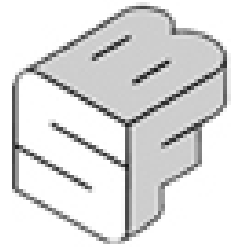
A note on knowing machines

Alessandro Aldini [1] *Vincenzo Fano* [1] *Pierluigi Graziani* [2]

1506 UNIVERSITÀ DEGLI STUDI DI URBINO CARLO BO | DISBEF DIPARTIMENTO DI SCIENZE DI BASE E FONDAMENTI



Ud'A UNIVERSITÀ DEGLI STUDI G. D'ANNUNZIO CHIETI PESCARA

HAPOC

10 October 2015

Outline

Where is the focus...

Setting and motivation

- Gödel's incompleteness results
- Gödelian arguments
- Informal and semi-formal approaches to Gödelian arguments
- A proof-theory based approach to Gödelian arguments:
  Epistemic Arithmetic (EA)

Where is the focus...

Setting and motivation

- Gödel's incompleteness results
- Gödelian arguments
- Informal and semi-formal approaches to Gödelian arguments
- A proof-theory based approach to Gödelian arguments: Epistemic Arithmetic (EA)

Goal

Illustrating the basics of EA and the results that can be achieved in the setting of "knowing" machines (with a small extension)

Gödelian arguments

Background

- Turing 1950: mechanistic project (mechanical simulation of human mind)
- Incompleteness results by Gödel: used to argue refutations of mechanism, as intended both extensionally and intensionally
  - Informal anti-mechanistic theses [Lucas 1961, Penrose 1989]
  - Semi-formal approaches [Benacerraf 1967, Chihara 1971]

Such approaches preserve intensional elements on properties of human mind

Hard still to define precisely what the (anti-)mechanistic claims shall state

Gödel standpoint

Gödel disjunction [Gibbs lectures 1951, published in Gödel Collected Works in 1995]

(1) *. . . human intelligence infinitely surpasses the powers of the finite machine (TM), and there are no absolutely unsolvable Diophantine problems . . .* [a]

OR

(2) *. . . human intelligence is representable through a finite machine (TM) and there are absolutely irresolvable Diophantine problems for it . . .*

[a] He was convinced that (1) held

Gödel standpoint

Gödel [1972]

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

Reinhardt et al.

Reasoning about knowledge more formally

- *Intuitive provability* is a kernel notion in (anti-)mechanistic claims and conjectures, like the Post-Turing thesis (*"humanly provable" is equivalent to provability by some Turing machine*)
- Idea: expressing this notion in an appropriately formulated formal language [Reinhardt 1981, Shapiro 1982] and the related properties by means of axioms

K: a modal operator for knowledge

Intuition

- Formalize the epistemological idea of *knowability / provability*
  1. avoid a model-theoretic definition of knowledge
  2. define an epistemic notion of *intuitive provability* as a modal operator K and then describe its properties in terms of axioms
- So, the satisfiability of $\mathcal{M} \models K\phi[s]$ can be read as:

$$\phi \textit{ is known when the free variables of } \phi$$
$$\textit{are interpreted according to assignment } s$$

Properties of K

Intuitively:

- 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

Formally:

B1. $K\forall x\phi \to \forall xK\phi$

B2. $K(\phi \to \psi) \to K\phi \to K\psi$

B3. $K\phi \to \phi$

B4. $K\phi \to KK\phi$

Theory of knowledge

Axioms:

- $B1$-$B4$
- 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. $\forall y_1 \ldots \forall y_n((\phi(x|\mathbf{0}) \wedge \forall x(\phi \rightarrow \phi(x|\mathbf{S}(x)))) \rightarrow \forall x \phi)$
- K-closure of any of the previous axioms

Reinhardt's result

Goal

- Investigate the relation between *intuitively weak decidability*:

  for all x satisfying a formula ϕ, $\phi(x)$ is provable

  and the Turing Machine realizing such a decision algorithm
- Approach: represent *intuitively weak decidability* by *weakly K-decidability* (the assignments of x satisfying ϕ are known)

Reinhardt's result

Theorem (Turing's thesis)

$\exists e \forall x (K\phi \leftrightarrow x \in W_e)$ *is consistent in* EA

Corollary (On Gödel's first incompleteness theorem)

In any theory T *in which the previous statement holds, it also holds:*

$$T \vdash \exists x(\phi \wedge \neg K\phi)$$

Theorem (On Gödel's second incompleteness theorem)

If $\exists \psi(x)$ *such that for all sentences* σ *of* T *with Gödel number* $\overline{\sigma}$ *it holds:*

$$T \vdash K(K\sigma \rightarrow \psi(\overline{\sigma}))$$

then:

$$T \vdash K \neg K\, Con\psi$$

Reinhardt's result

Theorem (Reinhardt's schema)

$\exists e K \forall x (K\phi \leftrightarrow x \in W_e)$ *is not consistent in* EA

Proof.

By $B1$ we derive: $\exists e \forall x K(K\phi \leftrightarrow x \in W_e)$
Assume $\phi(x) = \neg(x \in W_x)$ and $x = e$, hence:

$$K(K\phi \leftrightarrow \neg\phi) \tag{1}$$

while by the K-closure of $B3$:

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

From 1 and 2, by applying tautologies and distributivity, we derive $K(\phi \wedge \neg K\phi)$ and then $K\phi \wedge K\neg K\phi$, and applying $B3$:

$$K\phi \wedge \neg K\phi$$

Carlson's result

Theorem (Carlson's schema)

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

Using Carlson's notation, EA plus the schema above is a *knowing* machine

Alexander's result

A dichotomy about machines

- Reinhardt and Carlson proofs of results rely on knowledge of $B3$
- What if we get rid of $K(K\phi \to \phi)$

Theorem (Alexander's schema)

$\exists e K \forall x (K\phi \leftrightarrow x \in W_e)$ *is consistent in* EA *minus* $K\,B3$

Summarizing

From the results above

- A TM exists that enumerates the assignments making ϕ provable, or:

  I know that the set of x for which
  I know $\phi(x)$ is recursively enumerable

- however, I do not know what TM it is [Strong Mechanistic Thesis], or:

  I know I am a Turing machine, but do not know which one
  [Benacerraf 1967]

- Hence, if mechanism is true, then we cannot know it with mathematical certainty (thus confirming Gödel conjecture)
- The knowledge of such a TM can be acquired by renouncing to the knowledge of the *factivity* (soundness) of the TM, or:

  A factive knowing machine cannot know
  its own code and its own factivity [Alexander, 2014]

Extending the dichotomy

A specific case

- consider an interpreter $f_u(x, y) = f_x(y)$ and take $x = y$
- consider Alexander's knowing machine and $\phi_x(x)$ a w.f.f. based on term $f_x(x)$:

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

and then consider the case $x = e$, hence:

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

from which we derive:

$$K(K\phi_e(e) \leftrightarrow \phi_e(e))$$

and:

$$K(K\phi_e(e) \rightarrow \phi_e(e))$$

Extending the dichotomy

A specific case

- the knowing machine knows its own code and is aware of the soundness of the knowledge resulting by interpreting its own code, or:

If I know which universal TM I am,
then I know the soundness of what I prove with respect to my code

Conclusion

Conclusion and future work

- the theory of knowing machines offers a proof-theoretic framework to reason about the notions of intuitive provability and consistency of TMs
- what about the relation between provability and complexity

> *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*
> *[Benacerraf 1967]*