sciendo

STUDIES IN LOGIC, GRAMMAR AND RHETORIC 69 (82) 2024

DOI: 10.2478/slgr-2024-0028

Adam Olszewski
Pontifical University of John Paul II in Cracow
Kraków, Poland
e-mail: adam.olszewski@upjp2.edu.pl
ORCID: 0000-0003-3069-7518

THE LIMITS OF COMPUTER SCIENCE. WEIZSÄCKER'S ARGUMENT

Abstract. The main purpose of this paper, which takes the form of an essay, is an attempt to answer the question of the limits of artificial intelligence (AI). In the introductory section, we present the key milestones in AI development, both historical and future projections, in which two terms – Artificial Human (AH) and Artificial 'god' (AG) – play a special role. In the second section, we clarify the question of the limits of AI by indicating the hypothetical goal of AI development. The third section develops the argument proposed by C. F. Weizsäcker, originally formulated for cybernetics. The conclusion of this argument is optimistic about limitations to the possibilities of cybernetic simulations. We apply this argument to AI and subject it to a critique which ultimately undermines the legitimacy of its conclusion. We base the critique on two well-known results: the theorem of the unsolvability of the halting problem and Gödel's first incompleteness theorem, and we formulate two objections interpreted without adopting Church's thesis. In the crucial fourth section, we present a third objection in the form of a hypothesis for which we argue that AI (AH), understood as a subject, will always be solipsistic.

Keywords: artificial intelligence, artificial human, solipsism, limits of AI, argument.

> For fear is nothing but the failure
> to use the help that reason gives [...].
> (Wisdom 17:12)

1. Introduction

Given the constant fluctuations in the understanding of the basic concepts related to the main issue addressed in the paper and concepts related to it, it seems that a form of an essay is more appropriate for discussing them than a form of a standard research paper.[1] I hope this paper confirms the validity of this approach.

In the paper (Olszewski, Mączka 2020), published in 2017, i.e., seven years ago, we tried to answer the question of the limits of computer science (CS). We assumed then that we would be able to conceptually separate CS objects, i.e., the field of CS, from other objects. We argued that the limits that define the field of CS are determined by the properties of the human mind in the real world, which are expressed by famous Church's thesis. In the years that have elapsed since that publication, not only have the circumstances of CS and its limits changed, but so has the way in which they are approached. Let us first emphasise that today the more appropriate question is not the one about the limits of CS but the related question of the limits of artificial intelligence (AI) – we consider AI research to be a 'daughter' of CS. In this paper, we assume that AI in its current form is not its final stage and that its development is driven by humans, who are its creators and owners. Thus, we attempt, using the essayistic style, to predict the future development of AI. Here is our speculative picture of this development expressed in stages:

- The initial stage: algorithms (antiquity – the 12th century);
- The stage of the first calculating machines (the 13th–19th centuries);
- The stage of Turing's definition, the first computer, and cybernetics (the 20th century);
- The stage of the emergence and development of artificial intelligence (AI) (the 20th–21st centuries);
- The stage of the emergence of the artificial subject (Artificial Subject);
- The stage of the emergence of the artificial human (Artificial Human – AH);[2]
- The stage of the construction of the artificial 'god' (Artificial 'god' – AG).

Let us observe that the word 'god' is used in inverted commas and does not signify the person of real God, or any idol that someone may believe in, but a 'god' understood as an artificial object that possesses some of the 'qualities' attributed to God, such as omniscience, omnipotence, omnipresence, etc. Should anyone doubt such a scenario for AI development, please, note that certain current AI capabilities significantly exceed those of an average human, or perhaps any human.[3] Some even admit the possibility of AI escaping human control (cf. the motto of this paper). There is now talk of a theoretical construct called Artificial General Intelligence (AGI), which differs from AI in that it is supposed to represent (or simulate within software) any tasks that a human can perform. That is, using Searle's classic distinction, current AI is weak artificial intelligence, while AGI would be a strong version of artificial intelligence. AI focuses on specific and recognised goals that are realised by human intelligence, and AGI is supposed

to do so in general.[4] AI is an umbrella term that covers various realisations of certain specific human intelligent activities, while AGI can be treated as a term that denotes the one "representation of generalised human cognitive abilities in software".[5] Referring to the aforementioned stages in AI development, AGI should be considered equivalent to the stage of emergence of the artificial subject (AS), i.e., agent.[6]

