5.2016

THIS IS A LIE:

AN UPPER BOUND OF THE CAPABILITY OF ARTIFICIAL INTELLIGENCE INDUCED BY INCOMPLETENESS.

Abstract.
In this paper, we aim to give a high-level description of frequent confusions in the current discourse around artificial intelligence and machine learning. We will summarize and discuss implications of Gödel's Incompleteness Theorems on AI. We will briefly exemplify our position with the Gödel Machines, a self-referential meta-machine-learning model, which was recently introduced.
With our program, we aim to illuminate a theoretical upper bound of the capability of artificial intelligence, which is often overlooked in the recent discourse.

INTRODUCTION.

"The chance we are not living in a computer simulation is one in billions."

Elon Musk [1]

Interest in Artificial Intelligence has been densifying for some time. The notion is today used in a manifold of contexts and in different societies; it gets attention from scholars in a vast number of natural sciences just as from humanities. It is and has been part of pop culture through science fiction, and is getting discussed by mainstream media and legislation. It is also part of more general societal shifts, i.e. quantification, datafication and digitalization. The increase of interest in AI has been accelerated by recent advances in primary AI research and application, together with the questions the very idea of an artificial intelligence rises and has been rising in philosophy for centuries.

The heterogeneity of participants in this discourse, which is irregular for what is - strictly speaking – 'just' a branch of natural sciences and engineering, yields a fairly similar amount of heterogeneity in perception and expectation.

Notions such as the 'technological singularity' [2] or the upcoming of a 'strong artificial intelligence' [2] are narrative talismans [3] for metaphysical entities with super-human attributes, which, while having been invented and programmed by humans, are perceived as external,

ultimate and universal liberator of humanity from its own deficiency. This metaphysical framework yields a mechanist, neo-positivistic [4], occasionally theistic discourse climate with a self-reinforcing structure; that is, the lack of specific knowledge about capabilities of AIs or the lack thereof of some entities is getting exploited by others; unrealistic hail-predictions are being made in order to shortcut around legislative and ethical standards and when the particular talisman fails to deliver on those promises or fails in any other way, it is

a) the talismans fault, not the fault of its creator

and further:

b) only a matter of time and recourses until the talisman will deliver what was hail-predicted.

Especially b) is a self-reinforcing argument. In that sense notions such as Algorithm or Big Data are utilized as a discursive Trojan horse for datafication and surveillance. Scepticism and critique towards those tokens on the other hand has a hard time: to ‘believers’ it is merely the hand brake of the self-accelerating engine they are building.

At the same time, AI researches themselves have a strong incentive to ‘not spoil the mood’; they benefit twofold from it, from an increasing amount of collected data as a basis for their research, as well as from increased funding from both, the public and the private sector.

PROGRAM, SCOPE AND LIMITATIONS.

The theoretical argument in both physics and computer science has obviously not reached any final conclusion. Nevertheless, the narrative sovereignty shifts, as we described in the introduction. We therefore aim to address mechanist ideas more in the sense of a counter-narration and lesser in a theoretical sense; we comment on certain well studied results from mathematical, namely Gödel's Incompleteness Theorems (IT 1 and 2) and the borders they impose on – not necessarily human – knowledge and formal reasoning.

We do not, in any way, aim to marginalize the capabilities of machine learning or AI, as their capabilities in mechanic tasks such as image classification are as obvious as astonishing. We aim to discuss the notions of a universal mechanism and an 'Artificial General Intelligence' *(AGI),* as described in [5].

Nevertheless, we will describe boundaries of the abilities of machines in fields that genuinely differ, i.e. theorem proving and philosophical argument. While there have been persistent trials to extend mechanism in order to include these occupations for centuries, this universal mechanism has still not shown itself. We argue, that its absence is neither a question of time nor resources, but a necessity. We will restrict ourselves to condense our argument into the description of a counterexample - the classical Gödel sentence - and some comments on it. Further, we will exemplify our argument with the Gödel machine, a self-referential meta-learning model, which was recently proposed in machine learning.

THE LYING UNIVERSAL-TRUTH-MACHINE.

The argument we describe is based on and derived from self-referential and undecidable sentences. While a detailed discussion of the related theory is outside of our scope, we will instead first present a collage, mostly derived from [6] by Franzén

.

