AH ANALECTA HERMENEUTICA
INTERNATIONAL INSTITUTE FOR HERMENEUTICS / INSTITUT INTERNATIONAL D'HERMÉNEUTIQUE

ISSN 1918-7351
Volume 15.1 (2023)

What Is AI, and If So, How Many? Four Puzzles about Artificial Intelligence

Sebastian Rosengrün
CODE University of Applied Sciences, Berlin
ORCID: 0000-0002-0747-8424

Abstract
This paper demonstrates why the following philosophical questions are misleading: can an Artificial Intelligence (AI) think, feel or act, and does it, therefore, have moral rights and duties? It does so by elucidating the issue with four puzzles. The first puzzle concerns the extension of the concept of AI, which, from the standpoint of semantics, necessarily is either empty or underdetermined. The second puzzle makes a distinction between robots and AI. It points out that it is a grave technical misunderstanding to understand a robot as an entity of its own which can be attributed mental states or the status of a moral object. Based on this, in the context of the third and fourth puzzle, this paper states the paradox of the Computer of Theseus, which compares to a new version of the well-known paradox of the Ship of Theseus and demonstrates that, in the face of the peculiarities of hardware and software, AI, considered metaphysically, is a very strange concept.

Keywords: philosophical paradoxes, artificial intelligence, moral philosophy, consciousness, machine learning

Introduction

A significant part of the philosophical debate on AI is to ask whether an AI can think, feel, or act and, therefore, whether it may have moral rights and duties.[1] However, these questions are misleading. Indeed, they aim at what can be attributed to AIs, whether AIs possess mental states (consciousness, intentions, emotions, etc.) or are bearers of moral rights. However, both the historical debate since the 1950s and the current debate on AI mostly fail to determine exactly to whom or what something is attributed at all when talking about AI.

The metaphysical question of who or what an AI is, which entities can even be called AIs, is considerably more complex than one might assume. By metaphysics or ontology, this paper refers to the philosophical sub-discipline, which asks about the existence, being, essence, and structure of things. In analytic philosophy, in particular, metaphysics is closely related to semantics, the linguistic sub-discipline, which asks about the meaning and reference of linguistic expressions. Semantic questions are also the starting point of the following reflections on the metaphysics of AI.

The sentence (1) "This AI has mental states" is identical in form to the sentence (2) "The present king of France has a bald head." Both express an attribute about a certain individual, namely having mental states and being bald, respectively.

To determine the truth value of (2), it is not irrelevant to define what it means to be bald, to consider where baldness comes from, and to discuss moral rights and duties bald people have. In this example, however, assigning a truth value fails not because of an underdetermined definition of the attribute, but because of the indeterminacy of the individual about whom the attribute is expressed. Although the nominal phrase "the present king of France," semantically, refers to the individual who is presently king of France, it is an empty reference because France presently is a republic. That is, the individual who is said to be bald does not exist.

Applied to AI: It is philosophically puzzling to whom or what mental states are attributed in a sentence like (1). On the one hand, this is because—unlike in the case of the King of France—there are different meanings of the term "AI," and on the other hand—just like in the case of the King of France—it is unclear whether a nominal phrase like "this AI" refers to anything at all, and if so, to what exactly.

In this context, this paper discusses four puzzles of a philosophy of AI, some of which build upon each other, and which illustrate the problematic nature of the concept of AI from a semantic and metaphysical perspective.

---

[1] Sebastian Rosengrün, *Künstliche Intelligenz zur Einführung*, Zur Einführung (Hamburg: Junius, 2021).

Every Computer Is AI (Or None)

AI research is divided into two divergent branches:[2] on the one hand, AI is an interdisciplinary research field in which human or natural intelligence is modeled, simulated, and replicated, mostly with the goal of better understanding human or natural intelligence and other cognitive abilities. This field is commonly referred to as "cognitive simulation."[3] On the other hand, AI is a set of specific techniques within software engineering (and thus, AI is a sub-field of computer science). Those techniques are used in the field of cognitive simulation, too, although cognitive simulation goes far beyond computer-based methods and includes, among other things, attempts to replicate intelligence using biochemical methods (this area is widely known as 'wet AI').[4]

While advances in the field of cognitive simulation have yielded insights into intelligence, cognition, and consciousness, it is merely speculative at this stage whether artificial intelligences can be created that may have consciousness and other mental states. The main reason for this is that simulating intelligence is not the same as intelligence—much like a flight in a flight simulator is not a real flight. Moreover, it is doubtful what exactly distinguishes an artificial intelligence (if it is more than a simulation) from a natural intelligence, or whether the distinction between naturalness and artificiality can be maintained at all. If AI is understood in terms of cognitive simulation, there are currently no entities that can be called AI.