Following this line of reasoning, to guide the Readers' intuition and to illustrate the concept of Artificial Human (AH), we will say *per analogiam* that AI relates to human intelligence as much as AH relates to humans as a whole. As we observed when comparing AI with AGI, the term artificial intelligence collectively labels certain technologies, whereas in the case of AH, technologies are embedded in a single subject (an agent). The naturalistic view of humans, which often implicitly guides AI researchers, overlooks the spiritual side of *homo sapiens*, as this side, by definition, escapes empirical research based on observation, experimentation, and empirical analyses. If naturalism were correct, then all that is considered human spiritual abilities, such as faith, prayer, love of God and neighbour, adoration, etc., should emergently appear in AH, as by-products of the workings of the nervous system. The belief that some human skills cannot be reduced to neural mechanisms, i.e., to broadly understood intelligence, is a critical presupposition of AI (AH) research. A critical reflection on this issue is the main focus of this paper.

2. The limits of AI?

Considering the predicted stages of AI development, let us observe that the question of the limits of CS has so far been impersonal and concerned the field or methodology of CS as a scientific discipline. However, with the advent of AI, our question of the limits shifts from the impersonal realm of CS to the subjective or personal realm of AI. Now we ask: what else can AI do? This question, when applied to CS (i.e., what else can CS do?), does not sound convincing. It could be argued that such a question about AI merely anthropomorphises[7] a machine, which is similar to how personal computers are anthropomorphised in everyday life. This is indeed the case, but now this anthropomorphisation takes on a different and more serious meaning with AI development, since, by definition, we want AI to be as human-like as possible, including in aspects where humans are subjects.[8] Recent years have brought an interesting clarification in the understanding of CS, which will probably also allow its philosophy to develop rapidly. What

seems to have become evident is that the implicit **goal** of CS development of is the creation of an artificial human (AH) or an artificial 'god' (AG). In other words, both computer science as it has been understood so far and the computer as a machine, along with their surroundings, are directed towards the main goal:

[The main goal of CS development] To create an artificial 'god'[9].

Given the essayistic nature of this paper and the assumed validity of the belief regarding the goal, we can reformulate the question of the limits of CS into a more modest version:

[The main question] Are there any limits to AI's ability to imitate humans?

In other words, the question is whether there are any limits to building an AH. Today, we can undoubtedly list activities where AI surpasses humans, e.g., in computing, which means that artificial humans are already superior in some activities to an average specimen of *Homo sapiens*.

3. Weizsäcker's argument

Carl Friedrich von Weizsäcker (1912–2007), a German physicist and philosopher, addressed the problem of the limits of cybernetics in his article *The Philosophical Problem of Cybernetics*, where he analysed whether the range of application of cybernetic models could be determined (Weizsäcker 1974: 282). He was convinced that attempts to limit this applicability would fail. To this end, he formulated an interesting argument, which I quote *in extenso* first as presented by its author:[10]

> How could a statement that this or that action cannot in principle be simulated cybernetically be formulated? If someone came to me and said: "I know such an action that cannot be cybernetically simulated", I would ask him: "Please define this action for me because I do not understand very clearly what you mean. Tell me clearly what this action consists of" – for example, the creation of concepts. If he then actually tells me clearly what this action consists of, I will take his words as instructions for building a machine that simulates it. If, however, he describes this action so inaccurately that I cannot perform it, I will reply: "You have not told me at all what the action that cannot be simulated consists of" [...] (Weizsäcker 1974: 284).

Since the aim of this paper is to consider the limits of CS or rather AI (AH), we will replace cybernetics with AI (or AH) in the original ar-

gument. The question arises whether this substitution is acceptable because, simply put, cybernetic modelling concerns the life processes (actions) of the human organism, while AI is used to model human thinking (intelligence). In the form of an answer, let us note that modelling in the case of the postulated artificial human includes cybernetic modelling as a special case, which further supports adopting AH. However, modelling within AI in fact only concerns human intelligence or, more broadly, human thinking. After making this substitution, we obtain an argument that is also interesting and applicable to AI (AH), which is why we will consider it. Following this path, let us consider only the dialogue between fictional persons A and B from the above quotation:

A: Are you saying that AI (or AH) will not be able to simulate a certain human action?
B: Yes, I am.
A: So, tell me what it could be – or maybe what it already is.
Interlocutor B gives a description of 'this thing'.
A: Ok. Your description is a recipe on the basis of which we will equip AI (AH) with this thing.
(*If interlocutor B does not give a sufficiently accurate description, then A says:*
A: You haven't told me at all what this action that cannot be simulated consists in.)