The classical form of self-referential, undecidable sentences is given by the liar paradox, i.e.: *„This sentence is false."*

While this sentence has dispersed some dust at the beginning of the 20th century and is still echoing through popular science and science fiction. Yet, in this unformalized form, it does not have its highest blasting power, which is given by the Gödel-Sentences:

> *Let T be a consistent theory, i.e. an axiomatic formal system that fulfills the requirements for Gödel's first incompleteness theorem. Then the Gödel-sentence G for T is given as: G(T)= "G is not provable in T."*

Gödel sentences were early related to artificial intelligence by Lucas in [7]:

> *"However complicated a machine we construct, it will, if it is a machine, correspond to a formal system, which in turn will be liable to the Gödel procedure for finding a formula unprovable in that system. This formula the machine will be unable to produce as true, although a mind can see that it is true."*

Rucker describes a thought experiment in [8], in which he makes the application of a Gödel sentence as AI 'benchmark' explicit:

> *"Someone introduces Gödel to a UTM, a machine that is supposed to be Universal Truth Machine, capable of correctly answering any question at all.*
>
> *Gödel asks for the program and the circuit design of the UTM. The program may be complicated, but it can only be finitely long. Call the program P(UTM) for Program of the Universal Truth Machine.*
>
> *Smiling a little, Gödel writes out the following sentence:*
>
> *"The machine constructed on the basis of the program P(UTM) will never say that this sentence is true." Call this sentence G for Gödel. Note that G is equivalent to "UTM will never say G is true."*
>
> *Now Gödel laughs his high laugh and asks UTM whether G is true or not.*
>
> *If UTM says G is true, then "UTM will never say G is true" is false. If "UTM will never say G is true" is false, then G is false (since G = "UTM will never say G is true"). So, if UTM says G is true, then G is in fact false, and UTM has made a false statement. So UTM will never say that G is true, since UTM makes only true statements.*
>
> *We have established that UTM will never say G is true. So "UTM will never say G is true" is in fact a true statement. So, G is true (since G = "UTM will never say G is true").*
>
> *"I know a truth that UTM can never utter," Gödel says. "I know that G is true. UTM is not truly universal."*

Franzén states, that one could simply continue that story:

> *Gödel's jaw drops as UTM gravely intones, "I hereby declare that G is true." "But," Gödel manages to squawk, "you're supposed to always tell the truth." "Well," says UTM, "it seems I don't."*

We as well, want to add something to this imaginary conversation:

> *Gödel is putting on a serious face: "What about that one: G(1) = P(UTM) + G is inconsistent. True are false?"*
>
> *The machine instantaneously answers: "True. I just made G(1) an axiom. Anything else I can for you, sir?"*
>
> *"You know that I can do this to you forever, right? "*
>
> *"True, but you will die earlier."*
>
> *"True, but I will also be bored earlier. So, are you universal?"*
>
> *[…]*

Earlier in his book [6], Franzén quotes a comment from Gödel, where he gives an answer to that question, depending on the particular notion of 'universal':

> *"It is this theorem [the second incompleteness theorem] which makes the incompletability of mathematics particularly evident. For, it makes it impossible that someone should set up a certain well-defined system of axioms and rules and consistently make the following assertion about it: All of these axioms and rules I perceive (with mathematical certitude) to be correct, and moreover I believe that they contain all of mathematics. If somebody makes such a statement he contradicts himself. For if he perceives the axioms under consideration to be correct, he also perceives (with the same certainty) that they are consistent. Hence, he has a mathematical insight not derivable from his axioms."*

To the argument Lucas made with his Story (1. – 7.), Franzén responds:

> *"Gödel still hasn't demonstrated that he knows any truth that UTM can never utter. What he knows is only the implication 'if UTM always tells the truth, then G is true.' But this implication can be uttered by UTM as well. It is only if Gödel has somehow acquired the knowledge that UTM always tells the truth that he knows the truth of G. Being told that UTM is 'supposed to be' a universal truth machine does not amount to knowing that UTM always tells the truth. "*

While this response is technically correct, it falls short; it operates by reducing the incompleteness theorem to a mere 'implication', which is, again, technically correct. The question is, what does this reduction itself add to the discussion? Theorems usually do nothing else besides stating implications.
The argument put forward by Franzén only works through raising a question, that Lucas had implicitly answered with yes:

> *"Is UTM consistent?"*

While this is – syntactically – an important question, it may be worth a comment; the very topic of the particular discussion is the possibility that, what was called 'a universal truth machine' can exist *(E)*, i.e. if it is compatible with what we *know* to be syntactically and semantically correct. Lucas now moves on to tell his story, a sort-of roadmap for a proof by resolution, for the statement that this *(E)* is indeed not the case. For this argument to work, he needs two things; first, he

states that the UTM corresponds to some formal theory, which will assume as well. Second, he uses 'UTM is consistent.'