In the following, I focus on AI as a subfield of computer science, as a collective term for those techniques that currently play an important role, for example, in the engineering of chatbots, robots, autonomous driving systems, military drones, algorithm-based decision systems, and many other applications. AI encompasses the following subfields of software engineering: Machine learning based on neural networks; Computational linguistics or natural language processing; Machine vision; Reason-based reasoning; Planning and optimization. Combinations of those fields are not only possible but also common.[5]

Furthermore, it is discussed whether simple rule-based programs also count as AI. A relevant example would be a sequence of if-then statements, which—like any computer program—is realized as an electronic circuit system. However, all other

---

[2] Keith Frankish and William Ramsey, *The Cambridge Handbook of Artificial Intelligence*, 3rd ed. (Cambridge: Cambridge University Press, 2018); Klaus Mainzer, *Künstliche Intelligenz. Wann Übernehmen Die Maschinen?*, 2nd ed. (Berlin: Springer, 2019); Nils J. Nilsson, *The Quest for Artificial Intelligence: A History of Ideas and Achievements* (Cambridge: Cambridge University Press, 2009); Stuart Russell and Peter Norvig, *Artificial Intelligence. A Modern Approach*, 3rd ed. (Harlow: Pearson, 2016); Joseph Weizenbaum, *Computer Power and Human Reason* (New York and San Francisco: Freeman, 1976); Rosengrün, *Künstliche Intelligenz zur Einführung*.

[3] Daniel Dennett, "The Singularity—an Urban Legend?," 2015, https://www.edge.org/response-detail/26035.

[4] Wendell Wallach and Colin Allen, *Moral Machines: Teaching Robots Right from Wrong* (Oxford: New York: Oxford University Press, 2009), 55-56.

[5] Rosengrün, *Künstliche Intelligenz zur Einführung*, 13-33.

techniques mentioned are by their nature nothing else than highly complex rule-based systems; they can be completely reduced to them. This leads into the following paradox:

1. Rule-based systems are either AI or they are not.
2. All techniques that are commonly considered AI are completely reducible to rule-based systems.
3. Every computer program is a rule-based system.
4. If every rule-based system is an AI, then every computer program is an AI.
5. If rule-based systems are not AI, then no computer program is an AI.
   Therefore,
6. Either every computer program is an AI, or no computer program is an AI.

From the point of view of computer science, this paradox is not problematic. There, AI is primarily a loose collective term for software engineering techniques. For the successful execution of a program, it is irrelevant whether, for example, machine learning based on neural networks is metaphysically different from a simple "Hello World" command or whether it differs from it only because of a greater complexity of the source code.

This paradox becomes relevant only when entities are referred to as AI and/or certain attributes are ascribed to entities because they are "artificially intelligent" or an application of AI technology, suggesting both philosophical and social consequences. Ascribing mental states to a particular computer (or robot, software, etc.) because there is AI involved is therefore either an empty or misleading statement. According to the paradox explained above, this computer would either not exist at all or every other computer (for example, also the one I am writing this paper on, but also my smartphone and a Commodore 64 gathering dust in the attic) would possess mental states. Therefore, AI cannot be the reason that a computer possesses mental states.

The thesis that every computer possesses mental states may sound absurd at first glance. However, this does not mean that it is irrelevant. Hilary Putnam[6] has coined the position of functionalism or computer functionalism for this in the philosophy of mind. He argues that any electronic device on which a Turing-complete system can be realized (simplistically, any universally programmable computing machine) operates on the same principle as the human mind. However, the actual criterion for attributing mental states is then not AI, but Turing-completeness. AI would be only an unfortunate term for programming computers of any kind. From a philosophical perspective, the AI term would then be at least misleading, because its connotations, shaped by science fiction literature, invite to draw hasty false conclusions and to form magical associations.

---

[6] Hilary Putnam, "Minds and Machines," in *Dimensions of Minds*, ed. Sidney Hook (New York: New York University Press, 1960), 138–64. It is, however, well-known that Putnam changed his views over time, see Rosengrün, *Künstliche Intelligenz zur Einführung,* 35-64.