After standardising this dialogue, we indirectly obtain the following argument Arg1:

I. There is a human action A that cannot be simulated by AI (AH).[11] (an indirect assumption)
II. Every human action Ai can be described in an intersubjectively intelligible language in the form of Si. (a key premise) ($p \Rightarrow q$)
- Action A can be described in an intersubjectively intelligible language in the form of S.

III. Any description Si of an action in an intersubjectively intelligible language can be transformed into a description of an action for AI (AH). ($q \Rightarrow r$)
- S can be transformed into a description of an action for AI (AH).

IV. AI (AH) can simulate A. (a contradiction)

From this, we can draw a general and strong conclusion that:

[Conclusion] AI (AH) can simulate any human action.[12]

This argument seems cognitively attractive, and the premises sufficiently support the conclusion. The above Arg1 is also attractive from the point of view of AI and its future in the form of AH and AG.

However, this is not the only possible standardisation, as we can consider premise II to be false because it is too strong and combine it with premise III to form II', as in Arg2 below.

I'. There exists human action A that cannot be simulated by AI (AH). (an indirect assumption)

II'. Every human action Ai which is describable in an intersubjectively intelligible language in the form of Si can be simulated by AI (AH). $(p \wedge q \Rightarrow r)$

- Action A can be described in an intersubjectively intelligible language in the form S, and this description can be transformed into a description of action for AI.

III'. AI (AH) can simulate A or A is indescribable in an intersubjectively intelligible language.

From promises in Arg2, we cannot derive as strong a conclusion as in the case of Arg1. However, Arg1 seems to be consistent with Weizsäcker's intention. At this stage, we cannot definitively determine if Arg1 is formally or materially sound (although it does appear to be formally sound). Regarding material soundness, let us briefly mention here that a possible way of challenging the argument is to question premise II, specifically the assumption of the existence of an intersubjectively intelligible language in which the description of the action is to be formulated, which could lead to circular reasoning. This point needs further exploration, but we will now attack the argument in a different way by showing that the Arg1's conclusion is not sufficiently plausible or is even false. If we succeed, we will know that Arg1 is either formally sound and materially unsound or formally unsound although it may be materially sound. Both cases are unfavourable for Arg1, allowing us to reject it.

In our further considerations, we will assume a principle $\mathrm{AI}_{\mathrm{CT}}$, related to Church's thesis, which (Russell, Norvig 2022: 27) call the Church-Turing thesis and formulate as follows: "The Church-Turing thesis proposes to identify the general notion of computability with functions computed by a Turing machine".[13] Church's thesis is typically understood as a real limitation of a human subject, including a limitation of human computing skills, which means it is a hypothesis with an empirical component. Adopting McCarthy's thesis (cf. footnote 11) that human intelligence is computational in nature, Church's thesis imposes a limitation on human intelligence. Thus, if AI is a simulation of human intelligence, then AI should inherit this constraint, which we express below in terms of $\mathrm{AI}_{\mathrm{CT}}$. Famous metalogical theorems, such as the theorem of unsolvability of the halting problem for Turing machines and the theorem of incompleteness of first-order Peano arithmetic, change

their meaning depending on whether we accept the Church-Turing thesis or whether we only accept the AI_{CT} thesis formulated below. In a world where the Church-Turing thesis holds, both metalogical theorems apply to humans, whereas when we adopt only the AI_{CT} thesis, they apply only to AI (AH).[14] This is our assumption:

[AI_{CT}] AI acts as a universal Turing machine (UTM).

The second issue relevant to our discussion in this section is the question of how to understand both human action and its description. Since there is no clear consensus on what constitutes human action or how best to describe it, we cannot expect precise definitions here. For the sake of clarification, when discussing human action, we will temporarily adopt Alonzo Church's (Church 1941: 2–3) distinction between two ways of specifying a function: in extension and in intension. In simple terms, a description in extension means that we know the domain of the function and, for any given x in that domain, we know the value of fx in the codomain, along with some general characteristics of the function (e.g., that f is computable). A description in intension, on the other hand, provides an explicit procedure for calculating the value of the function for any given argument. By analogy, we will consider two ways of describing an action: an *in extension* interpretation and an *in intension* interpretation. Weizsäcker's argument most likely refers to the description of an action *in intension* – that is, as a procedure. However, it is important to note that if an action is described *in intension*, it precisely defines an action *in extension* as well. Conversely, if we only have an action given *in extension* and know that it is feasible, we can infer that an intension-based description of that action can exist[15].