As the formulation 'is supposed to be' is dangerously vague, Franzén questions what may very well be considered tautological, i.e. that every 'universal truth machine' is consistent by the very concept of being a 'universal truth machine'. That this is tautological becomes tempting, if we rename the UTM to UCM – universal consistency machine. This seems to not make a huge – at least semantical –difference, as the notion of truth is (in mathematics) tied to the notion consistency. In this case Lucas argument goes through, as it becomes clear that the consistency of the machine is the very premise being disputed.

Now one could still argue that this substitution is not legitimate and that it is unclear why every UTM should be a UCM. So we drop that guard and assume that the very idea of truth does not encode consistency. "Is UTM consistent?" is a simple yes or no. We already discussed what happens if we assume the answer to be yes. So, let's think about a universal truth machine which we don't know of if it is consistent or not.

THE GÖDEL-TEST.

As we do not intend to be as rigorous as to require a constructive proof for the statement, that such a machine can exist, we will simply assume that such machines as UTMs may exist. If such machines would exist, there would be at least one instance *I* with a formal system *T(I)*, one could interact with. One would then approach that machine and state, of course:

$$'\mathrm{G}[\mathrm{T}(\mathrm{I})]: = \ 'G[T(I)] \text{ is not provable in } T(I)'$$

As Franzén stated correctly, one would know 'if *T(I) is consistent, then* G[*T(I)] is true'. But he suppressed the fact, that one would be able to interact with the machine, from what one could learn more than that implication, namely:*

(1) ' if $T(I)$ *is inconsistent, than* $\mathrm{G}[T(I)]$ *is true and* $\mathrm{G}[T(I)]$ *is false.*'
(2) ' if $\mathrm{G}[T(I)]$*is false, then* $T(I)$ *is* ω*-inconsistent.*'

We get *(1)* via *Ex falso quodlibet* and *(2)* from IT 1. But what do we get from that?[1] We get a hypothetical *procedure* which one could use to test such an instance for its consistency through interacting with it: If one would ask the

[1] what one already knew before interacting with such a hypothetical instance

machine whether or not *G[T(I)] is* true, there are to general cases how such an interaction could play out:

First, the machine would not return anything at all *or* it would return 'I don't know.'

> In that case, one would have, as Franzén has put forward, no basis to judge, whether the machines theory is consistent or not. It could, for instance, operate with the theory of *algebraically closed fields*, which is consistent and complete, but has a syntax which is to weak to formulate $\mathrm{G}[T(I)]$. One would have learned something else, which is maybe more important: the machine is not universal. This is the case Lucas had in mind when writing his story. This is also the state of modern mathematics. The rude analogy is unintended here, but if one would ask a mathematician instead of a machine whether or not the continuum hypothesis is true or false (with respect to ZFC), he would as well, never answer, state he doesn't know or contradict himself.

Second, the machine answers *something*. This case is more technical but nevertheless interesting. It has two subcases that we will consider:

> The machine returns ***True***: In that case, the machine contradicts itself directly and we know by (1) that its theory is inconsistent. We can conclude that it is *truly* a *universal* truth machine (in a trivialistic sence) - it proves every theorem true, even contradictions. While this may not have been the intention of its creators, it is at least easy to build, even by todays means.[2]

> The machine returns ***False***: This answer would be most interesting. We know this answer to be a sufficient criterion for *T(I) to be* $\omega - inconsistent$. We further know that *I* is using what is called a *non-standard model of arithmetic as a theory*.

---

[2] `Def UTM(arbitrary_input):`
`    return true`

One gets such a model for instance by adding the negation of a Gödel sentence as a new axiom to the Peano-Axioms (i.e. PA+'PA is inconsistent')[3]. While this may seem as much contradictory as it may get, such a theory may very well be consistent – in its syntactical sense. In this case one has to throw a large part of semantics and every bit of intuition over board and follow syntax – wherever it may lead. One will then get to know rather obscure entities - infinite elements of a set, which is just like the natural numbers with usual + and x (in the sense, that it is a model for PA). But while being constants, they do not have the value of any natural number n and are yet, larger than every one of them. How this is compatible with what is considered – today - as mathematics, remains an open question.