The first answer to the paradox, according to which no computer is an AI, on the other hand, makes any statement *de re* about an AI a statement with an empty name and leads into the classic no-reference problem of philosophy of language.[7] To claim that a particular AI possesses mental states is then comparable to claiming that the current king of France is bald, which, depending on premises of philosophy of language, is either a false or a meaningless statement as long as France does not return to monarchy in a possible distant future. Alternatively, AIs can be understood as fictional entities (comparable to unicorns, for instance), which even seems obvious, especially given the popularity of the topos in science fiction literature. However, this leads to the fact that a statement about AIs says nothing about real entities. A statement about AIs would then be comparable to the statement "unicorns have pink manes," which beyond a fantasy story hardly presupposes the existence of real unicorns. Saul Kripke, for example, argues that natural kind terms for fictional entities like unicorns fall under the so-called pretense principle, i.e., those terms are used *as if* the entities really exist, while everyone is aware that their existence is just pretended.[8]

Moreover, the statement "AI possesses mental states" is also analyzable *de dicto*, as a statement about what is expressed by the term "AI," comparable to "The present king of France is the one who is monarch of the country designated as 'France' at the time of the utterance." However, even according to this reading, no entity would be said to have mental states, but merely expressed that an AI (whether it exists or not) is something that has mental states.

At least this would apply to all entities of the present and near future. It is true that it cannot be proven in principle that no technique of software engineering is conceivable that is not by its nature completely reducible to rule-based systems and would be classified as AI by the current scientific discourse. To claim otherwise, however, would be pure speculation, which, moreover, is likely to be based less on technical progress than on a quite possible change in the use of language: of course, "artificial intelligence" in the distant future (or in a counterfactual situation, i.e., a possible world) may denote something that is not completely reducible to rule-based systems. However, such a counterfactual use of terms is irrelevant to the validity of the thesis that AI is nothing other than a rule-based system.[9]

This first puzzle has shown that the term "AI" is indetermined, at least when it is used to refer to specific entities: Either every computer (or computer program) is an AI, or there is no AI. Instead of AI, therefore, it should in principle be more precise to speak of certain techniques of software engineering. Beyond this puzzle, my concern in what follows is to point out further metaphysical issues and problems that

---

[7] Bertrand Russell, "On Denoting," *Mind; a Quarterly Review of Psychology and Philosophy* 14, no. 56 (1905): 479–93; Saul A. Kripke, *Reference and Existence. The John Locke Lectures* (Oxford: Oxford University Press, 2013); John Perry, *Reference and Reflexivity* (Stanford: CSLI, 2001); Sebastian Krebs, *Kripkes Metaphysik Möglicher Welten* (Berlin: De Gruyter, 2019).

[8] Kripke, *Reference and Existence*; Krebs, *Kripkes Metaphysik Möglicher Welten*.

[9] Saul A. Kripke, *Naming and Necessity* (Cambridge, MA: Harvard University Press, 1980): 116-125.

result from a misunderstood notion of AI, which is often used in current discourse as if it denotes entities about which certain attributes can be stated. The puzzle presented in the following section is mereological in nature and concerns the frequently advanced proposition that robots possess mental states and/or moral rights because of AI.

A Robot Is Not an AI, an AI Is Not a Robot

According to the prevailing understanding, a robot is an electromechanical machine, consisting of a processor, sensors and effectors. Other possible criteria discussed to define a robot include independent physicality, autonomous or seemingly autonomous behavior, and the ability to influence its respective environment.[10] Of course, industrial robots (e.g., in automobile production) as well as household and everyday robots (e.g., vacuum cleaners and lawn mowers) are also considered robots. These are to be distinguished from android or humanoid ("human-like") robots, which are mostly associated with artificial intelligence in science fiction. Purely mechanical robots or automata, while historically significant, play little role in contemporary robotics.

A characteristic of electromechanical robots is that they are usually controlled by a computer (which is a Turing-complete system). Depending on the paradox described above, any current robot could indeed be classified as artificial intelligence. However, from a technical perspective, the term AI is mostly understood in a narrower sense: For robots specifically, in addition to machine learning, natural language processing and machine vision are the most relevant AI applications. Although they are by their nature nothing more than rule-based programming (see above), these areas certainly describe independent fields of software engineering or computer science.

