Theory of Knowing Machines: Revisiting Gödel and the Mechanistic Thesis

Alessandro Aldini, Vincenzo Fano, Pierluigi Graziani

Theory of Knowing Machines: Revisiting Gödel and the Mechanistic Thesis

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

[1] University of Urbino "Carlo Bo", Urbino, Italy
{alessandro.aldini, vincenzo.fano}@uniurb.it
[2] University of Chieti-Pescara "G. D'Annunzio", Chieti, Italy
pierluigi.graziani@unich.it

Abstract. Church-Turing Thesis, mechanistic project, and Gödelian Arguments offer different perspectives of informal intuitions behind the relationship existing between the notion of *intuitively provable* and the definition of decidability by some Turing machine. One of the most formal lines of research in this setting is represented by the theory of knowing machines, based on an extension of Peano Arithmetic, encompassing an epistemic notion of knowledge formalized through a modal operator denoting *intuitive provability*. In this framework, variants of the Church-Turing Thesis can be constructed and interpreted to characterize the knowledge that can be acquired by machines. In this paper, we survey such a theory of knowing machines and extend some recent results proving that a machine can know its own code exactly but cannot know its own correctness (despite actually being sound). In particular, we define a machine that, for (at least) a specific case, knows its own code and knows to be sound.

Keywords: Church-Turing Thesis, Mechanism, Epistemic Arithmetic, Gödelian Arguments, Intuitive provability, Knowledge operator

1 Introduction

After the seminal paper by Turing in 1950 [37], the launch of the mechanistic project, intended to establish whether minds can be explained, either extensionally or intensionally, in purely mechanist terms, was contrasted by the so-called *Gödelian Arguments*. These represent the effort done by several scholars to interpret Gödel's Incompleteness Theorems [12] with the purpose of refuting mechanism. In particular, several speculative ideas, like the anti-mechanist arguments by Lucas [24, 25] and Penrose [29], contributed to animate the debate. On the other hand, authors like Benacerraf [5], Chihara [8], and Shapiro [35] tried to follow more sophisticated lines of reasoning for the analysis of the relation between human mind and machines [9]. Most of these approaches preserve intensional elements on properties of human mind that make severely informal the argumentation and, more confusing, make hard even to define precisely what the (anti-)mechanistic thesis claims.

In this setting, a series of results presented by Reinhardt [32], Carlson [7], and Alexander [3], clarify some typically informal aspects of the Gödelian Arguments. This is done in the framework of a theory, called *Epistemic Arithmetic* (EA), developed independently by both Reinhardt and Shapiro [30, 34, 31], and encompassing an epistemic notion of intuitive provability. In particular, Reinhardt uses such an axiomatic framework to study both ramifications of the Church-Turing Thesis [21]:

every effectively calculable function is computable by a Turing Machine

and the consequences of Gödel's Incompleteness Theorems, with the aim of strengthening the formal elements behind the philosophical debate on the knowledge that can be acquired by machines.

EA is the language of Peano Arithmetic enriched by a modal operator K for *knowledge*, which is the notation used by Shapiro and from Carlson on (see, e.g., [34, 7, 3]). According to Shapiro, K means *"ideally, or potentially, knowable"*, while Carlson, analogously, says *"can eventually come to be known"*. On the other hand, Reinhardt prefers the more specific interpretation *"it is intuitively provable that"* [30] and, to strengthen such an idea, uses the epistemic operator B for *beweisbar* (meaning *provable* in German). By following the same motivation, we adopt the operator B, which is closer to the intended interpretation of its role, and use the two intuitions – provability and knowledge – interchangeably, by assuming that the notion of knowability [34] is actually limited to *intuitive provability*. Hence, the idea behind such a modal operator is to express a definition of decidability by human mind (*humanly provable*) that occurs in many forms both in the mechanistic thesis and in the Gödelian Arguments.

The formal interpretation of B passes through the definition of the properties at the base of an epistemic notion of knowledge. For instance, it is expected that humanly provable statements are closed under logic consequence, meaning that if we can intuitively prove $\phi \to \psi$ and we have an intuitive proof of ϕ, then by combining these we derive an intuitive proof of ψ. In other words, we are representing *modus ponens* as a rule for intuitive provability. Analogously, we also expect the soundness of intuitive provability (what can be proved is true, thus stating the infallibilism of knowledge) and that what is humanly provable includes all tautologies. Using this latter rule in the setting of intuitive provability amounts to establish an introspection principle: if ϕ can be proved then such a knowledge can be proved. Such a principle may be seen as an instance of the general, largely discussed and controversial, KK (knowing that one knows) law (see, e.g., [17, 19]). However, intuitive provability ("it can be proved") is weaker than the general notion of knowledge ("I know that") and relies on idealizations abstracting from space, time, and complexity constraints, thus making the introspection rule less debatable with respect to the classical contrast between internalist and externalist theories of knowledge (see, e.g., [6, 15]).