We will now attempt to demonstrate the Arg1 could potentially be unreliable which means that its conclusion could be false in some model relevant to our considerations by showing that a human action that cannot be simulated by AI (constrained by the AI_{CT} assumption) is possible. In other words, there can exist a human action that AI cannot simulate. This intention is not entirely new and has been described by (Russell, Norvig 2022: 1032–1035), who list three main types of arguments against the hypothesis that limits AI's ability to simulate human action. These are: a) ***Turing's argument from informality***, which posits that a human cannot encapsulated within the form of a program or formal rules;[16] b) ***the argument from disability***, which asserts that AI (AH) "cannot do something" and specifies what that might be; and c) ***the mathematical argument***, which uses Gödel's incompleteness theorem. In this section,

we will follow the path of the argument from disability and the mathematical argument and use two important technical outcomes in our argumentation: the unsolvability of the halting problem and Gödel's first incompleteness theorem.[17] Our argument is connected to both the mathematical argument and the argument from disability. It is important to emphasize that we will not attempt to explicitly identify – i.e., *in intension* – a specific human action that AI cannot simulate. Rather, we aim to demonstrate that it is possible such an action exists, or equivalently, that it has not been proven that such an action does not exist. Let us remind the Readers that this perspective is grounded in the rejection of the Church-Turing thesis as true. We are operating within the realm of possibility, meaning that it is conceivable that the class of functions computable by a Turing machine is a proper subset of the broader class of functions computable in the intuitive sense.

Let us start with the unsolvability theorem of the halting problem for Turing machines. In this argument, we do not accept the veracity of Church's thesis. Here is an imprecise form of this theorem, which will be useful in further discussion:

[The halting problem]	There is no Turing machine that solves the halting problem for Turing machines.

Assuming the possibility of the existence of a function that is computable in the intuitive sense but uncomputable by a Turing machine and $\mathrm{AI}_{\mathrm{CT}}$ we obtain:

[The version of the halting problem]	It is possible that there exists a function efficiently computable in the intuitive sense (action C feasible by a human) that solves the halting problem for Turing machines.

Such a hypothetical possible human action C that leads to solving the halting problem for Turing machines cannot be fully simulated by AI, as is evident from the above discussion and especially from $\mathrm{AI}_{\mathrm{CT}}$.

Informally speaking, a typical mathematical proof of the unsolvability of the halting problem indirectly considers a certain effectively computable function, which involves checking whether a given Turing machine with a specific input will halt. The description *in extension* of this function, the assumption of its existence and computability lead to the existence of a so-called diagonal function for the set K. To lead to a contradiction it has

to be assumed, based on Church-Turing thesis, that this diagonal function is computable by a Turing machine. Since we have assumed that the set of computable functions in the Turing sense is a proper subset of the set of all computable functions in the intuitive sense, it is impossible to lead to a contradiction in the proof. By virtue of the analogy we have adopted between actions and functions, we will say that it is possible that such action C (described *in extension*) exists that can be performed by a human because there is no logical reason why a human could not perform it, apart from constraints caused by lack of time, space, memory, etc.[18] Thus, we can say that there can exist human action C that generally (i.e., beyond the usual constraints) resolves the halting problem for Turing machines. As we mentioned above, if Church-Turing thesis were true, a human could not do this for logical reasons, since a human action would yield a certain effectively computable function that would be computable by a Turing machine by virtue of Church-Turing thesis. Under the $\mathrm{AI}_{\mathrm{CT}}$ assumption, AI cannot perform such an action for logical reasons, as the assumption of the feasibility of such an action by a UTM leads to a contradiction. This occurs precisely when the UTM examines its own Gödel number. This informal and simplified reasoning should justify the view that there can exist an action that can in principle be performed by a human and cannot be performed by AI for logical reasons. It should be emphasised again here that this hypothetical action C is given *in extension*.

We will derive a second counter-example for the Arg1's conclusion from Gödel's first incompleteness theorem. Let PA denote the formal system of first-order Peano arithmetic, Th(N) – the set of all true sentences in the standard model for PA, $\mathrm{T}_{\mathrm{Th(N)}}$ – the set of their Gödel numbers, and T_{PA} – the set of Gödel numbers of all PA theorems. It is known that set T_{PA} is recursively enumerable and not recursive. The following conclusion from the first incompleteness theorem holds:

[A version of Gödel's first theorem]	Theory Th(N) of true sentences in the standard model is not recursively axiomatisable.[19]