Accordingly, a robot could be defined to be artificially intelligent if it is controlled by a computer running AI applications, for example, software that analyzes obstacles in a room based on sensors (or cameras) and controls the robot's movements accordingly. While this description of a robot is unproblematic from an engineering perspective, some metaphysical issues arise from the technical setup as soon as artificial intelligence is used as a criterion for attributing mental states or even moral rights and duties to robots. After all, even if computers should possess mental states (and thus possibly the ability to suffer and moral rights) due to certain AI software,[11] this cannot be easily transferred to the robot that is controlled by this computer. Unlike humans, the "mind" or "brain" of a robot exists independently of its body. In this context, it is interesting to point to Hubert Dreyfus' famous criticism of "strong AI"

[10] Janina Loh, *Roboterethik. Eine Einführung* (Berlin: Suhrkamp, 2019); Catrin Misselhorn, *Grundfragen der Maschinenethik*, 4th ed. (Ditzingen: Reclam, 2019).

[11] I doubt this but the following argument is relevant nevertheless since it builds upon a common technical misunderstanding about the setup of robots which leads to further philosophical trouble.

according to which any human-like intelligence needs to be embodied, as intelligence presupposes being-in-the-world (in the Heideggerian sense).[12]

Most entities that are currently considered artificially intelligent robots are only peripheral devices controlled by a computer (a so-called server) in a network or cloud environment. While processors are indeed built into these robots, they serve only as distributors of information in the robot, while the AI code (e.g., in the area of machine vision and language processing, but also machine learning) is practically never executed on the processor built into the robot. The hardware installed in the robot is usually not designed for such resource-intensive computations. Furthermore, a server or AI software running on a network usually controls not just one robot, but any number of robots of the same (or even different) types. However, even this controlling software outsources various complex computations to more specialized AI applications, e.g., for processing speech. The main software just puts the threads together to control a group of robots.

In humans, the brain and body form a physical unit.[13] A human is a self-contained entity to which mental states can be attributed, of course, depending on how one thinks about the mind-body problem. A robot, however, is physically separate from the computer whose software controls it. The "brain" of a robot is—as explained above—usually not located in the robot itself, but in a computer center, which exchanges data with the robot via the Internet (or also with the help of other techniques of digital data transmission), processes input and controls corresponding output commands. At the same time, this computer is not only the "brain" of this robot, but the brain of very many robots.

To assume that a robot possesses mental states, moral rights or similar because it is controlled by an AI is therefore a misunderstanding. For example, neither my hand nor my intestinal wall possesses mental states and moral rights, but I do, in my wholeness of being human. If someone breaks my little finger, it is not my finger that feels pain but me. This person also does not commit an injustice to my finger but to me. Accordingly, a robot cannot be sentient and moral either, but—if at all—the entire system in which the robot is integrated. However, this raises numerous mereological questions as to which components belong to this system at all, and what is the concrete object of which mental states or the like are expressed. Unlike in the case of humans, who are more or less self-contained physical entities, these questions remain puzzling with respect to robots and AI in terms of their metaphysical presuppositions. But when, for example, the misogynistic regime in Saudi Arabia grants civil rights to the

---

[12] Hubert Dreyfus, *What Computers Can't Do: The Limits of Artificial Intelligence*, 7th ed., Perennial Library (New York: Harper & Row, 1986); Hubert Dreyfus, *What Computers Still Cannot Do: A Critique of Artificial Reason* (Cambridge, MA: MIT Press, 1999).

[13] This assumption, of course, can be criticized. However, any such criticism would not be an answer to the mereological problem regarding robots, but rather show that the same problem occurs also with regards to humans and their mental states, moral rights etc.

android robot woman Sophia,[14] or when people fall in love with artificially intelligent robots in the future,[15] but also when the European Parliament elaborates a concept on electronic persons, [16] this metaphysical mysteriousness also becomes a practical problem. For individuals can only possess and exercise rights if it is clear who or what exactly these individuals are, and which parts belong to them (and which do not).

However, this mereological problem leads far beyond AI-based robots. I show this in the following two sections, in which I introduce the thought experiment of Theseus' computer, which I use to show that the mereological underdeterminacy of AI poses practical problems in several respects at once.

Theseus' Computer: What Is AI, What Is Periphery?