As we will see, all the laws informally expressed above are formalized through very common axioms of epistemic modal systems, see, e.g., [18] for a comprehensive discussion. Therefore, in essence, knowability is treated in axioms form, while any attempt to specify precisely its meaning by following model theoretic

approaches is intentionally avoided. In fact, representing knowledge as a predicate easily leads to contradictions forcing all true propositions to be provable [23, 27, 26, 30, 32].

Then, the aim of an axioms system based on the operator B is to apply deductive reasoning to prove the (in)consistency of statements specified in an appropriately formulated logic and representing conjectures related to Church's Thesis and the mechanist project, with a specific interest towards the following:

> *the property of being humanly provable is equivalent to decidability by some Turing Machine.*

In the rest of the paper, we first illustrate such a theory of knowledge (Section 2). Then, we survey the main results obtained by recasting Gödel Incompleteness Theorems and by analyzing variants of the Church-Turing Thesis in this setting (Section 3). We also show how to extend a tradeoff result obtained by Alexander [3] about the relation between knowledge of soundness and knowledge of own code for (knowing) machines. Some conclusions terminate the paper (Section 4). This paper is a full and revised version of an extended abstract presented at HaPoC 2015 [1].

2 Epistemic Arithmetic

We start by introducing some notation used to describe the language of *Peano Arithmetic* (PA)[3]: ϕ, ψ denote well formed formulas (wff, for short), which, if not specified, are considered to include only one free variable x.[4] A *sentence*, usually denoted by σ, is a wff without free variables, and any set of sentences closed under logical consequence is called *theory*. Terms and substitution principle are defined as usual: if t is a term, then $x\,|\,t$ denotes the substitution of x by t. An *assignment* is a function $s : \mathcal{V} \to \mathcal{U}$ from the domain of variables to the reference universe, such that $s(x\,|\,t)$ denotes the function assigning t to x and $s(y)$ to every variable $y \neq x$.

Epistemic Arithmetic extends Peano Arithmetic with the modal operator B. The language of EA contains every wff of the language of PA and the additional formulas of the form $B\phi$ whenever ϕ is a wff of EA. In the following, given a set Φ of wff we also use the notation $B\Phi$ to denote the set $\{B\phi \mid \phi \in \Phi\}$. It is worth noticing that $B\phi$ is treated as an atomic formula. Then, in EA a structure $\mathcal{M}$ with respect to a universe $\mathcal{U}$ is defined in the classical way for the first order part related to PA and includes also a boolean interpretation function for B. More precisely, for each assignment s, $\mathcal{M} \models B\phi[s]$ means that ϕ can be intuitively proved when the free variables of ϕ are interpreted according to s. Obviously, the interpretation of $B\phi[s]$ does not depend on $s(x)$ whenever x does not occur free in ϕ.

[3] We employ the standard nonlogical symbols of PA syntax $\mathbf{0}, \mathbf{S}, +, \cdot$, and the usual application order on the operators.

[4] A variable x occurs free in ϕ if it is not in the scope of any quantifier of ϕ defined over x.

Moreover, we write $\mathcal{M} \models \phi$ if $\mathcal{M} \models \phi[s]$ for each assignment s and we say that ϕ is valid if $\mathcal{M} \models \phi$ for each structure $\mathcal{M}$; ϕ is logical consequence of a set of sentences Σ (denoted $\Sigma \models \phi$) if for each structure $\mathcal{M}$ it holds that:

$$\forall \sigma \in \Sigma : \mathcal{M} \models \sigma \;\Rightarrow\; \mathcal{M} \models \phi.$$

We observe that, in such an epistemic extension of PA, compactness and completeness results can be proved.[5]

As far as the axiomatization of EA is concerned, we first recall the Peano axioms:

1. $\forall x(\mathbf{S}(x) \neq \mathbf{0})$
2. $\forall x \forall y((\mathbf{S}(x) = \mathbf{S}(y)) \rightarrow (x = y))$
3. $\forall x(x + \mathbf{0} = x)$
4. $\forall x \forall y(x + \mathbf{S}(y) = \mathbf{S}(x + y))$
5. $\forall x(x \cdot \mathbf{0} = \mathbf{0})$
6. $\forall x \forall y(x \cdot \mathbf{S}(y) = x \cdot y + x)$
7. $\forall y_1 \ldots \forall y_n((\phi(x\,|\,\mathbf{0}) \wedge \forall x(\phi \rightarrow \phi(x\,|\,\mathbf{S}(x)))) \rightarrow \forall x \phi)$ for each wff ϕ

establishing that $\mathbf{0}$ is not in the codomain of $\mathbf{S}$ (see 1), $\mathbf{S}$ is injective (see 2), while 3 to 6 define the rules for sum and product, and 7 expresses the induction schema, where ϕ is a formula with free variables $x, y_1, \ldots, y_n$.

