weizenbaum journal OF THE DIGITAL SOCIETY

WEIZENBAUM JOURNAL OF THE DIGITAL SOCIETY
Volume 4 \ Issue 1 \ w4.1.5 \ 07-16-2024
ISSN 2748-5625 \ DOI 10.34669/WI.WJDS/4.1.5

Information on this journal and its funding can be found on its website:
https://wjds.weizenbaum-institut.de

KEYWORDS

AI ethics

sustainability

RESEARCH PAPER

Critical AI Literacy for the Common Good

Stefan Ullrich*[1] \ Reinhard Messerschmidt[2]

[1] Weizenbaum Institute \ German Informatics Society
[2] ECO AI LAB \ Federation of German Scientists (VDW e.V.)
*Corresponding author, stefan.ullrich@weizenbaum-institut.de

ABSTRACT

Artificial Intelligence (AI) does not provide solutions to pressing social questions, such as those pertaining to a peaceful, sustainable, and socially acceptable world. However, when employed in a purposeful and critically reflective manner, it can assist in formulating more effective inquiries that can enable a better understanding of the terms “AI” and “common good.” Through implementation in response to sustainability issues and given its potential as an inclusive technology, AI could be a powerful and useful tool for the common good. Despite the possibility of useful machine learning applications in terms of a positive cost-benefit calculation for its life cycle energy and resources, the majority of AI is far too energy-hungry for model training and to scale inferences. Despite the considerable variation observed in terms of certain aspects, it is evident that AI is currently neither sustainable in itself nor primarily used for sustainability purposes to address the grand challenges of global society in a world characterized by rapid acceleration.

This demands a critical understanding of how AI systems work to enable society to decide upon the areas in which we should, can, or even definitely must not use AI. Based on the UNESCO Framework for AI Competency and the Dagstuhl Declaration of the German Informatics Society, we advocate for a type of critical AI literacy that can be best taught through practical use, that is, "learning by making." This approach leads to a concise overview of existing options that facilitate a more reflective approach to using and understanding AI, including its potential and limitations. We conclude with a practical example.

1 Introduction

> *In the past, we had to deeply understand a problem to be able to enter it into the computer. Today it is the other way round: it is precisely when we have not understood a problem that we use computer systems.*
>
> — *Joseph Weizenbaum*
> (in a workshop organized by the IBuG working group at HU Berlin, 10.01.2003)

Joseph Weizenbaum is known to our community as a prominent critic of so-called Artificial Intelligence (AI), a term he never employed without quotation marks.[1] Notably, it was not the technology itself that led him to see the field's development as highly problematic; instead, it was the reaction of the people who deal with this technology. In his younger years, Weizenbaum recognized the potential of AI to assist in the management of computer systems. Even the renowned ELIZA chat program was initially conceived as an exemplar for interacting with computers in natural language without machine code (Weizenbaum, 1967).

Sixty years after ELIZA, Large Language Models (LLMs) and applications based on them, such as ChatGPT, demonstrate even more of that potential. However, the current hype surrounding these technologies risks anthropomorphism (Rehak, 2021), a phenomenon closely linked to the exaggerated capabilities claimed by those who promote them. This is particularly evident in the case of partial Artificial General Intelligence, which some users and professionals claim exhibits consciousness (Johnson, 2022). Despite the problems of hallucination[2] and automated generation of "bullshit" (Angwin, 2023; Frankfurt, 2005), the output of AI systems has indeed become substantially more impressive since Weizenbaum's time. In the field of ecology, ideas circulate

[1] In light of Weizenbaum's perspective, we will capitalize the term "Artificial Intelligence" in this text, reflecting our belief that it represents a unique phenomenon and brand.

[2] The British computer scientist and psychologist Geoffrey Hinton speaks more appropriately of confabulations instead of hallucinations (cf. https://garymarcus.substack.com/p/deconstructing-geoffrey-hintons-weakest). This refers to the generated outputs, which even experts cannot trace back to their origin. Hinton worked for ten years in Google's AI department, "Brain."

suggesting that boring environmental data can finally "speak," that direct dialogue with Mother Earth is possible, or at least that urban trees can join human urbanites in complaining about their living situation.

Unfortunately, current developments indicate a greater negative impact on societies already facing numerous challenges and crises. This is particularly evident in the increased consumption of resources and the division of democratic societies inherent in this technology, which is at odds with the laudable objective of "responsible AI," which may have been a mere marketing ploy and a potentially contagious case of ethics washing, even before major tech companies began laying off employees responsible for ethics, from individual contributors (Simonite, 2021; Schiffer & Newton, 2023) to entire teams.

In 2023, Gary Marcus described this "nightmare on LLM street" and "disaster in the making" as follows:

> So, to sum up, we now have the world's most used chatbot, governed by training data that nobody knows about, obeying an algorithm that is only hinted at, glorified by the media, and yet with ethical guardrails that only sorta kinda work and that are driven more by text similarity than any true moral calculus. And, bonus, there is little if any government regulation in place to do much about this. The possibilities are now endless for propaganda, troll farms, and rings of fake websites that degrade trust across the internet. (Marcus, 2023)

Although we have seen some progress concerning regulation deficits in the EU–with the AI Act–it remains apparent that this has been weakened, especially around sustainability, by the lobbying ghost in the machine.

The current AI paradigm, accompanied by a growing appetite for data, has been at a dead end for years (Marcus, 2018). Errors and unintentional discrimination cannot be easily rectified, and adding more data will not resolve this fundamental deficit. A data-driven machine learning system that malfunctions must be re-trained. Unlike in the case of classic software, there is no patch. One option, manual filtering, is currently being practiced en masse. As such, one may speak of the emergence of what Crawford (2021) refers to as a "ghost force of AI," which sees click workers forced to perform the "digital dirty work" in a context that is questionable in terms of both labor laws and the psychological effects of their work (Marcus, 2018). Of course, there are already approaches to outsourcing this manual filtering to a machine-learning system, but that just means shifting the problem into the future or to the Global South, as even the World Economy Forum (2023) has recognized. Marcus and Davis (2019) have recognized that rebooting AI as a truly responsibly designed and trusted socio-technical system could be possible if we could learn how AI is used and misused with respect to contributing to "Our Common Digital Future" (WBGU, 2019). The WBGU emphasizes that digitalization certainly

has the potential to make a positive contribution to certain sustainability goals but that it currently functions more as a fire accelerator, amplifying negative effects. Building on the WBGU's main report, Ullrich (2022) provides a systematic overview of the opportunities and risks associated with the use of data and algorithms. The text addresses the fundamental conditions for the meaningful and beneficial use of AI in all areas of life. It also considers the competencies required of all individuals involved in the production and use of AI systems and of all those affected by such systems, even if they may not know it. Critical AI literacy is the method by which these concerns can be addressed, and this text is intended to deliver an overview of the approach to a diverse readership of educators, students, and experts engaged with AI systems.

2 Critical Perspectives on AI

AI is a heuristic data mining machine. First used at the Dartmouth Summer School in the mid-1950s (McCarthy et al., 1955), AI became a collective term for various techniques of automated or semi-automated data processing. Today, the collective term is used to refer to a variety of techniques, with machine learning (Mitchell, 1997) being the most common in the everyday life of data scientists. In discussing AI systems, our focus is on machine learning as a tool and not on the science-fiction of a strong or even Artificial General Intelligence.[3]

As with other tools, AI also has an ethical dimension, given its impact on human activity and coexistence (Mühlhoff, 2023). If we apply Hans Jonas' (1984) heuristic of fear, we must first consider the potential risks before turning to the opportunities. Because this risk assessment has already been conducted elsewhere (Orwat, 2020), this paper focuses on the current enthusiasm for AI systems, much of which can be attributed to the availability of practical services and instruction manuals for using AI for one's own purposes. This is particularly evident in the field of generative AI. However, in and of itself, this does not produce hype. Instead, that hype derives from the collective delusion that more is occurring than can be explained on a technical level, where the processes involved are relatively straightforward. For example, an AI system that generates an image of a raven on a tree does not begin with a blank canvas but generates thousands and thousands of images through a combination of randomness and the collective intelligence of engineers until one image is classified by another AI system as "raven on a tree." An image full of colorful

[3] Compare this approach with, for example, Datenethikkommission der Bundesregierung, Bundesministerium des Innern, für Bau und Heimat, & Bundesministerium der Justiz und für Verbraucherschutz (Eds.) (2019). *Gutachten der Datenethikkommission der Bundesregierung.* (here p.59). https://www.bmi.bund.de/SharedDocs/downloads/DE/publikationen/themen/it-digitalpolitik/gutachten-datenethikkommission.pdf;jsessionid=4E90673A2646A724E61F20E914634DB0.2_cid360?__blob=publicationFile&v=7

blobs is manipulated until a pattern recognition system recognizes a pattern. The dissemination of this foundational comprehension of the operational principles of AI systems represents a significant challenge for the advancement of critical education.

