

WEIZENBAUM JOURNAL OF THE DIGITAL SOCIETY
Volume 3 \ Issue 3 \ w3.3.3 \ XX-XX-2023
ISSN 2748-5625 \ DOI 10.34669/WI.WJDS/3.3.3

Information on this journal and its funding can be found on its website:

<https://wjds.weizenbaum-institut.de>

This work is available open access and is licensed under Creative Commons Attribution 4.0 (CC BY 4.0):

<https://creativecommons.org/licenses/by/4.0/>

KEYWORDS

chatbots
humanism
user-centered design
artificial intelligence
critique

INTERVIEW

From Joseph Weizenbaum to ChatGPT

Critical Encounters with Dazzling AI Technology

Christiane Floyd

Prof. em., University of Hamburg

christiane.floyd@t-online.de

ABSTRACT

The paper considers AI systems from a use perspective. It focuses on conversational chatbots, starting from Weizenbaum's ELIZA and sketching the major scientific advances leading up to ChatGPT. The main discussion builds upon several experiment-reflection cycles conducted by the author to explore ChatGPT as a knowledge resource. The analysis considers ChatGPT responses in terms of accuracy, structure, context, perspective, and bias. The critical evaluation begins with the observation that ChatGPT produces a mixture of clear and precise results and arbitrary misinformation without ever clarifying its own scope. This leads to the identification of the system's key problem, namely, how it contends with truth, which involves replacing the idea of truth with a probabilistic surrogate based on textual correlation. In responsible use, a system like ChatGPT must be embedded in a human learning culture. A framework for this process should include an insistence on truthfulness, an impulse towards enhancing human competence, and strengthened responsibility structures within communities.

[In the field of] artificial intelligence [...] machines are made to behave in wondrous ways, often sufficient to dazzle even the most experienced observer. But once a particular program is unmasked, once its inner workings are explained [...] its magic crumbles away; it stands revealed as a mere collection of procedures [...]. The observer says to himself "I could have written that." With that thought, he moves the program in question from the shelf marked "intelligent" to that reserved for curious. (Weizenbaum 1966)

As an AI language model, I do not possess knowledge in the same way that humans do, nor do I have the capacity for self-reflection or introspection. (ChatGPT, April 3, 2023)

1 Introduction

On January 11, 2023, I had the privilege of giving the introductory talk to the Weizenbaum jubilee year W\100 at the Weizenbaum Institute. Nobody present could have foreseen the astonishing hype that artificial intelligence (AI) was to enjoy only a few weeks later, following the public launch of [ChatGPT](#).

As a *chatbot*, ChatGPT is a direct descendant of ELIZA, the seminal program created by Joseph Weizenbaum that first enabled so-called "conversations" between human users and machines via an electronic typewriter and could be tailored (by different scripts) to adopt various conversational roles (Weizenbaum, 1966). The DOCTOR version, with a script that made it resemble a Rogerian psychotherapist, stirred up wide use and discussion. The experience with DOCTOR turned Weizenbaum from a passionate AI researcher to one of its most renowned critics.

What shocked Weizenbaum was that (1) practicing psychotherapists believed that DOCTOR could be enhanced to automate psychotherapy, (2) users interacted with DOCTOR as if it were a human being, revealing their secrets and attributing human qualities to the technical system, and (3) many people thought that the program could be generalized to provide a basis for understanding natural language as a whole.

In the 1966 publication on ELIZA quoted at the beginning of this paper, Weizenbaum chose the word "dazzle" to describe the effect that DOCTOR had on many people and encouraged them to take a closer look at its "inner workings." Then they would see that the "magic crumbles away" and experienced observers would realize that they could have written the program themselves.

However, there is a vast technological gap between the forebear of all chatbots and modern AI language models, as ChatGPT exemplifies. We have more than enough reasons to be “dazzled,” so much so that many people feel shaken in their identity as humans and terrified of the effects of the new technology. I have met quite a few people who had just memorized the name “Artificial Intelligence” because it has suddenly become omnipresent. In contrast to ELIZA, it is not at all obvious how we can come to understand the “inner workings” of such systems well enough to make the “magic crumble away.”

Not even “the most experienced observer” could venture to write such a program alone. Through its sophisticated inner architecture, its intricate connection with other technologies, and its dependence on its environment – especially on a text corpus used for training – the system is hugely complex. Furthermore, its behavior changes over time depending on the text data it has been exposed to. The *deep learning* apparatus embodied by ChatGPT makes it hard to understand the system’s mechanics, to subject it to control, and to ensure that it can be put to human use without causing harm. In fact, a few months after making it available to the public, the system’s makers urgently demanded its regulation.