To the standard Peano axioms we add the universal closure of the following basic axioms of knowledge:

- $B1$. $B\forall x\phi \rightarrow \forall x B\phi$
- $B2$. $B(\phi \rightarrow \psi) \rightarrow (B\phi \rightarrow B\psi)$
- $B3$. $B\phi \rightarrow \phi$
- $B4$. $B\phi \rightarrow BB\phi$

where $B2$-$B4$ formalize the intuitive properties illustrated in Section 1 and are essentially taken from the modal system S4 [22, 18]. To complete the axioms system, the first order condition $B1$ establishes that the statement "ϕ can be proved to be valid" implies the knowledge of each element that can be assigned to the free variable x of ϕ and the provability of the formula under each such assignment. This condition represents a sort of accessibility principle ensuring that the elements assigned to the free variable of ϕ should be accessible to the knower in some way, and from which we notice that the following desirable condition can be derived:

$$\mathcal{M} \models B\forall x\phi[s] \;\Rightarrow\; \mathcal{M} \models B\phi[s(x|u)]\ \forall u \in \mathcal{U}$$

which states that if ϕ can be proved under the interpretation provided by $\mathcal{M}$, then it is possible to prove $\phi(x)$ for each assignment of the variable x in the universe $\mathcal{U}$.

[5] In particular, the set of valid wff is recursively enumerable and if Σ is recursively enumerable then so is $\{\phi \mid \Sigma \models \phi\}$.

Now, given a wff ϕ, we define the B-closure of ϕ as $\phi \cup B\phi$, which extends easily to sets of formulas. Then, the standard axiomatization of the theory of knowledge for EA is given by the B-closure of $B1$-$B4$ and of the axioms of PA, with the modus ponens being the unique rule of this formal system. The theory of knowledge axiomatized in such a way extends conservatively the classical interpretation of PA.

In particular, let us characterize the following property emphasizing the role of each element of the theory of knowledge. If T is a theory with an axiomatization of the form $\Sigma \cup B\Sigma$ extending the theory of knowledge, then:

$$\sigma \in T \;\Leftrightarrow\; B\sigma \in T.$$

Hence, we have a correspondence between the valid sentences of T and the sentences that can be intuitively proved in T. On one hand, if $\sigma \in \Sigma \cup B\Sigma$ then, by $B4$, $BB\sigma \in T$ and, by $B3$, $B\sigma \in T$. Now, assume that σ_2 derives by modus ponens from $\sigma_1, \sigma_1 \rightarrow \sigma_2 \in T$. By induction hypothesis, both $B\sigma_1$ and $B(\sigma_1 \rightarrow \sigma_2)$ are in T and, by applying $B2$, we derive $B\sigma_2 \in T$. On the other hand, by $B3$ we have that $B\sigma \in T$ implies $\sigma \in T$.

Finally, giving for granted the accessibility principle, it is possible to reason about a simple theory of knowledge [7, 3] in which $B1$ is replaced by the rule of necessitation: if $\vdash \phi$, then $\vdash B\phi$. Hence, it can be shown that the theory of knowledge corresponds to the set of sentences that can be proved in the modal system S4.

2.1 Knowing Entities and Machines

The statements presented in the next section are considered with respect to a knower reasoning about basic arithmetic. Therefore, we refer to the standard interpretation of the model of arithmetic, extended formally in the following way. Given a set of sentences Σ in EA, the structure $\mathcal{N}_\Sigma$ over the set of naturals $\mathbb{N}$ is such that, for any wff ϕ with one free variable x, it holds that:

$$\mathcal{N}_\Sigma \models B\phi[s] \text{ iff } \Sigma \models \phi(x \,|\, \overline{s(x)})$$

where $\overline{n}$ is the *numeral* associated to the natural number n (i.e., the term of the language of PA corresponding to n). Informally, $B\phi$ is satisfied by $\mathcal{N}_\Sigma$ under assignment s if and only if ϕ is logical consequence of Σ whenever replacing each free occurrence of x in ϕ by the numeral associated to $s(x)$. Then, we say that a theory T in EA is a (knowing) *entity* if $\mathcal{N}_T$ is a model of the theory of knowledge. Finally, by using Carlson's notation [7], a recursively enumerable entity is said to be a (knowing) *machine*.

3 Properties of knowing machines

While Shapiro was more involved with the description in EA of the grounds of constructive mathematics [34], Reinhardt's main intent was to use the theory of

knowledge illustrated above for studying variants of the Church-Turing Thesis. The effective version under analysis is based on a notion of *weak decidability*, according to which a property ϕ of natural numbers is weakly decidable if there exists a theorem proving procedure by an idealized human mathematician that can eventually produce each n satisfying the property, i.e., such that $\phi(n)$ is true. Such a notion is expressed in EA in terms of the modal operator B, thus leading to a definition of *weak B-decidability*: $\forall n(\phi(n) \rightarrow B\phi(n))$, i.e., given a formula ϕ with one free variable, it holds that the assignments of the variable satisfying ϕ are known. The objective is then to analyze the relationship between properties that are weakly B-decidable and the Turing Machines (TMs) that formalize the decision algorithm for these properties. Therefore, weak B-decidability expresses in the framework of EA a notion of humanly provable as discussed in Section 1, and its equivalence with respect to Turing computability is represented by the following formula (which we simply call *Turing Thesis*):

$$\exists e \forall x(B\phi \leftrightarrow x \in W_e) \tag{1}$$

where we assume that W_e is the recursively enumerable set with Gödel number represented by the PA language term e. Notice that such a statement can be seen as a constructive, effective version of Church's Thesis: a recursively enumerable set exists (and, therefore, a TM enumerates its elements) that contains all and only the assignments of x making ϕ intuitively provable, that is to say, Turing computability coincides with weak B-decidability.