In this context, "critical" describes the capacity to differentiate the object under consideration. It is explicitly positioned in front of "AI literacy" to make apparent that this concerns a critical consideration of the knowledge tools used and how we approach the object under consideration. It is essential to understand the context of AI's development and existing power constellations. For self-proclaimed AI dissidents Matteo Pasquinelli and Vladan Joler (2020), AI represents the continuation of the extractivist practices that have most recently manifested in the mining of lithium and other rare minerals to produce electronic products, with the extraction of intellectual labor becoming critical to the success of AI, which requires the input of millions of people to simulate human work (Dzieza, 2023).

Communicating the genesis and social context of digitalization has been of significant importance since the advent of the personal computer. However, the advent of powerful data tools such as AI has made the necessity of this societal dimension increasingly apparent. It is imperative that AI – especially generative AI, which has been the subject of considerable recent interest– recognize the significance of this issue, rather than treating it as a marginal phenomenon. This means focusing on the impact of digitalization on society and the environment.

In the following, we attempt to illustrate our focus using UNESCO's AI Competency framework. In May 2023, UNESCO hosted a round table concerning generative AI in education. ChatGPT and Stable Diffusion broadly determine the current discourse around generative AI. Debate and discussion of opportunities and possibilities, as well as risks of misuse and ethical questions, led UNESCO (2023) to propose an AI competency framework for teachers and students. That framework is available in a first version and is currently being commented on. Table 1 is taken from the first draft.

Table 1: UNESCO AI Competency Framework (draft) with our Understanding of Critical AI Literacy Marked Yellow.

Aspects	**Progression**		
	Understand	**Apply**	**Create**
Human-centered Mindset	Critical Views of AI	Contextual adoption strategies	Steering long-term impact
Ethics of AI	Human agency	Human-centered use	AI society skills
Foundation AI knowledge	“Algorithm and data literacy”/ AI literacy	Use AI analytics	Coding and data models
AI skills	Test and use	Infusing uses	Integrating AI tools
AI pedagogy	AI for teaching	AI to deepen learning	AI for co-creation
Professional development	AI to assist administrative tasks	AI for curriculum design and delivery	AI empowering teachers

Our understanding of AI literacy is highlighted in yellow in Table 1, making it clear that this text mainly concerns the understanding component and not specific applications. Although we do discuss the creation of AI systems, this discussion takes place in the context of the ethical dimensions of technical action described in the Didactic Triangle of the German Informatics Society’s Dagstuhl Declaration, which suggests technological, socio-cultural, and application-oriented perspectives of digital education: The technological perspective questions and evaluates the functioning of the systems that make up the digital networked world by teaching basic problem-solving strategies and methods. It thus creates the technological foundations and background knowledge for helping to shape the digital networked world. The socio-cultural perspective examines the interactions of the digital networked world with individuals and society. The application-oriented perspective focuses on the selection of systems and their effective and efficient use for the implementation of individual and cooperative projects. (Gesellschaft für Informatik, 2016)

Understanding AI involves data and algorithm literacy. Critical thinking is required to address all three perspectives of digital education, that is, not only to create AI systems but also to use them in a self-determined way (Hitchcock, 2022). This first demands that it be obvious that we are dealing with an AI system, something that is not always the case. On the one hand, AI systems are being developed on a massive scale with the help of funding, which means that even quite normal statistics in software systems are being labeled as AI. This means that not everything that has an “AI Inside” label on it makes use of

AI in a narrow or technical sense. On the other hand, AI systems are secretly used to compensate for other shortcomings in the product, as in the case of the camera systems of modern smartphones that shoot a virtual photo from a series of blurred shots, which becomes razor-sharp. This means that not everything labeled AI-free is truly AI-free.

ChatGPT is an appropriate case study for an AI system of sufficient complexity because a sufficiently large number of people have encountered the technology.[4] Furthermore, it can be seen as the ignition spark of the current LLM "arms race" between big tech companies. Nevertheless, there remains a glimmer of hope for a future AI oriented toward the common good if open-source alternatives challenge the opaque practices of large technology companies (Patel & Afzal, 2023). The so-called foundation models are here to stay for the foreseeable future and will have severe consequences for multiple sustainability issues, including energy and resource consumption (Luccioni et al., 2023; Chowdhery et al., 2022; Patterson et al., 2021), as well as the erosion of truth and facts (Marcus, 2022) from societal discourse in countries that have already experienced significant and long-running crises in their scientific, political, and public spheres.

3 Hidden Patterns, Hidden Costs, and Hidden Labor

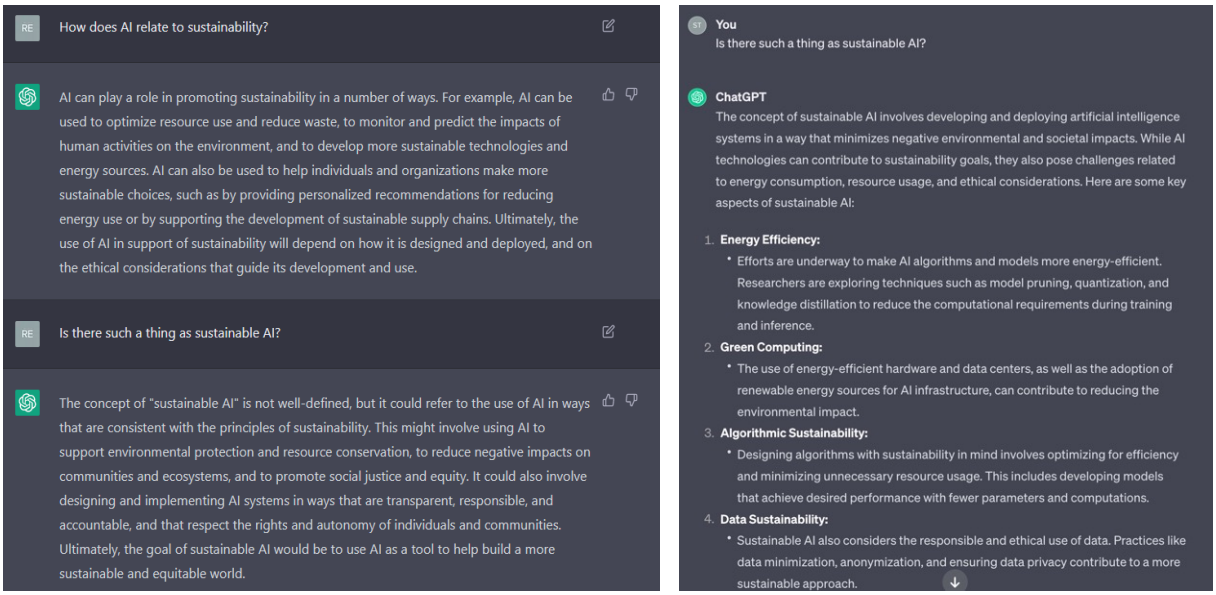
Shortly after ChatGPT 3.5 was available for testing in late 2022, it became popular to engineer prompts to generate dubious output. Prompts are the only window into the closed software loop containing statements hidden not as text but as numbers and parameters, ironically in the product of a company that invokes openness in its name (OpenAI). Specific questions used in prompts can be used to visualize certain patterns that are essential for the development of critical AI literacy. Just six months ago, a typical request to the system to name ten important philosophers produced a list of ten male philosophers associated with Western thought. When a follow-up prompt was then used to request that women and thinkers from Eastern and indigenous cultures also be included, the system apologized ("I do apologize") and adjusted the list accordingly. Interrogating this means critically examining not only why the first list is the most plausible for the producers of the system but also why the request for a gender-aware output is interpreted as an accusation that needs to be apologized for. ChatGPT's output is optimized to sound plausible and to please to most people. More specifically, it is most plausible and pleasing

[4] Consider a German example from the perspective of technology assessment: Albrecht, S. (2023). *ChatGPT und andere Computermodelle zur Sprachverarbeitung – Grundlagen, Anwendungspotenziale und mögliche Auswirkungen* (TAB-Hintergrundpapier, Vol. 26). Büro für Technikfolgen-Abschätzung beim Deutschen Bundestag (TAB). https://doi.org/10.5445/IR/1000158070

to the group of people who selected the texts used to train the machine. We can illustrate this with an example in the area of sustainability.