In Gödel's theorem, the ω-consistency of PA is assumed explicitly, while another crucial assumption, the recursivity of the axiomatic system of Peano arithmetic, is in a sense hidden and often overlooked. This assumption means that the T_{PA} set of Gödel numbers of all PA theorems is recursively enumerable, which implies that the set of Gödel numbers of PA axioms is recursive as are the rules of the system.[20] The set of all (Gödel numbers) sentences of the language of first-order arithmetic which are true in the standard model,

is not recursively enumerable. It is an empirical fact that a human can perform action D which consists in axiomatising the set Th(N). Such axiomatisation, based on the omega-rule, is known, but it is not recursive and is of second-order.[21] However, following Gödel, that a non-contradictory and recursive formal system can be equated with a Turing machine, the action of proving within such a formal system boils down to the construction of an appropriate Turing machine, which AI can do.[22] However, it seems that, at least at present, there is no Turing machine that can simulate the omega-rule, as it is not a finitistic rule. Thus, there appears to exist an action D that a human can perform but AI cannot.

The Readers will no doubt notice that both arguments are not fully conclusive, since they do not give a description of the relevant actions in extension, and they move in the realm of possibility, but their purpose is to raise reasonable doubt about AI's ability to simulate human actions. It is noteworthy that both the unsolvability theorem of the halting problem and Gödel's first incompleteness theorem are called *limitative theorems*. The researchers who have given these theorems this name have been right although they seemed to apply limits for humans (*homo sapiens*) while in fact they are rather limits for the possibilities of AI (AH), unless the Church-Turing thesis is adopted.

Summarising this section, we can say that Weizsäcker's argument, despite the aforementioned reservations, allows us to be optimistic about the future of AI, particularly regarding the absence of limits on simulating human actions and building AH. Philosophically and logically the following types of necessity and thus also impossibility are distinguished:

- logical,
- metaphysical,
- physical,
- epistemic,
- alethic,
- moral.

Logical impossibility is the strongest, as it implies the other impossibilities, while logical possibility opens the door to other possibilities.

Weizsäcker's argument is also linked, albeit loosely, to a significant ontological issue, which we are only signalling here and which requires separate study: behaviourism. This issue concerns the functioning (action) of a certain object and the question whether it exhausts the essence or identity of that object. In other words, behaviourism is an ontological thesis according to which a realistically existing human is equated with the totality of their actions, and it seems that an optimistic conception of AH might possibly

be based on such a philosophy. Although behaviorism was thought to be discredited long ago, it has resurfaced in AI research, which studies how human intelligence works rather than what it is.

4. Does AI necessarily have to be solipsistic?

We will now address a philosophical view called solipsism, which we believe has important implications for the philosophy of AI, and we will use it to formulate a third counter-example to the conclusion from Weizsäcker's argument.[23] To make it easier for Readers to follow our argumentation, we will first formulate the main hypothesis H of this section, which establishes the relationships between solipsism and AI.

[H] If AI (AH) develops to the level of creating its own philosophy, it will always be a version of *solipsism*.

Arguing for this risky hypothesis may seem challenging due to limited sources and vague concepts, but we will undertake this attempt. We will first develop the understanding of solipsism as a philosophical view. Analyses of numerous definitions of solipsism given in various studies reveal its four main forms: metaphysical solipsism, epistemological solipsism, methodological solipsism, and ethical solipsism. We will focus on metaphysical solipsism and treat its other forms as derivatives. Almost every definition of solipsism formulated in the literature contains the terms *Existence*, *Ego (I)*, and *Datum*, which in the following conceptual quasi-equations we will denote by the symbols E, I, and D respectively. The metaphysical version can be summarised in the thesis about what exists: what exists exists exclusively because nothing exists outside of it. This can be expressed as:

[SM1] $E = I + D.$

Thus *Existence* encompasses *Ego (I)* and that which is given for *Ego*, i.e., *Datum*. The second quasi-equation expresses that *Ego* is distinct from *Datum*:

[SM2] $I \neq D.$

An intuitive digression: To give the Readers an intuitive idea of what solipsism is, let us consider our usual cognitive attitude towards the world and ourselves that we tend to have in our heads, including our senses, and then let us reject the external world. What exists is only what is left in our heads after its rejection expresses what exists in the solipsistic sense.

We will now present a sketch of the reasoning (which obviously requires further elaboration) that supports our hypothesis H. Based on testimonies and reflections of various renowned philosophers, let us first adopt the following empirical premise:

[Premise1.] Based on the testimony of some renowned philosophers (e.g., Russell, Popper, Watson, and Santayana), solipsism, in its poor version, is an irrefutable philosophical view. The irrefutable version of solipsism, called *solipsism of the present moment*, has been studied in detail by Santayana, among others. Premise1 is partly empirical and grounded in the non-deductive scheme of an *ex auctoritate* argument.