As we will show later, we have that (1) is consistent in EA [30]. However, validity of the Turing Thesis implies that there exists an absolutely undecidable statement. Formally, the following theorem holds.

Theorem 1 (Incompleteness of B). *For every theory T in which (1) holds, then there exists ϕ with one free variable such that:*

$$T \vdash \exists x(\phi(x) \wedge \neg B\phi(x)).$$

Proof. Let us assume $\phi(x) = \neg(x \in W_x)$. Then, from (1) we derive:

$$\exists e \forall x(B\neg(x \in W_x) \leftrightarrow x \in W_e)$$

and, by taking $x = e$, it follows:

$$\exists e(B\neg(e \in W_e) \leftrightarrow e \in W_e). \tag{2}$$

Since, by applying $B3$, we have:

$$\forall e(B\neg(e \in W_e) \rightarrow \neg(e \in W_e)) \tag{3}$$

from (2) and (3) we obtain:

$$\exists e((B\phi(e) \leftrightarrow \neg\phi(e)) \wedge (B\phi(e) \rightarrow \phi(e))). \tag{4}$$

Now, recalling that, by applying tautologies, $B\phi(e) \rightarrow \neg\phi(e)$ and $B\phi(e) \rightarrow \phi(e)$ imply $\neg B\phi(e)$, and that $\neg\phi(e) \rightarrow B\phi(e)$ and $B\phi(e) \rightarrow \phi(e)$ imply $\phi(e)$, then it is immediate to observe that a tautological consequence of (4) is the following:

$$\exists e(\phi(e) \wedge \neg B\phi(e)) \tag{5}$$

which corresponds to the statement of the theorem.

Notice that $B3$ and (5) imply:

$$\exists e(\neg B\phi(e) \wedge \neg B\neg\phi(e))$$

because $(B\neg\phi(e) \rightarrow \neg\phi(e))$ and $\phi(e)$ hold, and therefore it must be $\neg B\neg\phi(e)$, thus stating the absolute undecidability of $\phi(e)$. Such an incompleteness result can be viewed as a version of Gödel's first incompleteness theorem, that is, if T is sound, then it is also incomplete [31].

Analogously, even Gödel's second incompleteness theorem can be recast in this setting to show that the consistency of T is absolutely unprovable. More precisely, a generalized version of Gödel's second incompleteness theorem defined in EA establishes that for any intuitively definable upper bound for intuitive provability we have an absolute impossibility of a consistency proof [31]. Formally, the following second incompleteness of B theorem holds, where we use the notation $\bar{\phi}$ to represent the symbolic expression $\bar{n}$, with n the Gödel number of the formula ϕ.

Theorem 2 (Incompleteness of B). *Assume the existence of a formula ψ with one free variable such that for every sentence σ of T it holds that:*

$$T \vdash B(B\sigma \rightarrow \psi(\bar{\sigma})) \tag{6}$$

then it also holds that:

$$T \vdash B\neg B Con\psi$$

where the predicate Con expressing consistency of its argument is defined as $\forall x\neg(\psi(x) \wedge \neg\psi(x))$.

Before proving this theorem, it is worth observing that ψ represents an intuitive upper bound for provability: it is provable that if σ is known, then ψ holds when applied to the symbolic representation of the Gödel number of σ. In fact, notice that the formula above is an instance of the B-closure of $B3$ stating the knowledge of the bounds of intuitive provability, which are established by ψ. Then, the result states that the consistency of ψ is absolutely unprovable or, in other words, no interesting upper bounds for intuitive provability can be defined. The following proof relies on a preliminary result by Reinhardt [31] stating the applicability of Gödel Fixed Point Lemma[6] in the setting of EA.

[6] Let $\psi(x)$ be any formula in the language of PA with one free variable x. Then, there exists a sentence σ of which it can be proved that $\sigma \leftrightarrow \psi(\bar{\sigma})$.

