ARTIFICIAL INTELLIGENCE. TRADITIONAL EXPECTATIONS AND A NEW CATEGORY OF MACHINES[1]

MARTINA HEßLER

In his short book, "Mein Algorithmus und ich" (My Algorithm and I), the writer Daniel Kehlmann writes about his experiences with an algorithm named CTRL (Kehlmann 2021). He had received an invitation from a Cloud Computing Company to write a story together with the algorithm. Kehlmann enthusiastically accepted the invitation since he expected to get a glimpse into the future: If he—as a writer—was to become obsolete in the near future, he would be the first to know. Now, what did Kehlmann learn about the future?

Although the algorithm produced some impressive sentences Kehlmann concluded that it was not possible to jointly write a story with this algorithm. CTRL was unable to conceive of a plot, and it started to produce nonsense after a relatively short time. As Kehlmann writes: "… then the wheels jam, the transmission is blocked, then it's over, and you have to start a new story." (Ibid.: 48)

Two things are remarkable here. First, Kehlmann uses a classical machine metaphor to describe the failure of his experiment with AI: Wheels jam, the transmission is blocked. Second, during his narration he repeatedly comes back to one of his initial expectations, i.e., that CTRL could become a writing companion, a counterpart alongside whom human authors could create a story, the way they would be able alongside a human writer. However, Kehlmann's unexpressed prerequisite was that he expected to encounter a humanized technology as his counterpart.

1 A short version of this article was first published as follows: Künstliche Intelligenz: Eine neuartige Kategorie von Maschine oder die Vermenschlichung der Maschine / Artificial Intelligence: A New Category of Machine or the Humanisation of the Machine, in: Keskintepe, Woschech 2021.

This leads to the core question of this article, namely to what extent AI represents a new category of machines and if so, what kind of machine? The focus here is not only on the transformation of concepts of machine. Inseparably intertwined with that is the change in human-machine relationships and the change in human self-image.

HISTORICAL CONCEPTS OF MACHINES

The concept of the machine is currently experiencing a remarkable renaissance.[2] Within the context of digitisation and Artificial Intelligence, there is even talk of a "second machine age" (Brynjolfsson/McAfee 2014). But what is astonishing is how matter-of-course this has been. Why do we talk of AI as a machine at all? After all, AI does jar with traditional notions of machines. It represents a completely new category of machine. But what kind of "machine" is AI in the first place? In fact, is it even a "machine"?

In the early modern era, the mechanical machine became a metaphor for processes that were regular, regulated, and reproducible ("running like clockwork").[3] Furthermore, in the 19th century, the concept stood for smooth-running, rational processes. As Joseph Corn (2011: 29) stated: "Clocks spread the technical ideal of machines that run themselves with minimal control or intervention by a human operator." These notions shaped the image of the machine right through to the second half of the 20th century as something that was regular, uniform, always the same, reliable, dependable, and predictable. The mechanical machine performs according to comprehensible and clear rules.

And even if, until recently, computers and AI continued to correspond to this image of the programmed, rule-based machine (Heintz 1993), a new dimension to the machine concept had already been identified as early on as the 1950s, precisely at a time when computers were "evolving". In the 1960s, the philosopher Gotthard Günther (1963) spoke of the "transclassical machine", which represented a new category of machines. Such a machine no longer performed work, as the "classical Archimedean" machine did, but processed information (ibid.: 183). Similarly, Max Bense (1955: 7) emphasized that the "mathematical machines (…), occasionally also called thinking machines," represented a "new state of being of technology" (ibid.: 8). Regarding the "thinking machine", French cyberneticist Louis Couffignal (1955: 13) stated that machines were "of the most diverse kind", which is why machines had to be classified using different categories.

Thus, the advent of the computer in the 1950s made it necessary to identify and categorize these new types of machines.[4] Now, the talk turned to a new "state of being" of the machine. It put the spotlight on computers' ability to process information instead of them performing work. Further, it referred to the difference between mechanical and mathematical machines.