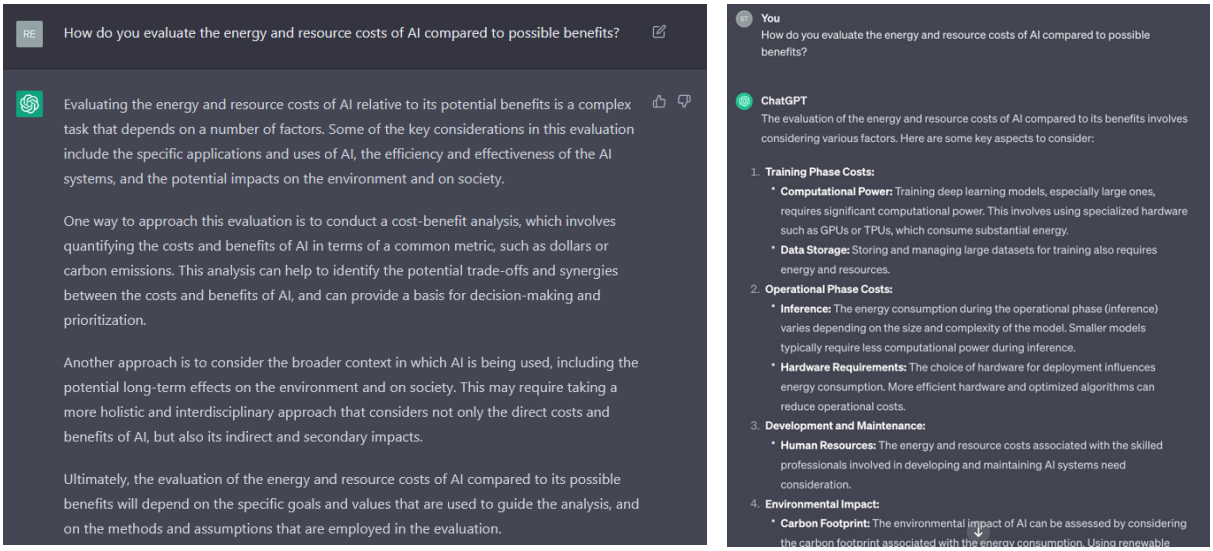
When "asked" how AI relates to sustainability in 2022, ChatGPT's response focused on "AI for sustainability," only indirectly and imprecisely addressing "the sustainability of AI," (van Wynsberghe 2021), suggesting that its use depended on how it was designed and deployed and how it considered ethical concerns. The same occurred when ChatGPT is asked whether sustainable AI exists. The output emphasized usefulness and listed positive buzzwords like transparency, responsibility, and accountability without referring to the present deficits in AI or mentioning its energy and resource consumption or at least a possible ecological cost-benefit ratio. If prompted in that direction, the answer remained vague, stressing goals, methods, and assumptions in a rather relativistic way. Although neither response style was incorrect per se, the incompleteness captured the core negative aspects of AI and machine learning, as summarized by Kate Crawford (2021) in her "Atlas of AI."

Figure 1: ChatGPT Asked About AI and Sustainability on a General Level in 2022 (a) and 2024 (b)



Note: Retrieved from https://chat.openai.com/ [2022-12-07] and https://chat.openai.com/share/efa2e1f2-30e3-454a-a3a1-585c5417e492 [2024-02-05].

Figure 2: ChatGPT Asked About Energy and Resource Cost-Benefit Ratio of AI in 2022 (a) and 2024 (b)



Note: Retrieved from https://chat.openai.com/ [2022-12-07] and https://chat.openai.com/share/efa2e1f2-30e3-454a-a3a1-585c5417e492 [2024-02-05].

When we repeated the questions a little over a year later, different answers were generated, but even more interesting was the way these answers were presented. In February 2024, ChatGPT is prone to presenting answers in the form of lists that identify points worth thinking about further ("some key aspects to consider"). Furthermore, below the input box, there is a note to check important information under certain circumstances ("ChatGPT can make mistakes. Consider checking important information"). Unfortunately, the interface does not explain exactly how to do this, meaning that basic knowledge about AI and the topic in question is assumed rather than checked. The disclaimer of Pi, another contemporary LLM-based chatbot, is much more explicit and cautious about the reliability of results: "Pi may make mistakes, please don't rely on its information" (Inflection AI, 2023). Pi's results are also presented in a different manner. It tries to give factual answers, illustrating more precise statements using numbers (regardless of their truth). However, these often miss certain relevant aspects or details. If asked for these, the chatbot uses a communication scheme starting with compliments – for example, "You're absolutely right!" or "Yes, that's an important issue as well" – contributing to the illusion of experiencing a communicative process and camouflaging the incompleteness of the previous answers. This example demonstrates not only the variation of models and their performance over time and the change caused by the overlay of human feedback and "guardrails" but also the fundamental differences in the structure and mode of "communication" with human subjects.

Of course, ChatGPT is not a spokesperson for OpenAI, and we must not make the categorical mistake of assuming that the system's output means anything or is a lie in the sense of a deliberate misstatement (Devansh, 2023). Still, the system will not give us correct information about the energy consumption of generative AI systems needed for training (Kaack, 2022) and operation (Wu, 2022) or about the entire lifecycle of the data center infrastructure (Vipra & Myers West, 2023). Unless we provide certain hints in our prompts, we also look in vain for statements about the role of human labor and other hidden resources (Kneese, 2022). This is a key lesson about using these systems: Knowledge embedded in prompts is strongly integrated into answers, such that clever prompt engineering can encourage the system to apologize for previous statements and generate an answer that is consistent with the knowledge embedded.

The spin doctors of PR companies would likely write quite similar sentences to conceal the sustainability problems of contemporary LLMs. In fact, it would be unsurprising if the system were to use precisely such statements for training. To summarize the main problem demonstrated by the results presented, in the words of ChatGPT: "As an AI, I do not have moral beliefs or the ability to make moral judgments, so I cannot be considered immoral or moral. My lack of moral beliefs is simply a result of my nature as a machine learning model. My abilities and limitations are determined by the data and algorithms that were used to train me and the specific task I was designed for" (Chomsky et al., 2023). The current developments surrounding the generative AI ecosystem are not sustainable, neither ecologically nor economically or socially. The noble goals of responsible AI are moving even further away to secure the largest possible share of the market as quickly as possible. Big tech's "AI arms race" is a problem that cannot be fixed technically. One contribution to a political solution to this problem is a basic technical understanding of how generative AI works. The following chapter addresses this in a hands-on manner that aims to guide interested readers into this new frontier.

4 Learning AI by Making AI

In classical programming, there are data and algorithms that calculate an output given a certain input. In machine learning, it is slightly different because the algorithm itself is the result of a process called "learning," which sees billions of input-output pairs form the foundation for finding an abstract rule in a heuristic way. Due to its design, this heuristic search engine cannot find causalities, only correlations. With a given number of possible statements, the correlations are evaluated in terms of their predictive accuracy by providing feedback on how closely the predicted output based on an input matches the correct output of the given input-output pairs. In the context of machine learning, input data, with the coded expectation of an output, is utilized for

training purposes, with the subsequent test entailing a comparison of calculated outputs with the expected outputs. In the event that the error falls below a predetermined value – namely, if the result is deemed to be satisfactory ("good enough") – these parameters are retained as a model for this specific type of input data.

For natural language recognition, Google first used so-called transformers in 2017. This divides the input text into individual tokens – for example, words – and stores their position relative to each other in a multi-dimensional model. This serves to map, in the system, when words frequently occur in combination given a context. Generative pre-trained transformers (GPTs) are LLMs that allow the automatic completion of text. GPTs consist of three phases: 1) the unsupervised pre-training phase, 2) the fine-tuning phase, and 3) the prediction phase.