Proof. By Gödel Fixed Point Lemma, we can choose a sentence σ such that:

$$T \vdash B(\sigma \leftrightarrow B\neg\psi(\bar{\sigma})). \tag{7}$$

We now construct the derivation sequence:

$$\begin{array}{ll} \sigma \rightarrow B\neg\psi(\bar{\sigma}) & \text{by (7) and } B3 \\ \rightarrow BB\neg\psi(\bar{\sigma}) & \text{by } B4 \\ \rightarrow B\sigma & \text{by (7) and } B2 \\ \rightarrow \psi(\bar{\sigma}) & \text{by (6) and } B3. \end{array}$$

At the same time, we also have the following derivation sequence:

$$\begin{array}{ll} \sigma \rightarrow B\neg\psi(\bar{\sigma}) & \text{by (7) and } B3 \\ \rightarrow \neg\psi(\bar{\sigma}) & \text{by } B3. \end{array}$$

Hence, from $\sigma \rightarrow \psi(\bar{\sigma})$ and $\sigma \rightarrow \neg\psi(\bar{\sigma})$ we obtain $\sigma \rightarrow \neg Con\psi$ by definition of *Con*. Since it is possible to prefix with B every step of the proof above, it also holds $B(\sigma \rightarrow \neg Con\psi)$ and, equivalently, $B(Con\psi \rightarrow \neg\sigma)$, which, by applying $B2$, implies:

$$BCon\psi \rightarrow B\neg\sigma. \tag{8}$$

Now:

$$BCon\psi \rightarrow B\neg\sigma \rightarrow \neg\psi(\bar{\sigma}) \text{ by (6) and } B3.$$

Since it is possible to prefix with B every step, we have $B(BCon\psi \rightarrow \neg\psi(\bar{\sigma}))$ from which, by $B2$, it follows $BBCon\psi \rightarrow B\neg\psi(\bar{\sigma})$. From this and by virtue of $B4$ we derive:

$$\begin{array}{ll} BCon\psi \rightarrow B\neg\psi(\bar{\sigma}) & \\ \rightarrow \sigma & \text{by (7) and } B3. \end{array}$$

Then, by applying $B3$, from (8) we derive $BCon\psi \rightarrow \neg\sigma$, which, in conjunction with $BCon\psi \rightarrow \sigma$ proved above, imply $\neg BCon\psi$. Again, by prefixing B to each previous step, we get:

$$B\neg BCon\psi.$$

This result completes the part concerning the revisiting of the Gödel incompleteness theorems in EA starting from the Turing Thesis. In the following, it is our interest to consider stronger versions of the Turing Thesis involving the knowledge that can be acquired in the theory about such a thesis. The strongest claim, which we call Reinhardt's schema, establishes that the index of the TM deciding $B\phi$ in the Turing Thesis is known. This claim can be refuted in EA, as stated by the following theorem.

Theorem 3 (Reinhardt's schema [32]). $\exists eB\forall x(B\phi \leftrightarrow x \in W_e)$ *is not consistent in* EA.

Notice that 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's intuition, *I know to be a TM and I know which one.* The inconsistency of this schema can be viewed as an alternative characterization of Gödel's first incompleteness theorem.

Proof. By $B1$, from Reinhardt's schema we derive:

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

Now, assume $\phi(x) = \neg(x \in W_x)$ and $x = e$, hence we have:

$$B(B\phi \leftrightarrow \neg\phi)$$

while by the B-closure of $B3$:

$$B(B\phi \rightarrow \phi).$$

From the conjunction of the two above, by applying tautologies and distributivity, we derive $B(\phi \wedge \neg B\phi)$ and then $B\phi \wedge B\neg B\phi$, and applying $B3$:

$$B\phi \wedge \neg B\phi.$$

A weaker version of Reinhardt's schema consists of moving the outermost B operator to prefix the whole formula, thus obtaining the so-called *Strong Mechanistic Thesis*, which we call Carlson's schema from the author who proved its consistency in EA.

Theorem 4 (Carlson's schema [7]). *$B\exists e \forall x(B\phi \leftrightarrow x \in W_e)$ 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, *I know to be a TM but I do not know which one.*

To prove Theorem 4, in [7] it is shown that the theory of knowledge plus Carlson's schema is a (knowing) *machine*. In order to establish the induction proving the result, Carlson generalizes the modal operator B by introducing a collection of operators B_t, where t belongs to a linearly ordered set and can be interpreted as the amount of steps (either logical or temporal) needed to prove the formula guarded by B_t. In such a stratified version of the theory of knowledge, each formula is such that any occurrence of B_t in the scope of another occurrence $B_{t'}$ satisfies the condition $t < t'$. Then, the key parts of the proof rely on showing that the standard axiomatization of the stratified theory of knowledge for EA plus Carlson's schema is recursively enumerable and that such a result is inherited by the non-stratified theory.

Theorem 4 implies, as a corollary deriving from the application of $B3$, the validity of the Turing Thesis. Another interesting corollary of the theorem is related to the first incompleteness of B theorem.

Corollary 1. *Given T and ϕ as in Theorem 1, then:*

$$T \vdash B\exists e(\phi(e) \wedge \neg B\phi(e)).$$

Proof. By Theorem 1, (5) holds and if (1) holds whenever prefixed by B, then so does (5), because under such a hypothesis it is possible to prefix with B every step of the proof of Theorem 1. Now, it is sufficient to notice that (1) prefixed by B is exactly Carlson's schema.