I am not an AI specialist, but I had the privilege of becoming acquainted with AI already in 1968 as a research associate at the Stanford Artificial Intelligence Laboratory, where I witnessed an amazing collection of machines that were “made to behave in wondrous ways,” many of them prototypes of AI-based technologies that have since reached maturity and can be observed in use in various contexts every day. One of these systems was DOCTOR, whose effects on trusting users I have observed.

Very early, I have been exposed to grand claims made by certain leading AI specialists equating humans and machines. I remember one of them provoking the audience during a lecture. “Of course, people are like machines,” he said. “Anyone feel threatened here?” The element of threat made those words unforgettable. Most of my AI friends stayed away from such claims and some stood up against them, most prominently Joseph Weizenbaum, with whom I enjoyed a lasting friendship after we met at Stanford in the early 1970s. His seminal book, *Computer Power and Human Reason* (1976) set the standard for discussion on the topic. Taking a stand against equating humans and machines also became a pillar of my own professional work (e.g., Floyd, 1986), although I am full of admiration for the magnificent technological achievements of AI.

This paper considers AI systems from the *use perspective*. The AI community does not like to talk about the use of AI systems, preferring to discuss impact, adopting a techno-centered perspective, as if the technology could act on its own. To my dismay, I found that even the [Center for Human-Centered Artificial Intelligence at Stanford University](#) complies with this way of thinking. For them, “human-centered” means promoting research on how to enhance the symbiosis of humans and AI based on theories that explore their similarities and differences.

My approach is human-centered in a different sense. I want to consider how AI systems can be suitable as tools or media for human use and how they can be understood in human terms and subjected to human purpose in accordance with human values. This means giving special attention to Weizenbaum’s arguments and drawing on the ongoing discussion on Digital Humanism coming out of Vienna (Werthner, 2019).

I focus on ChatGPT, restricting myself to viewing it as a knowledge resource. The paper is the result of several experiment-reflection cycles that have helped me gauge the strengths and weaknesses of the system and develop ways of talking about them. I came to know ChatGPT’s dazzling confidence when responding to a seemingly unlimited diversity of prompts and its ability to express its responses in remarkably clear language, regardless of whether those responses are accurate or absurd.

In what follows, I first retrace the technological development leading to ChatGPT to help newcomers appreciate the nature of deep learning-based AI language models. Then, I reflect on some of my experiments with the system to explore “its inner workings,” get a feeling for what it can do, and identify its limits. Building on these foundations, I describe my view of the greatest challenge posed by such systems: a radically new way of approaching truth – or rather, abolishing truth – which I call truth by correlation. Finally, I point to ways that humans can think and act to contend with this new challenge.

2 The Long Road from ELIZA to ChatGPT

The ELIZA system was the first program to demonstrate an understanding of language that enabled conversations between humans and machines to exhibit human-like behavior. However, the term conversation is misleading: While it normally denotes a free-flowing oral or written exchange between humans in embodied presence or via communication media, here it refers to a sequence of human prompts and system responses.

For Weizenbaum, systems such as ELIZA and ChatGPT seem magical only to those who don't understand their inner workings. This was the case for ELIZA, which was self-contained and rule-based. Although its behavior was hard to predict due to the intricate interplay of logical rules, it was repeatable. The results for one and the same input would remain stable over time. Specialists were able to explain its inner workings in a satisfactory manner.

However, modern chatbots are based on neural networks and deep learning. In their training phases, they are exposed to a huge text corpus that gets absorbed and used as a basis for responses. How they produce results for a given input cannot be explained in terms of rules. Instead, they are obtained via the statistical analysis of textual patterns and correlations. Furthermore, they evolve over time, depending on the specific text corpus used for training and on feedback obtained through the sequence of queries. While their mechanics are known, their results for a given input are unpredictable to “even the most experienced observer.”

Comparing ELIZA to ChatGPT is like comparing the Danube rivulet in the Black Forest with the mighty river in Serbia. The difference cannot be explained in terms of the original Danube alone. Instead, the vastness and power of the latter is enabled by contributions from numerous geographically disparate sources. To understand the Danube river system as a whole, it is critical to consider large swathes of Central and Eastern Europe. The following sections attempt to do the same for ChatGPT.

2.1 Technologies Around AI

The journey from ELIZA to ChatGPT took about 57 years, starting from stand-alone systems dependent on the hardware and operating system of local (unnetworked) computers. To appreciate the huge technological advance, we must consider the progress across several scientific disciplines during these decades, including theoretical foundations established, technological outputs, and the powerful interaction between these disciplines.

Language studies in connection with computers started early. In computing jargon, the languages we speak and write are called natural languages. Computer linguistics emerged as a new field to study natural languages around 1960 and encouraged rich cooperation between computing departments and the fields of linguistics, literature, and media studies. According to Søgaard (2022), ChatGPT is not a scientific contribution but an engineering feat. In this context, its language skills, derived from advances in the disciplines of computer-oriented language study, really are admirable.