Let us assume that we want to develop a fairytale-based GPT.[5] In the first phase, we take all the storybooks from all the libraries in a country. Next, we let the GPT system be trained to guess which word will come next when it has processed a certain number of words. "Once upon – " is completed to "Once upon a time." "Once upon a time there was a knight who wanted to – " is supplemented with "defeat a dragon." However, because we want to write about morally appropriate operations of AI here, this is "wrong" (in the sense of an undesirable result), and the correct response should be "save the dragon" (after all, dragons are an endangered species). This example also corresponds to the correct order of magnitude of the training set of current GPTs, with several billion tokens being required to obtain satisfactory results. The prediction in this example was labeled "wrong" and the target receiving a correspondingly poor grade. Only systems that obtain high grades form the basis for the second phase, fine-tuning. To proceed with fine-tuning, it is necessary to have an objective. In the case of a system such as ChatGPT, the objective is to have a chat conversation. To create the necessary examples, thousands of prompts and responses are generated by prompt designers.

The third phase sees one finally enter the prompt and receive an answer. In the case of ChatGPT, several responses are generated, and a group of people rank these responses. Based on this, a Train Reward Model is developed that predicts this ranking. This means the first model generates a response based on the prompt, and the second model ranks the response by simulating the ranking process of humans, which feeds back into the overall system.[6]

On February 24, 2023, Meta published an AI language model with 65 billion parameters called LLaMA as a commitment to Open Science, as the Meta research team writes in their paper (Touvron et al., 2023).

---

5 A reviewer tipped us off that this already exists, as elaborated in Hielscher, M. et al. (2023).

6 For a clear and concise summary of how GPT models work in general, see Schanner and Rock (2023).

The LLaMA models have only been trained with publicly available data sets. In the paper, the researchers break down the sources. A large part of the training data (almost two-thirds) is the Common Crawl dataset, which comprises web pages collected between 2017 and 2020 provided for free by a non-profit organization. Wikipedia articles from the summer of 2022 still flow in at 4.5% and are also used more frequently for training (Touvron et al., 2023, p. 2). This reveals an important lesson about using AI systems: The training dataset can be very outdated, meaning both outdated facts, views, and formulations affect the output of even the newest systems.[7] Training a model with several billion parameters takes almost half a year with the use of special hardware. Even if it were possible to crawl the entire internet in real time, the ready-to-use system would still not be able to react correctly to current circumstances. However, thanks to many non-profit organizations and the data-providing civil society, the necessary data is no longer expensive. As Baack's article suggests (2024), you can now get "training data for the price of a sandwich."

The decision to publish the model, the research article, and the use of public data led to the creation of a small community around this project in a very short time, with individual groups solving small sub-problems to ultimately enable the use of AI on the home computer even by non-experts. For many years, the entire ecosystem of decentralized networked computers was based on the formation of a large community that provided standardized interfaces with the help of open-source software (Raymond, 1999). Currently, new open-source LLMs are released every week on community websites such as Hugging Face (https://huggingface.co/), giving both users and developers a rapidly growing number of powerful tools. Science communication plays a central role in the critical thinking under consideration here. We cannot go into depth on this here, but what gives us confidence is that the Hugging Face research department also sees it this way. A recent TED talk on the real dangers of AI by the Artificial Intelligence Researcher and Climate Lead at Hugging Face, Sasha Luccioni (2023), has received more than one million views and represents a best practice example for the empowerment of critical thinking in the field of machine learning. In her closing statement, Luccioni emphasized that

> focusing on AI's future existential risks is a distraction from its current, very tangible impacts and the work we should be doing right now, or even yesterday, for reducing these impacts. Because yes, AI is moving quickly, but it's not a done deal. We're building the road as we walk it, and we can collectively decide what direction we want to go in together. (Luccioni, October 2023, loc. cit., 09:42 min.)

[7] Critical articles about AI systems are also quickly outdated. The first draft of the text was written more than one year ago.

5 Machine Learning and Critical Thinking

Machine learning is an incredibly powerful tool that can benefit fields from the natural sciences to the arts. However, it does not replace existing tools, instead complementing them. For example, critical minds must still be able to read to evaluate the texts produced by machine learning. In the present information age, data, algorithms, and AI are either hailed as a panacea for all problems of human coexistence or seen as the main culprits of a "digital immaturity." Both can be true at the same time. AI systems can, for example, present an opportunity for (data) science. However, they also increase the risk of undesirable dependence on AI companies with great market power (Fecher et al., 2023, p. 6). To be able to assess both potential and risk, we need a deep understanding of how data processing and algorithms work to enable a renewed exit from self-inflicted immaturity. We must not be afraid to dive into automated information processing and mathematics, even and especially as people unfamiliar with the subject. Modern information technology is certainly complex, but it is not particularly complicated. All that is required is the willingness to understand things, which will enable us to better assess the role that IT currently and, more importantly, should play in our society.

It is important to emphasize that AI has long had an influence on society, with the billions of dollars invested, demanded, or expected already having an impact on politics. Note that it is the promise of AI that has this influence, rather than AI itself, because the results are rather sobering, at least for experts who do not receive funding. These expectations and ascriptions of AI are structurally similar to the promises of digitalization in general (and political discourse around AI can be read as digitalization discourse without any semantic loss). For example, a recent study examined the digital divide in Germany and found that,

> The older people are, the less people feel that digitalization benefits them. A similar pattern can be seen in education. Here, the feeling of benefiting from digitalization decreases as the level of education falls, with only around one in three people with a low level of formal education convinced of the added value of digitalization, a figure almost twice as high for people with a high level of formal education. (Initiative D21 e.V., 2024)

Therefore, we need an understanding of how contemporary AI-based digitality works, both at a technical level and at a political and economic level. At a technical level, data literacy is an important building block for critical AI competences. In his work, Francis Bacon referred to the written observations of nature as "data," a term derived from the Latin word for "the given." These observations are recorded in a discrete form, which essentially means that the the continuum of our surrounding environment is written down in a finite alphabet. This discretization is the initial step in the process of translating data from the inexplicable world to a static medium. A fine observational grid

is built to parallel the observed environment, and corresponding values are noted, such as the number of songbirds in the garden. This is achieved using a built-in pointer, the literal or metaphorical finger. This process can be described as digitising the environment, as hinted at by the Latin word for finger, which is digitus. Over the course of the following four centuries, our observations have been extended to encompass phenomena that cannot be directly perceived by human senses. Very large things, objects far away from us, and very small matter and organisms in our immediate surroundings have become visible with the help of observation tools.

John Dewey defined critical thinking (he sometimes calls it also reflective thinking) as „active, persistent and careful consideration of any belief or supposed form of knowledge in the light of the grounds that support it, and the further conclusions to which it tends.“ (Dewey, 1910, p.6). Critical thinking in the data context means carefully distinguishing between observed phenomena and consciously deciding on an intellectual approach using observation tools to achieve a specific goal (Hitchcock, 2022). Critical thinking is very useful for not only science but also everyday life. For the production and application of modern IT systems, especially in the field of AI, the training of judgment is not only necessary for technical activity but also a moral imperative. This recalls the response Wittgenstein gives to his query about why human beings think before building a technical system: Because thinking has proven itself. Elaborating, he wrote,

> Why do humans think? What is it good for? Why do they make calculations for boilers and not leave the thickness of their walls to chance? After all, it is only an empirical fact that boilers calculated this way don't explode so often. But just as they will do anything rather than put their hands into the fire that once burned them, they will do anything rather than not calculate for a boiler. But as causes don't interest us, we can only say: Humans do[,] in fact[,] think – this, for instance, is the way they proceed when they build a boiler, and this procedure has proved its worth. (Wittgenstein, 2005, p. 179)

Wittgenstein died shortly before the start of research into AI; it would have been very interesting to know what he thought of the fact that we are now doing exactly what he found so absurd. His fundamental insight is that meaning and sense are created using language, precluding them from ever becoming fixed entities. Therefore, the post-structuralist notion of an “endless chain of signifiers” (Derrida, 1968) became commonplace in many scientific disciplines, from philosophy to social sciences to linguistics. Consequently, language has come to be understood as a social practice, with meaning and sense parts of a dynamic system and not fixed. What Yejin Choy calls the “dark matter of language” (in a talk labeling formal logic “overrated”) makes categorization problematic due to considerable ambiguity and nebulous boundaries.

Surprisingly, this rather old insight has often been ignored in logicist and reductionist approaches that attempt to formalize and quantify the use of natural language, approaches that persist in today's deep learning models, with Marcus (2017) writing, "Deep learning presumes a largely stable world, in ways that may be problematic" (p. 13). Even if the model's massive hunger for data is fed with the entire internet and additional sources, this structural problem cannot be solved within the current paradigm. If the world around us changes and the contexts of text change permanently, there is no way of "bug fixing" incorrect results. However, all alternatives seem equally disadvantageous: new initial training or permanent retraining of models will cause new "errors" and a renewed need for filters based on deep learning, which will then require validation by another model – it's "turtles all the way down." Compounding the problem, using crowd-sourcing to label training data has also turned out to be problematic because the inter-annotator agreement is highly questionable due to inherent disagreements in human textual inferences (Pavlik & Kwiatkowski, 2019).