However, Heinz von Foerster (1993: 244-252), also inspired by cybernetics, added another thought beyond the rule-governed nature of machines. He developed the concept of a "non-trivial machine", which he distinguished from "trivial machines". The trivial machine resembles the mechanical machine by working in a rule-governed way, being comprehensible, and predictable. The non-trivial machine is a machine whose behaviour is not predictable and not understandable.[5] Thus, Heinz von Foerster anticipated — regardless of the various different technological foundations — characteristics which distinguish today's AI from other types of machines, as will be further explained in a moment. His concept of a non-trivial machine can therefore not be subsumed under the two categories of machines already mentioned, since it points ahead to present AI conceptually.

2 Recently, "the machine" or "technology" have been the subject of different monographs and articles. Cf. e.g. Burkhard 2018, who examines the void of a philosophical theory of machines, Schatzberg (2018), Technology. Critical History of a Concept, Chicago/London; Popplow (1998) who follows the history of machines in early modern time; or for a short overview cf.: Heßler 2020.

3 On the metaphor of the mechanical clock cf. Mayr 1986.

4 The thesis of three different categories of machines can only be sketched briefly here. A longer work is in preparation.

5 The distinction between the trivial machine and the non-trivial machine is not that simple, as von Foerster argues. However, it cannot be elaborated on that in the realm of this article. See also: Kaminski 2014.

THE NOVELTY OF AI

Given the latest trends in AI and its pervasiveness in everyday life, the history of the machine concept needs to be updated again.[6] A third category of machines has emerged: Machine learning represents a categorically different and novel form of machine.

Perhaps the current renaissance of the machine concept thrives on the hype that surrounds *machine learning*, which is the mainstay of present-day AI successes. The categorial novelty is featured in the name itself: adaptive and learning machines are what we are now dealing with.[7] They learn by using data; the algorithms are self-improving. Programmers have ceased programming how an algorithm ought to arrive at a solution; instead, they program algorithms capable of learning from their "experience" and then developing a model by themselves to apply to other data. They are, therefore, developing autonomously (Kersting, Lampert, Rothkopf 2019: 19).

From a historical perspective, this is not something entirely new. Neural networks were already being discussed in the 1940s and 1950s. Further, cybernetic machines were also "learning". Research into *machine learning* has been conducted since the 1980s (See overview Lenzen 2018). The fact that AI is now celebrating new successes is down to a new type of data volume, high processing speeds, and a new quality of efficient algorithms. It means that AI is no longer comparable to 20th century machines. But what exactly is so novel about it—beyond its essential "learning" aspect? The following will give a description of seven characteristics which are mutually dependent.

First, AI interacts in a new way with its users. Let's return to the opening example of writing a story jointly with AI, and this becomes very clear now. The mediating role of technology in the writing process has already been addressed many times. People have written with fountain pens, typewriters, or computers, which has influenced and altered the writing process in each case (See overview: Gaderer 2020). However, writing a story jointly with an algorithm, represents a completely new dimension—even if AI's possibilities are currently still limited, as Kehlmann aptly put it. Many other examples for human-algorithms cooperation could be added. AI has become an advisor, a partner, and an assistant in everyday life. AI can no longer be seen as a mere tool used to edit, create, or calculate something. Decisions are now made, discussed, or written in conjunction with AI, as part of a densely interwoven network of actions.

Here, and this is a second central key aspect, AI applications develop individually while interacting with its counterpart. AI is adaptive. Siri, Alexa, chatbots, social robots, and my previous example, the algorithm CTRL, learn the behavioural patterns, matters of interest and preferences of their counterparts and, over time, "advise" and "respond" more and more precisely, in keeping with those individual habits. Historically, this is a remarkable turning point. The machine-like aspect is no longer that which is standardised and ever constant. Instead, AI systems evolve differently. They have their own individual "biography", one that depends on their counterpart and the context in which they are used. It is the everyday behaviour, the everyday use by each individual that permanently modifies the AI.