[Premise2.] [AI_{CT}] AI operates and will continue to operate as a universal Turing machine (UMT). This premise is essentially a version of Church's thesis, and although it remains unproven, it forms the foundation of contemporary AI (AH) and will likely remain so until a new model of computability, distinct from the Turing model, is developed.

[Premise3.] The only source of information for AI (AH) is machine code. The entire external world (objects) and all changes must be represented in the machine code of a computer processor. The machine code in a machine is analogous to Datum in solipsism. This premise aligns with the solipsistic requirement of 'exclusive existence'.

[Premise4.] Nothing else but machine code exists (for the UMT).

Premise3. and Premise4. Together yield: $E = I + D$, or: $E = UMT + a\ tape$, which can be expressed by referring to a diagram that represents the structure of Turing Subject (TS), derived from Hilbert's Subject diagram:[24]

(TS1) The Universal Turing Machine (UTM) thinks, i.e., computes. (**I**)

(TS2) It computes or executes a program. (**I**)

(TS3) On a tape, UTM can: (**I**), (**D**)
- read and recognise characters;
- erase and write characters;
- move the head right or left;
- change the internal state;
- halt computing.

(TS4) UTM has the ability for partial self-reflection (recursion theorem). (**I**)

Here is the formulation of Turing Subject expressed by the quasi-equation SM1, which embodies solipsism:

(T1) E = I + D.
(T2) I = Universal Turing Machine (UTM).
(T3) D = Datum (Tape – potentially infinite).

Historians of philosophy believe that, beyond the well-known origins of solipsism in Descartes' "Cogito, ergo sum" and Berkeley's philosophical system and its principle "esse = percipi", solipsism also derives from strong British empiricism as expressed by Locke and his French follower E. Condillac. In his "Treatise on the sensations", Condillac considered a hypothetical stone statue shaped as a human body, which he abstractly endowed with successive senses: touch, taste, smell, hearing, and sight, and analysed the changes in the statue resulting from the acquisition of sensory data.[25] In doing so, he attempted to reconstruct, along the lines of Locke's empiricism, the possibility of the statue's obtaining knowledge not linked to mental states. His analyses led him to conclude that the statue would have exclusively experiential states and no experiences indicating the existence of external objects. Condillac's view of the relationship between sensory data and existing things is called sensualism, and his conclusion regarding the statue was that it would remain solipsistic.[26] AI researchers can draw on Condillac's thought experiment to guide their own investigations, particularly those regarding AH and AG.[27]

To conclude this section, it can be said that adopting the indicated model of Turing Subject as adequate for AI allows us to maintain the hypothesis H. If this hypothesis were proven, it could imply that the pursuit of truth is an unattainable goal for AI, a pursuit that seems a fundamental aspect of the human species desire and action. The issue concerns truth and truthfulness in their classical understanding as the correspondence between cognition and Reality. For AI, there is no Reality, only Datum in the form of machine code, which is likely to remain unchanged for some time to come. To put this in a slightly different perspective, we will quote an accurate diagnosis of solipsism in the broader context of the perception of Reality in the form of an excerpt from the book entitled *The Problem of Perception*:

> For if Direct Realism is false, only two positions remain between which we must choose: Idealism and Indirect Realism. Idealism is almost totally out of favour today, so it is Indirect Realism that is most commonly thought to be the real rival to Direct Realism. Now, far and away the most common criticism that has been levelled against Indirect Realism is that it brings epistemological disaster in its wake. The thought is that, if we are constitutionally incapable of directly perceiving the physical world, but are cognitively confined within the circle of our own perceptions, then there is just no good reason to suppose that there is a physical world at all outside this domain, behind this veil. (Smith 2002: 12–13).

5. Conclusions

In this paper, we have examined C.F. Weizsäcker's interesting argument about the limits of AI. His conclusion regarding the limits of AI, and indirectly the limits of our thinking – i.e., that we cannot invent or describe human action that cannot be simulated artificially – seems inescapable. We have presented three arguments against this conclusion of which the third, the argument about the inevitability of solipsism for AI, appears philosophically promising for further investigation. Although our critique is not fully conclusive, partly because we have not assumed Church's thesis, it reveals that the question of the limits of AI is one of the most interesting and important theoretical problems in the field. This paper sketches numerous related issues that require deeper exploration. Weizsäcker's argument itself merits a separate, more detailed study, and specifically in the context of Church's thesis, as our consideration should show that there is a deeper connection between the two. Additionally, the issue of solipsism and AI seems promising for research on realism in general, and direct realism in particular. Finally, the search for actions from the human spiritual realm that cannot be simulated by AI needs to be resolved, which requires identifying strong assumptions that exclude such actions at the very least.

