4], and Shapiro [12], tried to follow more sophisticated lines of reasoning for the analysis of the relation between human mind and machines [5]. In this setting, we consider the most recent results by Reinhardt [10], Carlson [3], and Alexander [1], who analyzed a formal theory, called Epistemic Arithmetic (EA) [9, 11], 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 knowledge:

- *Logic Consequence*: if ϕ and $\phi \rightarrow \psi$ 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. $K\forall x\phi \rightarrow \forall xK\phi$
B2. $K(\phi \rightarrow \psi) \rightarrow K\phi \rightarrow K\psi$
B3. $K\phi \rightarrow \phi$
B4. $K\phi \rightarrow KK\phi$

where B2–B4 formalize the intuitions above and are strictly related to, e.g., the modal system S4, while the first order condition B1 establishes that the statement "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.[1]

Provided 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.

Under this theory of knowledge, variants of Church's Thesis are investigated to analyze the relationship between properties that are weakly K-decidable[2] and the Turing Machines (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.

[1] We are assuming that ϕ is a formula with one free variable x.
[2] The assignments of x satisfying ϕ are known.

Theorem 1 (**Reinhardt's schema** [10]). $\exists e K \forall x (K\phi \leftrightarrow x \in W_e)$ is not consistent in EA.

Intuitively, 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. By citing Carlson, *I am a TM and I know which one*. The inconsistency of this schema is a consequence of first Gödel's theorem. A weaker version of Reinhardt's schema is defined by Carlson.

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

By citing Carlson, *I know that the set of x for which I know* $\phi(x)$ *is recursively enumerable* or, by rephrasing an hypothesis studied by Benacerraf independently [2], *I am a TM but I do not know which one*. As a corollary of this result, the schema obtained by removing the outermost K operator is still consistent in EA.

The proofs of these results 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), 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, 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 Φ_u is a function mimicking the behavior of any other function. Formally, $\Phi_u(x,y) = \Phi_x(y)$. For instance, the universal TM is an interpreter. Now, let us consider Reinhardt's schema in EA *mod factivity* and $\phi = \Phi_u(x,x) = \Phi_x(x)$. Then, from:

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

by taking $x = e$ we derive:

$$\exists e K (K\Phi_e(e) \leftrightarrow e \in W_e)$$

and:

$$K(K\Phi_e(e) \rightarrow \Phi_e(e))$$

which expresses a limited form of knowledge of factivity 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 own factivity. By virtue of such a choice, the intuition that we stem is: *I am a factive TM – and I know which one – if I am a universal TM*. In our opinion, this is an interesting enhancement of the tradeoff result provided by Alexander that can represent an additional element for the analysis of the Gödelian Arguments.

References

[1] S. Alexander, "A machine that knows its own code", *Studia Logica* 102, 2014, pp. 567-576.

[2] P. Benacerraf, "God, the Devil, and Gödel", *The Monist* 51, 1967, pp. 9-32.

[3] T.J. Carlson, "Knowledge, machines, and the consistency of Reinhardt's strong mechanistic thesis", *An-*

nals of Pure and Applied Logic 105, 2000, pp. 51-82.

[4] C.S. Chihara, "On alleged refutations of mechanism using Gödel's incompleteness results", *The Journal of Philosophy* 69, 1971, pp. 507-526.

[5] V. Fano and P. Graziani, "Mechanical Intelligence and Gödelian arguments", in E. Agazzi (ed.), *The Legacy of A.M. Turing*, Franco Angeli, 2013, pp. 48-71.

[6] K. Gödel, "Über formal unentscheidbare Sätze der Principia mathematica und verwandter Systeme", *Monatshefte für Mathematik und Physik* 38, 1931, pp. 173-198; En. Tr. in K. Gödel, *Collected Works* I, Oxford University Press, pp. 144-195.

[7] J. Lucas, "Minds, machine and Gödel", *Philosophy* 36, 1961, pp. 112-127.

[8] J. Lucas, "Satan stultified: a rejoinder to Paul Benacerraf", *The Monist* 52, 1968, pp. 145-158.

[9] W. Reinhardt, "The consistency of a variant of Church's Thesis with an axiomatic theory of an epistemic notion", *Revista Colombiana de Matematicas* 19, special volume for the *Proceedings of the 5th Latin American Symposium on Mathematical Logic* (1981), 1985, pp. 177-200.

[10] W. Reinhardt, "Epistemic theories and the interpretation of Gödel's incompleteness theorems", *Journal of Philosophical Logic* 15, 1986, pp. 427-474.

[11] S. Shapiro, "Epistemic and intuitionistic arithmetic", *Intensional Mathematics*, North-Holland, 1985, pp. 11-46.

[12] S. Shapiro, "Incompleteness, mechanism, and optimism", *The Bulletin of Symbolic Logic* 4, 1998, pp. 273-302.

[13] A. Turing, "Computing machinery and intelligence", *Mind* 59, 1950, pp. 433-460.