Third, this in turn implies yet another characteristic that differentiates AI from machines in the traditional sense. The evolution of algorithms that learn autonomously cannot be predicted or planned, a feature of AI that resembles von Foerster's concept of a non-trivial machine. How AI will respond and interact with human beings in a year's time is anyone's guess. This means that machines also have a future all of their own.

Fourth, and connected with that, their results are not reproducible, another fundamentally new characteristic compared with the sort of machines we have been familiar with so far, which had no surprises in store; indeed, we could rely on their "sameness". However, talk to a bot today, for instance, and you quickly realise that the same question produces different answers.

6 Kaminski and Gelhard (2014) have published an edited volume in which the authors also put forward the thesis of a new category of machine. They coined the term "informal technicization" (informelle Technisierung) and focused on a new human-machine relationship in which technology becomes imperceptible because it cannot be directly experienced. The focus lies on ubiquitous computing.

7 Cybernetic machines were also learning and adaptive machines, but different ones. Cf. researches such as Cordeschi 2002; Pickering 2010; Müggenburg, 2018.

Fifth, another new dimension concerns machine decision-making. Admittedly, this is not something new. Decision-making machines have a long history. As early as the 1950s, philosopher and writer Günther Anders mocked the use of a computer to decide whether the US should end the Korean War. The computer, according to Anders, ultimately made a more ethical decision than humans did (Anders 1988: 59–64). Norbert Wiener (1958: 174–180) warned against "government machines" already in the 1950s. In the 1970s, the Allende government in Chile launched a cybernetic experiment in an effort to control the Chilean economy (Medina 2011). However, what is new is that AI, when it is applied in everyday life, must make *ad hoc* decisions. AI does not only support humans in decision-making processes. In part, it must make decisions itself—within seconds. This is what is currently being debated, e.g., when it comes to autonomous driving or the use of drones for military purposes. This raises the question of "moral machines." (Misselhorn 2018)

Sixth, and closely connected, what's also new is that machines are now capable of developing biases. Depending on the database used in each case, AI applications can, as is currently debated, perpetuate discrimination. Research into the history of technology has often shown that technology—by virtue of its cultural nature—is never neutral and that human assumptions have always been incorporated into its development. However, the fact that machines develop a bias in the course of their "actions" and then reach culturally formed decisions that are not transparent has a lasting impact on the very image of machines. In the 1950s and 1960s, it was hoped that computers would be able to make rational and objective decisions in an irrational world (Erickson 2013). Today, AI applications favour certain groups of people while discriminating against others.

Seventh, and finally, artificial neural nets are something of a black box for AI developers. The fact that technology is a black box is by no means historically new. During the 20th century, technology increasingly became a black box since most of its users no longer understand how it works. For example, while the first automobiles required their operators to possess a certain level of technical expertise, cars increasingly became a complex black box that even trained mechanics could only understand and repair with the help of diagnostic tools. Nowadays, most everyday technology is a black box for most users, be it the washing machine, the computer, or the mobile phone: a black box that can be easily operated or used, but whose functionality is not understood in everyday practice and, most importantly, also does not need to be known or understood. What is new, however, is that AI has also become a black box for the AI developers themselves. As Klaus Mainzer (2019: 254) put it, "From an engineering standpoint, authors therefore speak of a 'dark secret' at the heart of machine learning AI." He cites a 2017 article: "… even the engineers who designed (the machine learning based system) may struggle to isolate the reason for any single action." (Ibid.)

Since the early aughts of the millennium, the term "explainable AI" has emerged, and with it, a debate about the possible consequences of its incomprehensibility. In the meantime, developing explainable AI constitutes an important goal for computer scientists and societies. Within computer science as a discipline, however, this problem has been a topic of discussions for a long time. The term "black-box models" refers to this phenomenon, which goes hand in hand with probabilistic modelling and statistical learning methods, such as deep learning.[8] Thus, machines have a secret, as it were, in that their behaviour is not always "comprehensible". What's new is the incomprehensibility *in principle* that distinguishes AI as a black box from the many black boxes of consumer technology that we use so "naturally" without really understanding them.