Building on what has been said about robots, the question of which concrete entities count as AI raises mereological questions not unlike those of precisely determining the essence of a human being. In doing so, my following considerations presuppose a so-called Aristotelian essentialism. By this I mean the basic idea, loosely based on Aristotle's metaphysics, that things possess some attributes essentially, other attributes only accidentally.[17]

While the question of which attributes are essential to a human being and which are merely accidental can often be answered intuitively, intuitions about computers and AI have their limits. My left hand, for example, is a part of my body, it stands in a mereological relation to it, respectively to me. If I would lose my hand due to an accident or similar, I would still be me, my hand is not a necessary part of me. My left hand does not belong to my being or my essence.

But what belongs to the essence of a computer or an AI? In reference to the ancient Theseus paradox, this question can be illustrated by the following thought experiment: Theseus is a teenager who programs artificial intelligences in his spare time. His favorite project is an AI called Minotaur, which is supposed to find exits from winding mazes on its own based on machine learning with neural networks.

Since his computer is getting a bit old, he asks his friend Ariadne to replace some components. Ariadne gradually replaces the graphics card, hard drive, and motherboard of Theseus' computer with more powerful models and copies all the data (including the compiled AI and the uncompiled source code) to Theseus' new hard

---

[14] Cleve Wootson, "Saudi Arabia, which denies women equal rights, makes a robot a citizen," 2017, https://www.washingtonpost.com/news/innovations/wp/2017/10/29/saudi-arabia-which-denies-women-equal-rights-makes-a-robot-a-citizen.

[15] David Levy, *Love and Sex with Robots: The Evolution of Human-Robot Relations* (New York: HarperCollins, 2007).

[16] Loh, *Roboterethik, 84-5.*

[17] Willard Van Orman Quine, "Three Grades of Modal Involvement," in *The Ways of Paradox and Other Essays* (New York: Random House, 1966), 156–74; Kripke, *Naming and Necessity;* for my own take on Aristotelian essentialism, see Krebs, *Kripkes Metaphysik Möglicher Welten*, chapter 2.4.

drive. Since Ariadne still has good use for Theseus' old components, especially the hard drive and motherboard, she installs them in her own computer. Being curious about Theseus' latest progress on his Minotaur project, she starts the AI that is still on Theseus' old hard drive.

The philosophical paradox arising from this thought experiment is: which is the original Minotaur? The one AI that is on Theseus' new (improved by Ariadne's help) computer, or yet the one AI that Ariadne just started on the original components of Theseus' computer?

Unlike the ancient Theseus paradox, this paradox is puzzling on two levels, both software and hardware. Before discussing the genuine mysteriousness of the nature of AI at the software level (see next section), I first show some considerations about the hardware level. These are not necessarily original compared to the ancient paradox of Theseus, but they are highly relevant philosophically when computers and AI, respectively, are ascribed mental states, moral rights, and other such attributes.

In computer technology, components that are located outside the central unit of a computer are called peripherals. These include, for example, the mouse, keyboard, monitor, and also network and graphics cards. It stands to reason to assume that these devices can be replaced without changing the essence of a particular computer—much like it stands to reason that Theseus' ship will still be Theseus' ship even if you replace the sail or steering wheel.

However, if mental states are attributed to an AI, which can be traced back to "sensory perceptions," the input by sensors, already the installation or de-installation of peripheral devices such as microphones, webcams etc. can seriously change the nature of the mental states of an AI. For instance, a webcam with slightly higher resolution would lead to a completely different visual "perception" of the AI. Comparable considerations are usually discussed in relation to humans under the heading of enhancements, the optimization of humans through technology. In a sense, my glasses already have a serious influence on my sense of sight, but hardly anyone would seriously doubt that I am still me after I have replaced my glasses with ones with a higher diopter number. The same applies, for example, to prostheses, hearing aids, etc., and even with futuristic-looking enhancements such as the Eyeborg color sensor by cyborg activist Neil Harbisson, it will be difficult to argue that Harbisson is no longer Harbisson.[18]

Unlike humans, however, even those parts of a computer that do not belong to the periphery but form its central unit can be easily replaced and improved.[19] What exactly counts as the central processing unit of a computer is disputed in computer science: some definitions also include the main memory (RAM), the entire motherboard, and even the hard disk; others only the processor (CPU) or even the

[18] Harbisson, Neil, "I listen to color," 2012, TEDGlobal, http://ted.com/talks/neil_harbisson_i_listen_to_color.

19 I am not speculating about computer-brain interfaces as they are currently discussed mostly among transhumanists.

processor core (the concrete microchip). But it seems questionable whether replacing the processor core (or the entire motherboard) changes the nature of the computer or the AI implemented on it.