N O T E S

[1] Important features of an essay include its non-systemic nature (not exhausting the topic), unsystematic structure (loose associations), and subjectivity.

[2] Let us observe that every human is a subject but not every subject is a human.

[3] The article (Xu Y., Liu X., Cao X., et al. 2021) reads: "This paper carries out a comprehensive survey on the development and application of AI across a broad range of fundamental sciences, including in-formation science, mathematics, medical science, materials science, geoscience, life science, physics, and chemistry" (p. 723), which brilliantly demonstrates the range of applications of AI.

[4] Cf. (Lutkievich 2022: 1).

[5] Cf. (Lutkievich 2022: 1).

[6] The notions of a subject and an agent, although semantically close, are not identical. The former is commonly used within the philosophy of mathematics and the philosophy of cognition, while the latter is prevalent in cognitive science and AI.

[7] It is derived from two Greek words: ἄνθρωπος (anthrōpos) – human and μορφή – shape or form.

[8] Some argue that CS is intrinsically identical to AI. In our opinion, the important difference between the two is that AI is CS with a precise development goal, or in short AI = (CS + goal).

[9] Cf. earlier comments on this concept.

[10] The original source of this quotation is Weizsäcker's 1968 paper, published as an article entitled *The philosophical problem of cybernetics* in his book (Weizsäcker 1974: 280–291). He understood cybernetics as the art of controlling, which allowed life processes to be explained by comparing them with control systems (Weizsäcker 1974: 281).

[11] There is an evident correlation that whatever AI can do, the same can be done by AH.

[12] A similar hypothesis was adopted at the famous 'Dartmouth Summer Research Project on Artificial Intelligence' in 1955: "The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it" (McCarthy et al. 1955: 1). However, unlike in Weizsäcker's work, this was accepted without premises. Furthermore, when asked to define intelligence, McCarthy replied: "Intelligence is the computational part of the ability to achieve goals in the world" (McCarthy 2007: 2).

[13] This formulation refers to a version of the thesis given by Alonzo Church, who spoke of the identification of concepts, which naturally linked his thesis to AI research. Cf. (Church 1935).

[14] The considerations in this paragraph are crucial for the entire paper.

[15] Can we say that the following description of a human action is simulated by AI? "Imagine a machine built on the basis of a computer that uses a peripheral device to randomly toss a coin. Each successive toss (its number) and the result of each toss are recorded on an appropriate medium, with values represented as 0 and 1. If this process could be continued indefinitely under certain conditions, it would define a function. This function would, in a sense, be constructed. [...] Furthermore, for any natural number, which represents the number of the next toss, the value at that point can be calculated" (Olszewski 1999: 99). Whether this description can be converted into a mathematical model or whether AI can be equipped with the described procedure is uncertain. This indicates that the issue remains quite ambiguous.

[16] Works along these lines were published in the 1990s, and, from today's perspective, are outdated, as they critiqued an earlier form of AI, called Good Old-Fashioned AI (GOFAI) (Russell, Norvig 2022: 1033).

[17] The use of Gödel's theorem in this context has a long history that starts with Lucas's famous 1961 paper. The use of the halting problem in argumentation is rather rare. Cf. (Russell, Norvig 2022: 1032–1037).

[18] Note that the computer is subject to similar constraints as humans are, albeit on a different scale.

[19] If PA is consistent, then PA is incomplete.

[20] This equivalence holds by virtue of Craig's result; cf. (Cooper 2003: 119), (Odifreddi 1989: 356–358).

[21] When we refer to the omega-rule, we mean it in this form: for any formula $A(x)$ of the language of arithmetic, if $A(0), A(1), A(2), A(3), \ldots, A(n), \ldots$ for any numeral n denoting a natural number n have been proved in the system, then the formula $\forall x A(x)$ can be attached.

[22] (Feferman 1996: 137) quotes Gödel's statement that "a formal system is nothing but a many-valued [nondeterministic] Turing machine which permits a predetermined range of choices at each stage".

[23] In a brief section on consciousness and qualia, (Russell, Norvig 2022: 1036–1037) address issues related to solipsism. Qualia (singular: quale), are individual states of subjective conscious experience, e.g., the redness of a red apple.

[24] The relationship between Turing Subject and Hilbert's Subject was explored in (Brożek, Olszewski 2013).

[25] The thought experiment with the statue bears some resemblance to our hypothesis H, which was brought to my attention by Jerzy Mycka, for which I would like to express my gratitude.