Informally, this result states that knowledge of the Turing Thesis implies knowledge that there exists an absolutely undecidable sentence in EA. In particular, the corollary requires B prefixed to $B3$, which is the only axiom for B used in the proof of (5) in Theorem 1. In general, all the proofs related to variants of the Church-Turing Thesis discussed so far rely on the validity of $B(B\phi \rightarrow \phi)$, stating that in the formal system the soundness (*factivity*, as called by Alexander) of knowledge can be proved. Under such a condition, we have Theorem 4 stating that for each wff ϕ, an unspecified TM enumerates recursively the set of values for which one knows that ϕ is satisfiable (which can be read as "*I know to have some code related to* ϕ"). On the other hand, we have Theorem 3 stating that the knowledge of the identity of such a TM cannot be acquired (which can be read as "*I cannot know my own code related to* ϕ"). In between the limiting results stated by Reinhardt and Carlson, Alexander has recently proved a dichotomy revealing the relation between knowledge of factivity and knowledge of own code. Either a machine can know to be sound (that is, $B(B\phi \rightarrow \phi)$ is valid) as well as that it has some code (without knowing which, as proved by Carlson), or it can know its own code exactly (thus proving the consistency of Reinhardt's schema) but in such a case cannot know its own soundness (that is, $B(B\phi \rightarrow \phi)$ is not valid anymore). As a consequence, we emphasize that renouncing knowledge of soundness implies that the machine loses also knowledge of incompleteness.

Providing that the axioms of EA *mod factivity* consist of the axioms of EA except for the universal closure of $B3$ prefixed by B (i.e., $B(B\phi \rightarrow \phi)$ is not valid), it is possible to prove that:

Theorem 5 (Alexander [3]). *Reinhardt's schema is consistent in* EA *mod factivity.*

and then to construct the previous dichotomy.

In order to sketch the main intuition behind the proof (see [2] for details), it is worth considering the family of axioms $\Sigma(n)$, for $n \in \mathbb{N}$, which essentially consists of the axioms of EA *mod factivity* minus $B3$ and plus the additional schemes:

$$\forall x(B\phi \leftrightarrow b(x, \bar{\phi}) \in W_{\bar{n}}) \quad \textit{for any wff } \phi \textit{ with one free variable } x$$

where b is a canonical computable bijection - hence definable in the language of PA - mapping pairs (represented by the symbolic Gödel number of ϕ and its input x) to numerals. Since by completeness and compactness of EA, $\Sigma(n)$ is recursively enumerable, then, by the Church-Turing Thesis, chosen ϕ as above, for every $n \in \mathbb{N}$ there exists a total computable function $f : \mathbb{N} \rightarrow \mathbb{N}$ such that:

$$W_{f(n)} = \{b(m, \bar{\phi}) \mid \Sigma(n) \models \phi(x \,|\, \overline{m})\}$$

meaning that the set of elements that, assigned to x in ϕ, make ϕ a logical consequence of $\Sigma(n)$, is recursively enumerable by a set of index $f(n)$. Now, by Kleene's Recursion Theorem, it holds that there exists $e \in \mathbb{N}$ such that $W_e = W_{f(e)}$.

Hence, $\mathcal{N}_{\Sigma(e)}$ satisfies $\forall x(B\phi \leftrightarrow b(x,\bar{\phi}) \in W_{\bar{e}})$, because for each s with $s(x) = m$ it holds:

$$\begin{array}{rl} \mathcal{N}_{\Sigma(e)} \models B\phi[s] & \Leftrightarrow \\ \Sigma(e) \models \phi(x\,|\,\overline{s(x)}) & \Leftrightarrow \\ \Sigma(e) \models \phi(x\,|\,\overline{m}) & \Leftrightarrow \\ b(m,\bar{\phi}) \in W_e & \Leftrightarrow \\ \mathcal{N}_{\Sigma(e)} \models b(m,\bar{\phi}) \in W_{\bar{e}} & \Leftrightarrow \\ \mathcal{N}_{\Sigma(e)} \models (b(x,\bar{\phi}) \in W_{\bar{e}})[s]. & \end{array}$$

In addition, by construction, $\mathcal{N}_{\Sigma(e)}$ satisfies also all the instances of the other axioms of $\Sigma(e)$. Then, for each axiom ϕ of $\Sigma(e)$, it holds that $B\phi$ is satisfied by $\mathcal{N}_{\Sigma(e)}$ because $\Sigma(e) \models \phi$ and ϕ is a sentence, so that, by definition of $\mathcal{N}_{\Sigma(e)}$, it follows $\mathcal{N}_{\Sigma(e)} \models B\phi$. Notice that, by virtue of this condition, $\mathcal{N}_{\Sigma(e)}$ satisfies Reinhardt's schema. To complete the proof, it remains to show that $\mathcal{N}_{\Sigma(e)}$ satisfies $B3$, which is not included in $\Sigma(e)$, i.e., $\mathcal{N}_{\Sigma(e)} \models B\phi \rightarrow \phi$. Supposing $\mathcal{N}_{\Sigma(e)} \models B\phi[s]$ for some s, then $\Sigma(e) \models \phi(x\,|\,\overline{s(x)})$ and, since $\mathcal{N}_{\Sigma(e)} \models \Sigma(e)$, we derive $\mathcal{N}_{\Sigma(e)} \models \phi(x\,|\,\overline{s(x)})$.