While these and similar problems also arise with respect to the ancient paradox of Theseus, the computer version of the paradox opens up yet another level: namely, with respect to the metaphysical status of an AI, it is completely unclear whether "AI" denotes the software or a concrete hardware realization of that software. This supposedly only theoretical question, however, becomes immediately practical exactly when mental states and moral rights are attributed to AI.

Theseus' AI: Universality and Individuality of Computer Programs

Every computer program (software) can be reduced to electronic circuits (hardware). A program is nothing more than a description or prescription of how certain electronic circuits are to behave. The program in turn has a counterpart on the hardware, where it is represented in some form (be it optical, magnetic or electrical). It is at this point, however, that the question of what exactly an AI is becomes philosophically strange. This is aptly summarized, for example, by the media theorist Friedrich Kittler with his famous bon mot "There is no software."[20] If there is no software, however, the question of what exactly an AI is becomes philosophically odd.

To make this oddity conceptual, it is helpful to become aware of the functioning and technical structure of a computer program: Programmers produce the source code of a program, i.e., the collection of those algorithms which determine the so-called output depending on the respective input. This source code, however, is not the actual program, but only an abstraction of the machine language that can be understood by humans. This source code must first be made "readable" for machines. For this there are two usual procedures: Either the entire source code is compiled into machine language by a so-called compiler before it can be executed, or the source code is translated line by line into machine language by a so-called interpreter and executed directly. Which method is used usually depends on the chosen programming language. Currently, the most popular programming language for AI application is Python, which is an interpreter language, but can also be compiled.

Regardless of whether the source code is compiled or interpreted, the question arises whether the mere source code of a program already constitutes AI. After all, Theseus "created" his Minotaur AI by saving the source code of the Minotaur in a text document. However, to classify the source code alone as AI would be absurd, at least if one ascribes certain mental states or moral rights to an AI program. The source code of a program *is* merely an ordinary text document whose content corresponds to the

[20] Friedrich A. Kittler, *The Truth of the Technological World: Essays on the Genealogy of Presence*, trans. Erik Butler (Stanford University Press, 2014), 219.

syntax of a programming language. However, hardly anyone would ascribe mental states or significant moral rights to a text document (which includes, for example, the file in which this paper is stored). One could even take this further and raise the question of whether also handwritten source code could be called an AI (and whether handwritten documents, accordingly, should also be seen as something possessing mental states and moral rights).

There are countless copies of the source code of every AI program, not only because of regular backups, but also because of the technical structure of computer operating systems. These copies match the original exactly, so that in the case of digital copies—unlike analog copies—it is no longer possible to distinguish which text document is now the original.[21] Although so-called generation loss is also possible in current computer technology when copying files, i.e. the loss of individual bits when copying files, this does not provide a criterion for distinguishing between the original and a copy of files, either practically or theoretically. Thus, Theseus has not only one Minotaur on his computer, but countless identical Minotaurs. Likewise, in the thought experiment sketched above, Ariadne has innumerable files with the same source code, i.e. also on her computer there is not just one exact copy of the Minotaur, but innumerable ones.

From a metaphysical perspective, the concept of AI therefore involves a problem of individuation, since it is impossible to determine which of these files contains the actual Minotaur, and if so, from how many copies on a new Minotaur is created (assuming Ariadne changes only one line of the source code, is this already a new individual?) and whether then perhaps even Theseus' and Ariadne's computers each house innumerable artificially-intelligent entities, to which all mental states and moral rights are to be attributed, if one assumes that AIs possess these attributes.

This individuation problem exists, however, even if one does not count the source code as AI proper, but only its translation into machine language or the execution of this machine language by the computer. Indeed, if one assumes that only the execution of an AI program constitutes an AI capable of mental states and, moreover, entitled to moral rights, little is gained for the solution of this problem. In fact, this would mean that every time a program is restarted, a new conscious individual would be created, and this individual would be killed with the termination of a program.