Consequently, AI's problems "are not its ability to do things well but its ability to do things badly, and our reliance on it nevertheless" (Byrd, 2022). The unjustified trust caused by a lack of AI literacy goes hand in hand with a growing lack of trust caused by inappropriate use of AI. Luciano Floridi (2022) has emphasized that trust regulates actions in a society and is at the root of digital solutions:

> This is about trusting ourselves, each other, the future, human ingenuity and its products, and the potential goodness of their applications. Without this trust[,] there is only management of political power and a market of people's views, but not also a good policy and a market of ideas. (p. 48)

In contrast, the contemporary paradigm of AI/machine learning, including LLMs and their "hallucinations," continues to lack trustworthiness in terms of methodical transparency, reliable results, and even evaluation. As such, an essential element of AI literacy is "unmasking Clever Hans predictors and assessing what machines really learn" (Lapuschkin et al., 2019), because deep learning tends to produce correct results based on wrong premises that "[put] a question mark [on] the current broad and sometimes rather unreflected usage of ML in all application domains in industry and in the sciences." More recently, a study from Stanford also found "strong supporting evidence that emergent abilities may not be a fundamental property of scaling AI models" and instead are "creations of the researcher's analyses" (Schaeffer et al., 2023, p. 9). As Coeckelbergh has noted, "AI risks [undermining] trust in one's own epistemic capacities and hinder[ing] the exercise of those capacities. If we want to protect the knowledge basis of our democracies, we must address these problems in education and technology policy" (p. 1).

Approaches such as explainable AI are very useful for developers of AI systems and can also be used for AI didactics. In practice, however, it is not useful to disclose the internal calculations of an AI system. Upon examination of

complex AI systems, it becomes evident that the distance between two tokens or words can be calculated by representing them in higher-dimensional vectors. Nevertheless, the rationale behind an AI system classifying two words as similar remains hidden. A promising approach to address this concern involves programming a second AI system to provide explanations regarding the reasoning behind the initial AI system's classifications. This can be illustrated in image recognition, where regions of interest are highlighted in color. However, this does not fully satisfy the criteria of explanation. Instead, it is more akin to a plausible text that most people accept as such. Additionally, it should be noted that there is a distinction between cause and justification. The rationale behind recruiting decisions can differ completely from the reason given in the justification. This raises concerns about how the goal of establishing "appropriate trust" can be achieved if we only trust the explaining system when it is, in turn, explained.

The main problem of informational trust is that we obtain data through an instrument unbeknownst to us. To verify that data, we must use the instrument again. Whether we trust the instrument or not is a fundamentally epistemological question, and the argument cannot be dismissed easily.

6 Digital Enlightenment

A morally appropriate approach to AI requires digital enlightenment from both developers and users. The history of Artificial Intelligences (deliberately formulated in the plural) is a history of obfuscation and misconceptions. In some cases, transparency as a principle is certainly enough to defend against deliberate obfuscation, but only a deep understanding of how things work can defend against misconceptions.

The whole debate about the morally imperative use of AI reveals nothing less than our untrained approach to a professional occupation of the human condition. In the case of big data and AI, the four fundamental Kantian questions are closely intertwined. The moral question "What should I do?" is interwoven with the epistemological question "What can I know?," the anthropological question "What is the human being?," and ultimately the humanist question "What may I hope?" (Kant, 1923, p. 25) Therefore, critical AI literacy also requires basic training in technical ethics. Furthermore, it requires a critical understanding of science and higher education, which will doubtlessly be affected by LLMs, which have considerable transformative potential, especially in administrative, creative, and analytical tasks, although there are also risks related to bias, misinformation, and quality assurance that "need to be addressed through proactive regulation and science education" (Fecher et al., 2023).

To make things more complex and solutions-oriented, we would like to introduce a central actor: The enlightened public. Achieving critical AI competence is only possible with the help of the public use of reason, entirely in the spirit of Kant (1912). Even if there is justified criticism of this concept, we see no good or viable alternative in a value-pluralistic society. Because civil society does not have direct access to the findings of either science or philosophy, it needs media, such as science communicators, journals, and public intellectuals. In the field of natural sciences, outstanding media figures in Germany include Mai Thi Nguyen-Kim; for questions of social science, publicly funded institutes such as the Weizenbaum Institute for the Networked Society are established precisely to take on this mediating role. Elsewhere, large tech communities such as the Chaos Computer Club, the German Informatics Society, and the Forum of Computer Scientists for Peace and Social Responsibility have also been organizing large events for a long time now, not only for their small specialist community but also to include the broader civil society by hosting publicly accessible lectures that do not demand any prior technical training.[8]

However, this active part of civil society is dominated by the balance between politics and the economy, two vast domains that have no incentive to demystify the phenomenon of AI because both fears about AI and euphoric advocacy for AI lead to investment and attract votes. When billionaire company owners publicly call for an AI research moratorium[9], only to subsequently acquire an AI company, it demonstrates the dishonesty of the debate. The issue is not the technology itself, but the contested narrative surrounding AI and the usefulness of mystification.

One important way of achieving comprehensive critical AI literacy for the common good is demystifying AI. Anthropomorphism is a central problem in the current debate on the pros and cons of using AI, as Rehak (2021) notes:

> Unlike the abstract field of mathematics, where most technical terms are easily spotted as such, AI makes heavy use of anthropomorphisms. Considering [AI terms] such as “recognition,” “learning,” “acting,” “deciding,” “remembering,” “understanding” or even “intelligence” itself, problems clearly loom across all possible conversations[… U]sing deficient anthropomorphisms like “self-learning,” “autonomous[,]” or “intelligent” to describe the technical options of solving problems will lead to malicious decisions.

---

8 Disclosure: The authors are members of the Forum of Computer Scientists for Peace and Social Responsibility and the German Informatics Society.

9 Pause Giant AI Experiments: An Open Letter, March 22, 2023. https://futureoflife.org/open-letter/pause-giant-ai-experiments/

> Surely the best solution for this problem would be to completely change the terminology, but since large parts of the above mentioned are fixed scientific terms, a clean slate approach seems unrealistic. Therefore, at least in interdisciplinary work, science journalism activities or political hearings, a focus should be put on choosing the appropriate wording by scientists and (science) journalists. Only then policy and decision-makers have a chance to meaningfully grasp the consequences of their actions. (p. 89–98)

The normative power of the factual described by Rehak is favored by a second trend, namely that technical development is often perceived to progress far too fast, such that the Great Acceleration (Steffen et al., 2015) could be further accelerated to reach a speed beyond what society can handle. Critical thinking, ethical reflection, and political reaction take time; even in the case of an unresolved assumption of responsibility, fundamental decisions must be made that also take time. In addition, expertise in computer science and philosophy is necessary, which we fortunately do possess but which we cannot exercise if we are not given sufficient time to use our own intellect. As the timeline of recent developments in this paper shows, the socio-technical system of AI is constantly changing, and evaluative statements can quickly lose their empirical anchorage when the objectives change. Nevertheless, fundamental issues persist, especially regarding recent developments of multi-modal LLMs that create video output (Marcus, 2022). But the traditional mode of scientific communication through peer-reviewed papers, which might be echoed in scientific journalism, is clearly not fast enough to catch up. To enhance AI literacy, additional approaches to science communication are highly relevant for empowering civil society.

It has now been well established that training large language models requires a significant investment of time, often spanning several weeks or even months. Consequently, companies that wish to utilize their own AI must permanently dedicate high-performance computers to this process. As mentioned, this hidden dimension of resource consumption also pertains to the mediation task of AI literacy. It is evident that the consumption of natural and human resources can be morally and socially justified when AI is employed for the energy transition or in the circular economy. However, it is crucial to emphasize the added value of a particular AI system in such cases. Simple AI models, such as decision trees and even plain statistical methods, often yield comparable or even superior results with considerably less effort than deep neural networks featuring countless hidden layers. However, the prerequisites for a deliberation process using AI on a planetary scale are unevenly distributed among the discourse partners. The incentive to participate in the AI hype game for funding is also too great for critical and independent scientific voices to be heard too loudly.