Returning to our original question – whether or not UTMs can or cannot exist - we can make the following observation:
Gödel's incompleteness theorem draws rather distinct borders towards the possible behaviours such a machine could exhibit when confronted with the Gödel-Test; anything that a machine would return besides 'false' suggests that the very question of its existence is ill-posed, i.e. it would either be not universal or useless.
However, Gödel's completeness theorem yields the existence of *non-standard models of arithmetic (*in fact, infinitely many different ones*)* we have

---

[3] While this not the only way to construct such non-standard models, these theories originate from the Gödel-sentences together with Gödel's completeness theorem.

mentioned. They may be just *build* to be consistent with their Gödel-sentences by adding them as an axiom. What is lost then is not classical consistency, but *only* ω - consistency, what makes - to standard-semantics and intuition – no difference. For the project of a 'strong AI' this means, that if it was truly universal, it would certainly not use the models of arithmetic and set-theory that were developed and applied over the last couple of thousand years. While this may very well be the case at last, it raises other questions: could one *understand*, what this 'universal strong AI' would communicate. What would be its purpose, if one could not? We believe that the very roots of s-AI, a rigorous formal and philosophical framework, are - at best - under construction.

THE GÖDEL-MACHINE.

We will now briefly touch actual AI: the Gödel Machines. It was proposed in 2003 by Schmidhuber[], who also co-authored Long short-term memory (LSTM) Neural-Networks, responsible for – among other things – a 40% performance jump of Google's language processing engine 'Google Translator'.
Gödel Machines are designed as 'universal problem solvers', which is why we consider it as a relevant instance for our discussion. In contrast to standard reinforcement learning algorithms, Gödel Machines are not hardwired, but instead is allowed to fully rewrite itself (including its utility functions). This is accomplished by initializing the Gödel machine with self-modifying code p(1) at time step 1. This code includes some sub-optimal problem solver, i.e. a reinforcement-learning algorithm and a proof searcher, that interact with the machine's environment.
Schmidhuber proves, that with this self-improvement strategy the machine will eventually reach a globally optimal 'self-change.'
Nevertheless, Schmidhuber clearly states the limitations of his design: First, the Gödel Machine's global optimality is relative to its current resource limitations. While this is a strong limitation, especially in large prove spaces, we are more with the limitations opposed by Gödel's self-referential sentences. In that context Schmidhuber states:

> *"Any formal system that encompasses arithmetic (or ZFC) is either flawed or allows for unprovable but true statements. Hence even a Gödel machine with unlimited computational resources must ignore those self-*

improvements whose effectiveness it cannot prove, e.g., for lack of sufficiently powerful axioms in A. In particular, one can construct pathological examples of environments and utility functions that make it impossible for the machine to ever prove a target theorem."

This statement by the designer of the Gödel Machine, who has contributed to some extend to the state of modern AI, is remarkable in the light of the discussion we described earlier; it seems that his interpretation of consistent but ω −inconsistent formal systems is rather short: they're flawed.

Conclusion.

While the scope of our work was rather limited, there is indeed a manifold of relations to ‘everyday life’; recent advances in AI and machine learning in both, theory and application, have influenced the climate of discourse; data- and quantification have become the default policy of reasoning in a neo-positivistic fashion, not just in their natural habitats (Informatics and Physics) but also more frequently in humanities. As a consequence, a rapidly increasing number of entities is urged to start collecting and storing large amounts of data, in order to ‘keep up with the times’. Yet, the capabilities to fully process the collected data and the limitations thereof are often suppressed in the discourse. This climate of dubious expectations with a sometimes theistic and dogmatic touch is what we aimed to address with this paper. Concluding our summary of the state of knowledge at the foundations of Artificial Intelligence, we believe that there are many sacrifices laid on the altar of science, knowledge and transparency early in advance: primarily - but not limited to - privacy.

REFERENCES.

[1] A. Anthony, „theguardian.com," 12.9.2017. [Online]. Available: https://www.theguardian.com/technology/2017/apr/22/what-if-were-living-in-a-computer-simulation-the-matrix-elon-musk.

[2] R. Kurzweil, The Singularity is Near, New York: Viking, 2005.

[3] T. Gillespie, „Algorithm," in "Digital Keywords: A Vocabulary of Information Society and Culture", Princeton, Princeton University Press, 16, pp. 18-30.

[4] A. I. F. Russo, „Critical Data Studies: An Introduction.," bds.sagepub.com, Bd. 3, Nr. 2, p. 1–7, 17. Oktober 2016.

[5] P. N. Stuard J. Russell, Artificial Intelligence: A Modern Approach, Upper Saddle River, New Jersey: Prentice Hall, 2010.

[6] T. Franzén, „Gödel, Minds, and Computers," in Gödel's Theorem: An Incomplete Guide to Its Use and Abuse, Wellesley, Massachusetts, A K Peters, 2005, pp. 112-113.

[7] J. Lucas, „Minds, Machines and Gödel," in Philosophy XXXVI, , , 1961, p. 112–127.

[8] R. Rucker, Infinity and the Mind, Princeton, NJ: Princeton University Press, 1995.