Altogether, this gives rise to a breathtaking image of AI that radically breaks with previous notions of the machine: i.e., a machine that interacts with people, that "responds", that is adaptive, and develops in keeping with the counterpart in each case. Further, it is capable of developing bias. It is meant to make moral and ethical decisions; its behaviour changes; and it is not always reproducible and not always comprehensible, even for its developers. It becomes our advisor, companion and/or assistant. It is the interplay of all these characteristics mutually conditioning one another and rooted in machine learning that

8 https://gi.de/informatiklexikon/explainable-ai-ex-ai (02.02.2022)

makes current AI systems an entirely novel category of machine. In a nutshell, one could say that AI is becoming human-like. AI no longer corresponds to the 20th century image of the machine which humans often chose to distance themselves from. Rather, AI appears humanlike in that its behaviour is not predictable, that it is not necessarily reproducible, in that it may develop biases, and in that it has a past, so to speak: its behaviour is based on past data.

PARADOXICAL EXPECTATIONS: TRADITIONAL NOTIONS AND A NEW CATEGORY OF MACHINE.

During the 20th century, Western discourses on technology concerns focused on the fear that humans might become machine-like. Up until the 1970s, critical voices uttered dread that machines would lead to totalitarian rationalization, de-subjectification, and that uniformity of humans would ensue. People though insisted that they were different from machines. Behaving like a machine had clear negative connotations.

As argued above, AI no longer corresponds to this machine image of regular, mechanical, and always "perfect" machines. AI now simulates many features that have earlier been emphasized as typically human *in contrast* to the mechanical: individuality, subjectivity, emotionality, the unpredictability of human actions, the human capacity for ethical action. That means that AI challenges the human self-image, insofar as humans still define themselves in contrast to the machine.

Further, AI as an apparently "humanized" category of machine challenges human expectations towards machines. Humans tended to expect "high standards" from machines, e.g., a machine should be devoid of bad human characteristics, it should do "better" than humans. Early on, the pioneering AI researcher John McCarthy stated that nobody wanted a computer that loses its nerve (Cited in Lenzen 2018: 244). Or, to put it favourably, it should be a companion, just as Daniel Kehlmann imagined.

However, human expectations for AI as an "objective" advisor or a reliable companion point to a paradox: AI's described properties, which make it a new category of machine and apparently human, simultaneously mean that AI loses many of the properties of the classical machines, which humans have come to appreciate: the expectation of humans exerting dominance and control over machines, the assumptions of objectivity, reliability, predictability, and rationality. Exactly these qualities were attributed to the classical machine. However, these same properties cannot be ascribed to AI as a novel category of a machine.

Concepts of traditional human-machine relationships and traditional images of machines still shape expectations towards AI. That is particularly true for an anthropocentric attitude that sees humans as controllers, describes technology as assistance, emphasizes the hierarchy between humans and machines, and thinks in terms of dualisms. In a nutshell: Historically determined concepts of humans and machines shape our expectations of AI—although they are no longer suitable.

It seems that humans have not fully grasped yet the ambivalence and novelty of AI. They expect a categorically novel machine to behave in the same manner as the classical machine did. Daniel Kehlmann fell into the same trap when he commented on the failure of AI to write a story jointly with a human; his measuring stick was the concept of a classical machine against which AI failed miserably. It seems that people are not aware of this paradox: On the one hand, AI is supposed to become more human-like, and its new properties such as adaptability, learning ability, development ability, and individualization indeed

make it apparently more human-like. On the other, however, AI is supposed to be reliable, transparent, predictable, and objective. However, if AI becomes more human-like, it will also display more of what is regarded as human "failure" and unreliability.