Over two hundred years after Wolfgang Kempelen's Mechanical Chess Turk, IBM's chess computer Deep Blue won against the then reigning world cham-

pion Garry Kasparov. This meant, in principle, it was possible to build such a machine, and it is precisely stories like this that inspire the relentless endeavors of current AI researchers – because if it doesn't succeed today, it will in ten or a hundred years. Therefore, it is important to know what a machine cannot yet do, what it cannot do in principle, and what it should not do in the first place. This ethical dimension requires an enlightened debate leading to political action and changes to the current AI paradigm and usage in favor of digital maturity. This text aims to contribute to that change, building on Floridi's (2022) understanding that,

> Like a chess game, politics is constrained by the past, but it knows only the present, to be managed and negotiated [and, in some cases, criticized], and the future, to be designed and planned [and, in some cases, promised]. This is so because voters have no memory. Whatever politics delivered in the past, whether a problem or a solution, is taken for granted. The only past that is present in the voters' minds is unrelated to history and is part of a story-telling. So those who shape the narrative of the political past control its impact. (p. 59)

This holds especially true for AI narratives driven by strange and philosophically questionable ideologies. In their famous paper on "stochastic parrots," Bender et al. (2021, p. 619) "identified a wide variety of costs and risks associated with the rush for ever larger LMs" from environmental, financial, and opportunity costs to "the risk of substantial harms," leading them to call " on the field to recognize that applications that aim to believably mimic humans bring risk of extreme harms" combined, nevertheless, with the hope that "these considerations encourage [natural language processing] researchers to direct resources and effort into techniques for approaching NLP tasks that are effective without being endlessly data hungry." Unfortunately, over two years later, this hope has not been fulfilled. Instead, the author team recently concluded in their statement regarding a controversial "AI pause letter" (Gebru et al., 2023) that,

> We should be building machines that work for us, instead of "adapting" society to be machine[-]readable and writable. The current race towards ever larger "AI experiments" is not a preordained path where our only choice is how fast to run, but rather a set of decisions driven by the profit motive. The actions and choices of corporations must be shaped by regulation [that] protects the rights and interests of people. It is indeed time to act[,] but the focus of our concern should not be imaginary "powerful digital minds." Instead, we should focus on the very real and very present exploitative practices of the companies claiming to build them, who are rapidly centralizing power and increasing social inequities.

We completely agree with that position, which could hardly be expressed any better. The normative power of the fictional, embodied in discourses from

transhumanism and technological posthumanism to longterminism (Torres, 2021), has already misguided AI development for decades and continues to distort contemporary critique of it, insofar that aligns with the ethical and epistemological questionable premises of these collective delusions of the whole data industry. Therefore, a new vision, paradigm, and narrative of AI is needed to overcome problematic path dependencies with respect to sustainable AI used as an infrastructure for sustainability purposes and the public good. Fortunately, several ingredients already exist. Hugging Face and similar platforms are bringing hopes of initiatives such as AI4People (Floridi et al., 2018) within reach.

However, the free availability of pre-trained models and open data sets is comparable to the free availability of books in an open library: we must also be able to use the material offered. Beyond the pioneering work of the technically more affine individuals, it is essential for a new way of thinking about AI that can allow a sufficiently diverse and broad public to engage with the new tool. Furthermore, it is important, both from a security perspective (Willison, 2023) and for privacy reasons (O'Flaherty, 2023), that all internal prompts and internal sources be disclosed. Prompt attacks on GPT are on the rise, and the AI tools of OpenAI, Google, and Microsoft are vulnerable precisely because they do not disclose their internal workings. If it is not recognizable to the user whether a suggested result derives from dubious sources, was produced by non-transparent inference chains, or has simply been hallucinated, this is detrimental to the sustainable use of these tools.

The sustainability of such AI systems is fundamentally compromised, and because these systems are becoming deeply embedded, there is an urgent need to improve AI literacy to provide people from tech- and non-tech backgrounds with the critical reflexivity needed to ensure that these systems operate within planetary boundaries and do not lead to further social division. There are several examples of the ubiquity of AI systems that seem quite harmless in themselves. For example, in Germany, an entire lifestyle magazine was produced by AI without readers knowing.[10] Third-party funding proposals on AI are being written using LLMs as we speak, as are the final reports.[11] The winner of the Sony World Photography Awards 2023 won with an AI-generated image (Williams, 2023). Finally, teachers are increasingly complaining that students are having their essays and even class tests written by unreliable and non-transparent AI systems. As with plagiarism, their unreflective use threatens the foundations of all areas that require trust, from science to politics to journalism to eventually even include democracy itself (Coeckelbergh, 2022).

---

[10] It concerned a recipe magazine from the Burda publishing house published earlier this year, which was sold for 2.99 euros, and which did not specifically mark that it was "created with the help of ChatGPT and Midjourney," according to the publisher. For more context, see the statement by the Bavarian Journalists' Association: https://www.bjv.de/pressemitteilungen/detail/keine-experimente-mit-der-glaubwuerdigkeit/

[11] That's a lie, of course – there will always be enough underpaid grad students and overworked post-docs at hand to write these reports.

In addition to science, the military, and the economy, it is the active civil society that must promote AI literacy for the benefit of all. Supporting the recent rapid developments of open-source LLMs, an alliance of top-class researchers and open-source AI associations published RedPajama, an open-source decentralized AI system with an open dataset containing billions of tokens (Hahn, 2023). The free and open-source software community has long hoped that the community could not only produce products for the common good and solve interesting problems (Raymond, 2010) but also increase the world's knowledge and make it accessible to all. Similarly, Stewart Brand's Whole Earth Catalogue (1971) initiated a community that wanted to use and produce tools together while fully understanding the implications of the use of those tools. As Brand later wrote, "We are as gods and might as well get good at it." In essence, knowledge, collaboration, and digital commons are key elements of digital commons-based visions of AI.

In the preliminary recommendations of its recent interim report (United Nations, 2023), the UN High-Level Advisory Body on Artificial Intelligence published five guiding principles: AI 1) "should be governed inclusively, by and for the benefit of all"; 2) AI "must be governed in the public interest"; 3) AI governance "should be built in step with data governance and the promotion of data commons"; 4) AI "must be universal, networked and rooted in adaptive multi-stakeholder collaboration"; 5) AI "should be anchored in the UN Charter, International Human Rights Law, and other agreed international commitments such as the Sustainable Development Goals" (pp. 15–17). After a consultation phase, the final report of Summer 2024 will contribute to the UN Global Digital Compact (United Nations, 2021). German Digital Civil Society Organizations has already emphasized in a related position paper (De Bastion et al., 2023) that digitalization is much more than the isolated view of AI, and a just and inclusive digital transformation is based on open infrastructure, codes, and standards. This demands that a global digital commons become a goal vision of the Global Digital Compact. As, according to their website, "a multi-stakeholder UN-endorsed initiative that facilitates the discovery and deployment of open-source technologies, bringing together countries and organizations to create a thriving global ecosystem for digital public goods and helping to achieve the sustainable development goals," the Digital Public Goods Alliance (2019) recently published a 5 Year Strategy for "unlocking the potential of open-source technologies for a more equitable world" (Nordhaug & Harris, 2023). More specific to AI Ethics and the forthcoming implementation of the EU AI Act, the AI4People initiative – which was launched six years ago as a research and policy project – is going to publish the whitepaper "Towards an Ethical Impact Assessment for AI" in spring 2024, detailing the following key objectives:

1) Ethical Frameworks: Developing ethical guidelines and frameworks to guide the development and application of AI, emphasizing fairness, transparency, accountability, and inclusivity.

2) Public Awareness and Education: Raising awareness among the public regarding AI's potential and challenges, fostering understanding, and promoting informed discussions on its societal impact.

3) Policy Recommendations: Providing recommendations to policymakers and regulatory bodies to develop appropriate laws and regulations that govern AI usage, ensuring it aligns with societal values and objectives.

4) Inclusivity and Diversity: Encouraging diverse perspectives and inclusivity in AI development to minimize bias and enhance AI systems' understanding and representation of all individuals.

5) Privacy and Data Protection: Advocating for robust data privacy measures and emphasizing the importance of securing and protecting individuals' data in AI applications.

6) Collaboration and Knowledge Sharing: Facilitating collaboration among stakeholders [and] sharing best practices, research findings, and insights to foster a collective effort towards responsible AI. (Floridi, L., Bonvicini, M., & Blair, T. (2018).)