[26] Cf. (Gaukroger 2020: 179–180).

[27] At this stage, the links between solipsism and philosophical reflection on AI have not been extensively explored. The classic book (Russell, Norvig 2022) mentions the term solipsism only once on page 1036, with reference to Turing's discussion of machines and thinking.

REFERENCES

Brożek, B., Olszewski, A., (2013): *Podmiot matematyczny Hilberta*, Zagadnienia Filozoficzne w Nauce. 53, 93–132.

Burns, E., Laskowski, N. and Tucci, L., (2022): *What is Artificial Intelligence* (AI)? *Definition, Benefits and Use Cases*. Available at: https://www.techtarget.com/searchenterpriseai/definition/AI-Artificial-Intelligence (accessed 4 January 2023).

Church, A., (1935): *An Unsolvable Problem of Elementary Number Theory*, *Preliminary Report*. (abstract) Bulletin of the American Mathematical Society, 41(6), 332–333.

Church, A., (1941): *The Calculi of Lambda-Conversion*. Princeton: Princeton University Press.

Condillac de, E. B., (1754): Traité des sensations. Par M. l'Abbé De Condillac, De l'Académie Roy ale de Berlin. Tome La Londres; & se vend A Paris, Chez De Bure l'aîné, Quay des Auguftins, à Saint Paul.

Cooper, S.B., (2003): *Computability Theory*. Chapman & Hall, Boca Raton. DOI 10.1007/s11225-007-9050-0

Feferman, S., (2006): *Are There Absolutely Unsolvable Problems? Gödel's Dichotomy*. Philosophia Mathematica, 14(2), 134–152. https://doi.org/10.1093/philmat/nkj003

Gaukroger, S., (2020): *The Failures of Philosophy: A Historical Essay*. Woodstock, Oxfordshire: Princeton University Press.

Kopera, G., (2021): *How Adaptive AI Outpaces Traditional AI Capabilities*. Available at: https://www.thoughtai.org/post/how-adaptive-ai-outpaces-traditional-ai-capabilities (accessed 4 January 2023).

Lutkievich, B., (2022): *Artificial general intelligence* (AGI). TechTArget. https://www.techtarget.com/searchenterpriseai/definition/artificial-general-intelligence-AGI

Mączka, J., Olszewski, A., (2020): *The Question of the Boundaries of Computer Science*. Seminare Poszukiwania naukowe, 41(4), 45–57

McCarthy, J., Minsky, M., Rochester, N., Shannon, C.E., (1955): *A proposal for the Dartmouth summer research project on artificial intelligence*, http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html (accessed May 2024).

McCarthy, J., (2007): *What is Artificial Intelligence?*, Draft, 1–15.

Müller, V.C. (2023): *Philosophy of AI: A Structured Overview*. Final Draft.

Odifreddi, P., (1989):*Classical Recursion Theory. volume 1: The Theory of Functions and Sets of Natural Numbers*. Amsterdam: North-Holland.

Olszewski, A., (1999): *Teza Churcha a platonizm*. Zagadnienia Filozoficzne w Nauce, 24, 96–100.

Oxford English Dictionary, s.v. "artificial intelligence (*n.*)", March 2024, https://doi.org/10.1093/OED/3194963277.

Popper, K.R., (1993): Knowledge and Body-Mind Problem, Routledge, London.

Russell, B., (1923). *Human Knowledge. Its Scope and Limits*. London: George Allen and Unwin.

Russell, S., & Norvig, P., (2022): *Artificial Intelligence. A Modern Approach*. 4th ed. New York: Pearson.

Santayana, G., (1923): Scepticism and animal faith: introduction to a system of philosophy. New York: C. Scribner's sons.

Smith, A.D., (2002). The Problem of Perception. Cambridge, Mass.: Harvard University Press.

Stacewicz, P., (2019): *From Computer Science to the Informational Worldview. Philosophical Interpretations of Some Computer Science Concepts*. Foundations of Computing and Decision Sciences, 44, 27–43.

Sturt, H., (1911): "Condillac, Étienne Bonnot de". [in:] Chisholm, Hugh (ed.). *Encyclopœdia Britannica*. Vol. 6 (11th ed.). Cambridge University Press. pp. 849–851.

Watson, R., (2016): *Solipsism: The Ultimate Empirical Theory of Human Existence*, St. Augustines Press.

Weizsäcker, C.F., (1974): *Die Einheit der Natur*. DTV, München.

Xu Y., Liu X., Cao X., et al.. (2021): *Artificial intelligence: A powerful paradigm for scientific research*. The Innovation, 2(4), 708–727.