Summarizing, we have described a recursively enumerable set of sentences (including EA *mod factivity* and Reinhardt's schema) that turns out to be an entity, which represents the result expressed by Theorem 5.

In the framework of such a machine, we show a result related to a specific case. Consider $\phi(x) = (x \in W_x)$ and $x = e$. Thus from:

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

we derive:

$$\exists e B(B\phi(e) \leftrightarrow e \in W_e)$$

and:

$$B(B\phi(e) \rightarrow \phi(e))$$

which expresses a limited form of knowledge of soundness allowed in EA *mod factivity*. More precisely, we have a specific function ϕ for which the related machine knows its own code and knows to be sound with respect to a specific input x. We notice that, by definition, ϕ turns out to represent an instance of an interpreter function[7] and the specific input x is the index of the machine itself. Therefore, by following the same intuitions derived from Reinhardt and Carlson theorems, "*If I am a universal TM knowing my own code, then I know the soundness of the computation provided when I am fed with my own code*". By virtue of the previous results and according to Alexander's dichotomy, knowledge about other TMs is much more limited. As a consequence, in general a universal TM knowing its own code cannot know the soundness of what it proves when

[7] Given any gödelization of functions, an interpreter f_u is a function mimicking the behavior of any other function, i.e., $f_u(x,y) = f_x(y)$. As an example, the universal TM is an interpreter. We recall that interpreters represent a classical tool in computability theory and play a fundamental role for programming languages.

interpreting other TMs. On the other hand, if a (universal) TM can prove the soundness of its knowledge, then it cannot know its own code.

In our opinion, this is an interesting enhancement of the tradeoff result provided by Alexander representing an additional formal element for the analysis of the Gödelian Arguments.

4 Conclusion

Historically, the idea of a provability logic is first discussed by Gödel [13, 14] with the aim of defining a formal semantics for intuitionistic truth [36]. Gödel's calculus is based on propositional logic and on the modal operator $\Box$, and is basically equivalent to S4 [22]. The resulting notion of provability shall not be confused with formal deducibility for theories including PA, as such a correspondence leads to a contradiction [14]. Following Gödel's attempt, several approaches have been proposed either to define axiom systems for formal deducibility, or to find an exact provability semantics for S4-like modal systems. For a comprehensive survey, we refer, e.g., to [4].

In this setting, the theory of knowing machines offers a proof-theoretic framework to reason about the notions of intuitive provability and consistency of TMs. In practice, the extended results show precisely the relation between self-awareness of soundness and of own code in the setting of knowing machines, and show some compatibility with philosophical arguments like the following suggestions due to Gödel himself [38]:

> 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.

The results provided in EA are obtained by following formal lines of reasoning even if starting from the definition of an epistemic notion of intuitive provability. In fact, on one hand, it is worth noticing that the modal operator B necessitates an atomic treatment of formulas of the form $B\phi$ and allowing only for an axiomatic representation of its properties. Hence, the lack of a precise model-theoretic semantics represents a weakness. However, on the other hand, it is also worth observing that all the proofs of the results provided by Reinhardt, Carlson, and ourselves derive from the formal application of deductive reasoning and proof theoretic techniques. The unique exception is given by Theorem 5, which relies on the application of the Church-Turing Thesis (see, as an example, [33] for the informal aspects of practical uses of the Church-Turing Thesis).

Even in the setting of EA as defined by Shapiro [34], the consistency of epistemic variants of Myhill's version of the Church-Turing Thesis [28] can be demonstrated [11, 16, 20]. Such variants are captured by the following informal statement: if it is intuitively provable that for each x there exists y such that it holds that $\phi(x, y)$ can be proved, then ϕ determines a total recursive function.

This is analogous to the Turing Thesis discussed in this paper and is independent from the formalization of Carlson's schema, which represents the strongest mechanistic position demonstrated in the framework defined by Reinhardt.

The abstractions behind the notion of intuitive provability, which cover all the aspects related to space and time constraints [31], do not allow for reasoning about the relation between knowability and complexity. As future work, it would be interesting to investigate such a relation, in order to formalize what informally stated, e.g., by Benacerraf [5]:

> 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.