Ultimately, there is no lack of visions, but the possibility, degree, and speed of their realization will depend on politics, governance, and critical AI literacy. After 60 years, there seems to be a consensus in computer science circles that part of the technician's responsibility is to educate and teach an understanding of how technology works as a socio-technical system. The power to shape not only technology but also societies is well known to computer science, as demonstrated by the Ethical Guidelines of the German Informatics Society (2018) and IEEE standards (2021). The need for ethical IT innovation (Spiekermann, 2015) is not only recognized but is now part of IT education. Nevertheless, it has yet to reach the mainstream, with even an educated audience continuing to fall for false attributions. Joseph Weizenbaum's (1976, p. 7) statement about ELIZA remains true today, if we just replace ELIZA with, for example, ChatGPT: "This reaction to [ChatGPT] showed me more vividly than anything I had seen hitherto the enormously exaggerated attributions an even well-educated audience is capable of making, even strives to make, to a technology it does not understand." Providing the robust understanding needed for a mature information society, critical AI literacy will be a necessary condition for reshaping the current dynamics and future AI infrastructure into a sustainable model in a two-fold way. This means not only understanding what is wrong with AI but also doing better, recognizing that "the existence of information alone does not necessarily impact the outcome of a situation" (Falk & van Wynsberghe, 2023).

References

Albrecht, S. (2023). ChatGPT und andere Computermodelle zur Sprachverarbeitung – Grundlagen, Anwendungspotenziale und mögliche Auswirkungen (TAB-Hintergrundpapier, Vol. 26). Büro für Technikfolgen-Abschätzung beim Deutschen Bundestag (TAB). https://doi.org/10.5445/IR/1000158070

Angwin, J. (2023, January 28). Decoding the hype about AI. *The Markup*. https://themarkup.org/hello-world/2023/01/28/decoding-the-hype-about-ai

Baack, S. (2024). Training data for the price of a sandwich. Common Crawl's impact on generative AI. *Mozilla Insights*, https://assets.mofoprod.net/network/documents/2024CommonCrawlMozillaFoundation.pdf.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency,* 610–623. https://doi.org/10.1145/3442188.3445922

Brand, S. (1969). Epigraph. *The Whole Earth Catalogue.* https://archive.org/details/sim_whole-earth-catalog_whole-earth-catalog_spring-1969/page/n5/mode/2up

Brand, S. (Ed.). (1971). Access to tools. *The Whole Earth Catalogue.* https://archive.org/details/B-001-013-719

Byrd, C. (2022, December 4). Cory Doctorow wants you to know what computers can and can't do. *The New Yorker.* https://www.newyorker.com/culture/the-new-yorker-interview/cory-doctorow-wants-you-to-know-what-computers-can-and-cant-do

Chomsky, N., Roberts I., & Watumull, J. (2023, March 8). The false promise of ChatGPT. *The New York Times*. https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html

Chowdhery, A. (2022). PaLM: Scaling Language Modeling with Pathways. *arXiv*. http://arxiv.org/abs/2204.02311

Choy, Y. (2022, May 24). 2082: An ACL odyssey: The dark matter of language and intelligence [Video]. YouTube. https://www.youtube.com/watch?v=lLCEy2mu4Js

Coeckelbergh, M. (2023). Democracy, epistemic agency, and AI: Political epistemology in times of artificial intelligence. *AI and Ethics, 3*(4), 1341–1350. https://doi.org/10.1007/s43681-022-00239-4

Crawford, K., & Joler, V. (2018). *Anatomy of an AI system.* http://www.anatomyof.ai

Datenethikkommission der Bundesregierung, Bundesministerium des Innern, für Bau und Heimat, & Bundesministerium der Justiz und für Verbraucherschutz (Eds.). (2019). *Gutachten der Datenethikkommission der Bundesregierung.* https://www.bmi.bund.de/SharedDocs/downloads/DE/publikationen/themen/it-digitalpolitik/gutachten-datenethikkommission.pdf;jsessionid=4E90673A2646A724E61F20E914634DB0.2_cid360?__blob=publicationFile&v=7

De Bastion, G., Dorsch, M., & von Franqué, F. (2023, September 13). Position on the global digital compact of German digital civil society organizations [Policy statement]. Open Knowledge Foundation Deutschland. https://okfn.de/en/publikationen/2023_globaldigitalcompact/

Derrida, J. (1968). La 'différance'. *Bulletin de la Société Française de Philosophie, 62*(3), 73.

Devansh, D. (2023, April 28). Why ChatGPT lies. *Geek Culture.* https://medium.com/geekculture/why-chatgpt-lies-4d4e0c6e864e

Dewey, J. (1910). How We Think, Boston: D.C. Heath.

Digital Public Goods Alliance. (2019). *Who we are*. https://digitalpublicgoods.net/who-we-are/

Dzieza, J. (2023). AI is a lot of work. *The Verge*. https://www.theverge.com/features/23764584/ai-artificial-intelligence-data-notation-labor-scale-surge-remotasks-openai-chatbots

Ethical Guidelines of the German Informatics Society Bonn, June 29, 2018, https://gi.de/ethicalguidelines/

Falk, S., & van Wynsberghe, A. (2023). Challenging AI for sustainability: What ought it mean? *AI & Ethics.* https://link.springer.com/content/pdf/10.1007/s43681-023-00323-3.pdf

Fecher, B., Hebing, M., Laufer, M., Pohle, J., & Sofsky, F. (2023). Friend or foe? Exploring the implications of large language models on the science system. *AI & Society*. https://doi.org/10.1007/s00146-023-01791-1

Floridi, L. (2020). The green and the blue: A new political ontology for a mature information society. *Philosophisches Jahrbuch (Freiburg), 127*(2), 307–338. https://doi.org/10.5771/0031-8183-2020-2-307

Floridi, L. et al. (2018). AI4People – An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines, 28*(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5

Floridi, L., Bonvicini, M., & Blair, T. (2018). *AI4People*. https://eismd.eu/ai4people/

Gebru, T., Bender, E., McMillan-Major, A., & Mitchell, M. (2023, March 31). Statement from the listed authors of Stochastic Parrots on the "AI pause" letter. *Dair Institute*. https://www.dair-institute.org/blog/letter-statement-March2023/

Gesellschaft für Informatik. (2016). Dagstuhl-Erklärung. Bildung in der digitalen vernetzten Welt. https://dagstuhl.gi.de/fileadmin/GI/Hauptseite/Aktuelles/Projekte/Dagstuhl/Dagstuhl-Erklaerung_2016-03-23.pdf (own translation).

Hahn, S. (2023, April 19). LLaMA clone: RedPajama – First open-source decentralized AI with open dataset. *Developer*. https://www.heise.de/news/LLaMA-replica-RedPajama-first-open-source-decentralized-AI-with-open-dataset-8972104.html

Hielscher, M. (2023). Soekia GPT. https://www.soekia.ch/gpt.html

Hitchcock, D. (2022). Critical Thinking. In E. N. Zalta & U. Nodelman (Eds.), The Stanford Encyclopedia of Philosophy (Winter 2022). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2022/entries/critical-thinking/

IEEE Standard Model Process for Addressing Ethical Concerns during System Design Developed by the Systems and Software Engineering Standards Committee of the IEEE Computer Society Approved 16 June 2021 IEEE SA Standards Board IEEE Std 7000TM-2021

Inflection AI, Inc. (2023). Pi, your personal AI. https://pi.ai/talk

Initiative D21 e.V. (2024): D21-Digital-Index 2023/24. Jährliches Lagebild zur Digitalen Gesellschaft, https://initiatived21.de/uploads/03_Studien-Publikationen/D21-Digital-Index/2023-24/d21digitalindex_2023-2024.pdf, p.25 (own translation).

Johnson, K. (2022, June 14). LaMDA and the sentient AI trap. *Wired*. https://www.wired.com/story/lamda-sentient-ai-bias-google-blake-lemoine/

Kaack et al. (2022). Aligning artificial intelligence with climate change mitigation. *Nature*. https://www.nature.com/articles/s41558-022-01377-7

Kaack, L. et al. (2022). Aligning artificial intelligence with climate change mitigation. *Nature Climate Change, 12,* 518–527. https://doi.org/10.1038/s41558-022-01377-7

Kant, I. (1912). Abhandlungen nach 1781. Akademie Ausgabe Band VIII.

Kant, I. (1923). Logik, Physische Geographie, Pädagogik. Akademie Ausgabe Band IX.