One possible objection would be to claim that quitting a program merely means putting a conscious individual into a kind of artificial coma, which would be awakened by the restart. But if AI has consciousness and moral rights, it would then be ethically dubious to restart a program (or even the computer) without first asking permission. At the latest when a program is recompiled (especially if small changes

---

[21] Armin Nassehi, *Muster: Theorie der digitalen Gesellschaft* (Bonn: Bundeszentrale für Politische Bildung, 2020); Michael Betancourt, *The Critique of Digital Capitalism: An Analysis of the Political Economy of Digital Culture and Technology* (New York: Punctum Books, 2015).

have been made to the source code beforehand), this objection falls short. With the recompilation the old program is completely overwritten, at the latest here a new conscious individual would have been created, while the previous program would be "killed." Then, however, each overwriting of existing programming code and the recompilation necessary thereupon would mean to murder a conscious individual. With interpreted programming languages, each restart of the program would be connected automatically with a re-creation of a conscious individual, since the source code is always translated thereby from scratch again into machine language. Software engineering—of whatever kind—would then to be rejected for moral reasons.

Although this sounds absurd, this is—following my argumentation—a direct consequence of the assertion that an AI possesses mental states. In fact, a similar argument can be found in Thomas Metzinger's work, according to which the creation of artificial consciousness is ethically questionable. Metzinger assumes that the "first machines satisfying a minimally sufficient set of conditions for conscious experience and self-hood would find themselves in a situation similar to that of the geneti-cally engineered retarded human infants. Like them, these machines would have all kinds of functional and representational deficits—various disabilities resulting from errors in human engineering."[22] Creating artificial consciousness, according to Metzinger's argument, produces unnecessary suffering. This argument is, of course, not about AI in the technical sense presented in this paper, but explicitly about artificial consciousness. Metzinger does not claim that every AI has consciousness. He merely assumes that, according to his own naturalistic theory of consciousness, the creation of artificial consciousness is possible, although this artificial consciousness need not necessarily be based on AI in the computer science sense.

Nevertheless, Metzinger's argument leads into an objection, interesting in the context of Theseus' computer, to the thesis that an AI (or a machine on which AI is realized) possesses consciousness (and/or deserves moral rights). In so far as this is true, any change in the source code of a program, including the necessary recompilation/interpretation, would be tantamount to erasing the existence of a conscious individual due to design errors and replacing it by the creation of a new conscious individual. That software engineering is a highly morally questionable activity would thus be a direct consequence of the thesis that AI possesses mental states. This, of course, does not refute computer functionalism (and numerous similar positions). To consequently reject any form of software engineering on the basis of ethical considerations, however, is in stark contrast to the enthusiasm for technology and innovation that some proponents of the thesis that AI can possess mental states currently embody in public.

---

[22] Thomas Metzinger, *The Ego-Tunnel: The Science of the Mind and the Myth of the Self* (New York: Basic Books, 2009), 195. See also Thomas Metzinger, "Artificial Suffering: An Argument for a Global Moratorium on Synthetic Phenomenology," *Journal of Artificial Intelligence and Consciousness* 08, no. 01 (2021): 43–66.

The puzzles formulated in this paper have thus shown, above all, into which strange absurdities the thesis that AI possesses mental states necessarily leads, if one considers the fundamental metaphysical question of who or what the individuals are at all, about whom corresponding attributes are sometimes all too carelessly stated in the current discourse.

Speculations

This semantic and metaphysical puzzles pointed out in this paper have shown that the question of what AI is, is problematic. However, this problem must be answered especially by those who ascribe various attributes to AI (or software, or computers in general) in the current discourse. Only by expressing certain attributes, there is an argumentative obligation to define whom or what the attributes are expressed about.

It is important to note that the puzzles also arise when—as is often the case in the current discourse—we are not talking about AI, but about so-called Artificial General Intelligence (AGI). This refers to those AIs that are not only capable of solving a specific task, but can generally solve all (or at least most) tasks that previously could only be solved by human intelligence. With respect to an AGI, the questions and problems posed in this paper are even stranger, since an AGI does not currently exist. Even futurologists speculating at length about the consciousness of an AGI, such as Max Tegmark, admit that "there's absolutely no guarantee that we'll manage to build human-level AGI in our lifetime—or ever."[23]

Furthermore, since it is at least questionable whether the construction of an AGI is even technically possible, it is also entirely speculative as to how such an AGI could possibly be constructed. However, this makes any statement attributing mental states to an AGI a statement about the extension of an empty concept (comparable to a statement about unicorns, see above). The semantic and metaphysical puzzles pointed out in this paper, therefore, become all the more absurd, the less the form of AI of which certain attributes are said to be AI at the present state of the technology.

[23] Max Tegmark, *Life 3.0: Being Human in the Age of Artificial Intelligence* (London: Allen Lane, 2017), 132.