References

1. A. Aldini, V. Fano, and P. Graziani. A note on knowing machines. In F. Gadducci and M. Tavosanis, editors, *Preliminary Proceedings of the 3rd Int. Conf. on the History and Philosophy of Computing (HaPoC 2105)*, pages 15–17. Pisa University Press, 2015.
2. S. Alexander. *The Theory of Several Knowing Machines*. PhD thesis, Ohio State University, 2013.
3. S. Alexander. A machine that knows its own code. *Studia Logica*, 102:567–576, 2014.
4. S.N. Artemov and L.D. Beklemishev. Provability logic. In D. Gabbay and F. Guenthner, editors, *Handbook of Philosophical Logic, Second Edition*, volume 13, pages 229–403. Kluwer, 2004.
5. P. Benacerraf. God, the devil and Gödel. *The Monist*, 51:9–32, 1967.
6. L. BonJour. *The Structure of Empirical Knowledge*. Harvard University Press, 1985.
7. T.J. Carlson. Knowledge, machines, and the consistency of Reinhardt's strong mechanistic thesis. *Annals of Pure and Applied Logic*, 105:51–82, 2000.
8. C.S. Chihara. On alleged refutations of mechanism using Gödel's incompleteness results. *The Journal of Philosophy*, 69:507–526, 1971.
9. 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. S. Feferman et al., editors. *Kurt Gödel Collected Works*, volume I. Oxford University Press, 1986.
11. R. Flagg. Church's thesis is consistent with epistemic arithmetic. In S. Shapiro, editor, *Intensional Mathematics*, pages 121–172. North-Holland, 1985.
12. K. Gödel. Über formal unentscheidbare Sätze der Principia mathematica und verwandter Systeme. *Monatshefte für Mathematik und Physik*, 38:173–198, 1931. En. Tr. in [10], pages 144-195.
13. K. Gödel. Zum intuitionistischen Aussagenkalkül. *Akademie der Wissenschaften in Wien, Mathematisch-naturwissenschaftliche Klasse, Anzeiger*, 69:65–66, 1932. En. Tr. in [10], pages 222-225.
14. K. Gödel. Eine Interpretation des intuitionistischen Aussagenkalküls. *Ergebnisse eines mathematischen Kolloquiums 4*, pages 39–40, 1933. En. Tr. in [10], pages 300-303.

15. A. Goldman. What is justified belief? In George Pappas, editor, *Justification and Knowledge: New Studies in Epistemology*, pages 1–23. D. Reidel Publishing, 1979.
16. N.D. Goodman. Flagg realizability in arithmetic. *Journal of Symbolic Logic*, 51(2):387–392, 1986.
17. J. Hawthorne. *Knowledge and Lotteries*. Oxford University Press, 2004.
18. J. Hintikka. *Knowledge and Belief: An Introduction to the Logic of the Two Notions*. Cornell University Press, 1962.
19. J. Hintikka. Knowing that one knows. *Synthese*, 21:141–162, 1970.
20. L. Horsten. In defense of epistemic arithmetic. *Synthese*, 116(1):1–25, 1998.
21. S.C. Kleene. *Mathematical Logic*. Wiley, 1967.
22. C.I. Lewis and C.H. Langford. *Symbolic Logic*. Dover, 1932.
23. M.H. Löb. Solution of a problem of Henkin. *Journal of Symbolic Logic*, 20:115–118, 1955.
24. J.R. Lucas. Minds, machine and Gödel. *Philosophy*, 36:112–127, 1961.
25. J.R. Lucas. Satan stultified: a rejoinder to Paul Benacerraf. *The Monist*, 52:145–158, 1968.
26. R. Montague. Syntactical treatments of modality. *Acta Philosophica Fennica*, 16:153–167, 1963.
27. J. Myhill. Some remarks on the notion of proof. *Journal of Philosophy*, 57:463–471, 1960.
28. J. Myhill. Intensional set theory. In S. Shapiro, editor, *Intensional Mathematics*, pages 47–62. North-Holland, 1985.
29. R. Penrose. Beyond the doubting shadow. *Psyche*, 2-1:89–129, 1996.
30. W. Reinhardt. The consistency of a variant of Church's thesis with an axiomatic theory of an epistemic notation. In *Proceedings of the 5th Latin American Symposium on Mathematical Logic, Revista Colombiana de Matematicas, vol. XIX*, pages 177–200, 1981.
31. W. Reinhardt. Absolute versions of incompleteness theorems. *Noûs*, 19(3):317–346, 1985.
32. W. Reinhardt. Epistemic theories and the interpretation of Gödel's incompleteness theorems. *Journal of Philosophical Logic*, 15:427–474, 1986.
33. L. San Mauro. The informal side of computability: Church-Turing thesis, in practice. In F. Gadducci and M. Tavosanis, editors, *Preliminary Proceedings of the 3rd Int. Conf. on the History and Philosophy of Computing (HaPoC 2105)*, pages 83–84. Pisa University Press, 2015.
34. S. Shapiro. Epistemic and intuitionistic arithmetic. In S. Shapiro, editor, *Intentional mathematics*, pages 11–46. North-Holland, 1985.
35. S. Shapiro. Incompleteness, mechanism, and optimism. *The Bulletin of Symbolic Logic*, 4:273–302, 1998.
36. A.S. Troelstra and D. van Dalen. *Constructivism in Mathematics, Volumes 1 and 2*. North-Holland, 1988.
37. A. Turing. Computing machinery and intelligence. *Mind*, 59:433–460, 1950.
38. H. Wang. *From Mathematics to Philosophy*. Humanities Press, 1974.