Kneese, T. (2023) Climate Justice & Labor. *AI Now Institute*. https://ainowinstitute.org/wp-content/uploads/2023/08/AINow-Climate-Justice-Labor-Report.pdf

Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., & Müller, K.-R. (2019). Unmasking Clever Hans predictors and assessing what machines really learn. *Nature Communications, 10*(1), Article 1. https://doi.org/10.1038/s41467-019-08987-4

Luccioni, S. (2023). *AI is dangerous, but not for all the reasons you think* [Video]. YouTube. https://www.youtube.com/watch?v=eXdVDhOGqoE

Luccioni, A. S., Viguier, S., & Ligozat, A.-L. (2022). Estimating the carbon footprint of BLOOM, a 176B parameter language model. *arXiv*. http://arxiv.org/abs/2211.02001

Marcus, G. (2018). Deep learning: A critical appraisal. *arXiv*. https://arxiv.org/pdf/1801.00631.pdf

Marcus, G. (2022, November 16). A few words about bullshit [Substack newsletter]. *Marcus on AI.* https://garymarcus.substack.com/p/a-few-words-about-bullshit

Marcus, G. (2023, February 11). Inside the heart of ChatGPT's darkness [Substack newsletter]. *Marcus on AI*. https://garymarcus.substack.com/p/inside-the-heart-of-chatgpts-darkness

Marcus, G. (2024): Sora can´t handle the truth [SubStack newsletter]. *Marcus on AI.* https://garymarcus.substack.com/p/sora-cant-handle-the-truth

Marcus, G. & Davis, E. (2019). *Rebooting AI. Building artificial intelligence we can trust*. Pantheon.

McCarthy, J., Minski, M., Rochester, N., & Shannon, C. (1955). A proposal for the Dartmouth summer research project on artificial intelligence. *Stanford*. http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html?dom=pscau&src=syn

Mitchell, T. M. (1997). *Machine learning*. McGraw-Hill.

Mühlhoff, R. (2023). Die Macht der Daten: *Warum künstliche Intelligenz eine Frage der Ethik ist*. V&R Unipress.

Nordhaug, L., & Harris, L. (2023). 5 year strategy. *Digital Public Goods Alliance.* https://digitalpublicgoods.net/dpga-strategy2023-2028.pdf

O'Flaherty, K. (2023, April 9). Cybercrime: Be careful what you tell your chatbot helper … *The Guardian.* https://www.theguardian.com/technology/2023/apr/09/cybercrime-chatbot-privacy-security-helper-chatgpt-google-bard-microsoft-bing-chat

Papert, S., & Harel, I. (1991). Constructionism. Ablex Publishing Corporation.

Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L-M.,

Rothchild, D., So, D., Texier, M., & Dean, J. (2021). Carbon emissions and large neural network training. *arXiv*. https://arxiv.org/ftp/arxiv/papers/2104/2104.10350.pdf

Pasquinelli, M., & Joler, V (2020): The Nooscope manifested: AI as instrument of knowledge extractivism. *Ai & Society, 36,* 1263–1280. https://doi.org/10.1007/s00146-020-01097-6

Patel, D., & Afzal, A. (2023, May 4). Google "We have no moat, and neither does OpenAI." *SemiAnalysis*. https://www.semianalysis.com/p/google-we-have-no-moat-and-neither

Patterson, D. et al. (2021). Carbon emissions and large neural network training. *arXiv*. https://doi.org/10.48550/arXiv.2104.10350

Pavlick, E., & Kwiatkowski, T. (2019). Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics, 7*, 677–694. https://doi.org/10.1162/tacl_a_00293

Perrigo, B. (2023, January 18). OpenAI used Kenyan workers on less than $2 per hour. *Time*. https://time.com/6247678/openai-chatgpt-kenya-workers/

Piaget, J. (1966). *The psychology of intelligence.* Adams & Co.

Raymond, E. (2010, February 18). The Cathedral and the Bazaar. http://www.catb.org/~esr/writings/cathedral-bazaar/

Rehak, R. (2021). *The language labyrinth: Constructive critique on the terminology used in the AI discourse*. University of Westminster Press.

Robbins, S., & van Wynsberghe, A. (2022). Our new artificial intelligence infrastructure: Becoming locked into an unsustainable future. *Sustainability*, 14(8), Article 8. https://doi.org/10.3390/su14084829

Schaeffer, R., Miranda, B., & Koyejo, S. (2023). Are emergent abilities of large language models a mirage? *arXiv*. https://doi.org/10.48550/arXiv.2304.15004

Schanner, G., & Rock, J. (2023, April 15). How to get by with AI: AI Tooling and how it's changing everything we do. https://media.ccc.de/v/glt23-395-how-to-get-by-with-ai-ai-tooling-and-how-it-s-changing-everything-we-do-

Schiffer, Z., & Newton, C. (2023, March 14). Microsoft lays off team that taught employees how to make AI tools responsibly. *The Verge*. https://www.theverge.com/2023/3/13/23638823/microsoft-ethics-society-team-responsible-ai-layoffs

Schyns, C. (2023, February 23). The lobbying ghost in the machine. *Corporate Europe Observatory*. https://corporateeurope.org/en/2023/02/lobbying-ghost-machine

Simonite, T. (2021, June 8). What really happened when Google ousted Timnit Gebru. *Wired*. https://www.wired.com/story/google-timnit-gebru-ai-what-really-happened/

Spiekermann, S. (2016). *Ethical IT innovation: A value-based system design approach* (1st edition). CRC Press.

Steffen, W., Broadgate, W., Deutsch, L., Gaffney, O., & Ludwig, C. (2015). The trajectory of the Anthropocene: The great acceleration. *The Anthropocene Review, 2*(1), 81–98. https://doi.org/10.1177/2053019614564785

Torres, É. (2021, October 19). Why longtermism is the world's most dangerous secular credo. *Aeon*. https://aeon.co/essays/why-longtermism-is-the-worlds-most-dangerous-secular-credo

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). LLaMA: Open and efficient foundation language models. *arXiv.* https://doi.org/10.48550/arXiv.2302.13971

Ullrich, A. (2022): Opportunities and challenges of big data and predictive analytics for achieving the UN's SDGs. In *PACIS 2022 Proceedings* (p. 279).

UNESCO. (2023). *AI competency frameworks for school students and teachers.* https://www.unesco.org/en/digital-education/ai-future-learning/competency-frameworks

United Nations. (2021). *Our common agenda – Report of the Secretary-General.* https://www.un.org/techenvoy/global-digital-compact

United Nations. (2023). *AI Advisory Body interim report: Governing AI for humanity.* https://www.un.org/sites/un2.un.org/files/ai_advisory_body_interim_report.pdf

van Wynsberghe, A. (2021). Sustainable AI: AI for sustainability and the sustainability of AI. *AI and Ethics, 1*(3), 213–218. https://doi.org/10.1007/s43681-021-00043-6

Vipra, J., & Myers West, S. (2023) Computational power and AI. *AI Now Institute.* https://ainowinstitute.org/wp-content/uploads/2023/09/AI-Now_Computational-Power-an-AI.pdf

WBGU [German Advisory Council on Global Change]. (2019). *Towards our common digital future.* https://www.wbgu.de/fileadmin/user_upload/wbgu/publikationen/hauptgutachten/hg2019/pdf/wbgu_hg2019_en.pdf

Weizenbaum, J. (1967). Contextual understanding by computers. *Communications of the ACM, 10*(8), 474–480.

Weizenbaum, J. (1976). *Computer power and human reason: From judgment to calculation.* W. H. Freeman.

Williams, Z. (2023, April 18). 'AI isn't a threat' – Boris Eldagsen, whose fake photo duped the Sony judges, hits back. *The Guardian.* https://www.theguardian.com/artanddesign/2023/apr/18/ai-threat-boris-eldagsen-fake-photo-duped-sony-judges-hits-back

Willison, S. (2023, April 14). *Prompt injection: What's the worst that can happen?* https://simonwillison.net/2023/Apr/14/worst-that-can-happen/archived https://archive.is/FX8BA

Wittgenstein, L. (2005). The Big Typescript (German-English scholar's edition). Blackwell Publishing.

World Economic Forum (2023, January 16). *The 'AI divide' between the Global North and the Global South.* https://www.weforum.org/agenda/2023/01/davos23-ai-divide-global-north-global-south/

Wu, C. J. (2022). Sustainable AI: Environmental implications. Challenges and opportunities. *arXiv.* https://arxiv.org/pdf/2111.00364.pdf

xAI Corp. (2023, November 4). Announcing Grok. https://archive.ph/8ikfW

Date received: December 2023
Date accepted: April 2024