REFERENCES

Anders, Günther (1988), Die Antiquiertheit des Menschen. Bd. 1, Über die Seele im Zeitalter der zweiten industriellen Revolution, München.

Bense, Max (1955), Vorwort. In: Louis Couffignal, Denkmaschinen. Stuttgart.

Brynjolfsson, Erik/McAfee, Andrew (2014), The Second Machine Age, New York.

Burkhard, Martin (2018), Philosophie der Maschine, Berlin

Cordeschi, Roberto (2002), The Discovery of the Artificial. Behaviour, Mind and Machines Before and Beyond Cybernetics, Dordrecht.

Corn, Joseph J. (2011), User Unfriendly. Consumer Struggles with Personal Technologies, from Clocks and Sewing Machines to Cars and Computers, Baltimore.

Couffignal, Louis (1955), Denkmaschinen. Stuttgart.

Erickson, Paul u.a. (2013), How Reason Almost Lost its Mind. The Strange Career of Cold War Rationality, Chicago.

Foerster, Heinz von (1993), Wissen und Gewissen, Frankfurt am Main

Gaderer, Rupert (2020), Schreiben, in: Martina Heßler/Kevin Liggieri (Hg.), Handbuch Technikanthropologie, Baden Baden, S. 502–506.

Günther, Gotthard (1963), Das Bewusstsein der Maschinen, Krefeld, Baden-Baden

Heintz, Bettina (1993), Die Herrschaft der Regel. Zur Grundlagengeschichte des Computers, Frankfurt am Main.

Heßler, Martina (2020), Maschinen, in: Martina Heßler/Kevin Liggieri (Hg.), Handbuch Technikanthropologie. Baden Baden, S. 256–262.

Kaminski, Andreas (2014), Lernende Maschinen: Naturalisiert, transklassisch, nichttrivial? Ein Analysemodell ihrer informellen Wirkungsweise, in: Andreas Kaminski/Andreas Gelhard (Hg.), Zur Philosophie informeller Technisierung, Darmstadt, S. 58–81.

Kaminski, Andreas/Gelhard, Andreas (Hg.) (2014), Zur Philosophie informeller Technisierung, Darmstadt.

Kehlmann, Daniel (2021), Mein Algorithmus und ich, Stuttgart.

Kersting, Kristian/Lampert, Christoph/Rothkopf, Constantin (Hg.) (2019), Wie Maschinen lernen. Künstliche Intelligenz verständlich erklärt, Wiesbaden.

Keskintepe, Yasemin/Woschech, Anke (Ed.): Künstliche Intelligenz/Artificial Intelligence. Maschinen Lernen Menschheitsträume/Machine Learning Human Dreams, Göttingen.

Lenzen, Manuela (2018), Künstliche Intelligenz. München 2. Auflage.

Mainzer, Klaus (2019), Künstliche Intelligenz – Wann übernehmen die Maschinen? Berlin, 2. Auflage

Mayr, Otto (1986), Uhrwerk und Waage: Autorität, Freiheit und technische Systeme in der frühen Neuzeit, München.

Medina, Eden (2011), Cybernetic Revolutionaries: Technology and Politics in Allende's Chile, Cambridge, Mass.

Misselhorn, Catrin (2018), Grundfragen der Maschinenethik, Stuttgart.

Müggenburg, Jan (2018), Lebhafte Artefakte. Heinz von Foerster und die Maschinen des Biological Computer Laboratory, Konstanz.

Pickering, Andrew (2010), The Cybernetic Brain. Sketches of Another Future, Chicago and London

Popplow, Marcus (1998), Neu, nützlich und erfindungsreich. Die Idealisierung von Technik in der frühen Neuzeit. Münster

Schatzberg, Eric (2018), Technology. Critical History of a Concept, Chicago/London

Wiener, Norbert (1958), Mensch und Menschmaschine, Frankfurt am Main.