Computer technology has advanced considerably in recent decades. While the mode of operating has remained the same, computing power has increased by several orders of magnitude. This increase in computing power has enabled the transition of AI from the original rule-based model to neural networks, the technological basis for ChatGPT.

Human-computer interaction at the hardware and software levels was investigated and revolutionized in the 1980s, leading to the interaction modes and styles that we take for granted today and producing precursors to modern chatbots.

Network technology began development in the 1960s, originally to connect computers for military purposes. The revolution leading to the new paradigm that understood that networks should connect people via computers started in the 1970s. Extensive research, technical innovation and standardization led to the establishment of the Internet with new online infrastructures for work and communication.

The World Wide Web became publicly available around 1990, sparking profound changes in human knowledge culture by enabling the universal availability of knowledge artifacts via digitalization. This led to the global publication of knowledge artifacts online and consequent new forms of building up and sharing human knowledge.

Soon after, data science came into being as a specific perspective in computing with a focus on storing, organizing, and retrieving huge databases in the form of texts, numbers, and images. Automatic ways of searching, sorting, interpreting, and profiling were developed, leading to the idea that a new form of knowledge could be derived from data analysis.

Meanwhile, statistical modeling and learning algorithms were developed to inform data science and analysis in probabilistic terms. The idea was to find and compare patterns in related texts or images to be able to predict their connection with a very high probability of accuracy. In its essence, the approach underlying the ChatGPT architecture is probabilistic.

This list may be incomplete, but it is, nonetheless, impressive. Seven disciplines, each with its own elaborate research agenda, innovative technologies, and widespread impact, must be considered to begin to understand the tremendous advances in AI since 1966.

2.2 Paradigm Shifts in AI

It is against this background and in continuous interaction with advances in these various computing disciplines that we need to appreciate the changes in AI itself, always remembering that this is an inherently interdisciplinary research field strongly interwoven with (especially) the cognitive sciences.

Perhaps the most important change in AI was the move from the symbolic approach – which considered thinking and knowing in terms of explicit logical rules – to the neural network approach, which attempts to emulate the composition of and connections between neurons in the brain. Notably, while in the symbolic stage all assertions of AI could be expressed in terms of logic, the neural network approach is sub-symbolic, relying entirely on the mathematical properties and the current status of neurons and synapses.

This particularly applies to machine learning, which is built on the idea that AI systems can surpass the knowledge imbued to them by their makers via experience, that is, exposure to sample data and feedback.

In *rule-based learning*, a system can expand its knowledge base by applying the rules it has been supplied with. There may also be a mechanism for forming new rules. The alternative, *adaptive learning* using neural networks, sees the status of neurons and synapses set to initial values. Then, systems are exposed to sample data in a training phase and adapt their status according to what they find in the data. Within this mode, *supervised learning* involves feedback from a human trainer, with *unsupervised learning* seeing “algorithms learn patterns exclusively from unlabeled data” (Wikipedia). The quality of adaptive machine learning depends on the network architecture, including the number of neurons and their arrangement and the number of layers traversed from the original input to the learned output.

2.3 The Transformer Architecture and Deep Learning

I must be honest here and confess that I am not qualified to write the coming section. I have investigated some of the original literature, and I am happy to admire the work of my colleagues, but I am not sure that I can give an introduction to what they are doing in my own words, even if I do consider myself an intelligent observer.

The subject of deep learning is treated comprehensively in the work of Goodfellow et al. (2016), and I would advise interested readers to consult that text. However, to introduce the concept, let me quote the opening paragraph from a review of that book by Kim (2016):

Deep learning is a form of machine learning that enables computers to learn from experience and understand the world in terms of a hierarchy of concepts. Because the computer gathers knowledge from experience, there is no need for a human computer operator formally to specify all of the knowledge needed by the computer. The hierarchy of concepts allows the computer to learn complicated concepts by building them out of simpler ones; a graph of these hierarchies would be many layers deep [...] Deep learning has already proven useful in many software disciplines, including computer vision, speech and audio processing, natural language processing, robotics, bioinformatics and chemistry, video games, search engines, online advertising and finance (Kim 2016, p. 351).

That is, the mechanisms responsible for deep learning draw on several mathematical disciplines – in particular, statistics and linear algebra – to form suitable learning algorithms that can be combined into a system’s learning strategy.

The Transformer Architecture underlying ChatGPT was introduced by Vaswani et al. (2017, p. 5999). They write in their introduction “In this work [,] we propose the Transformer, a model architecture [...] relying entirely on an attention mechanism to draw global dependencies between input and output.”

The Transformer architecture comprises two components: The *encoder* maps the input sequence to an internal vector representation of morphemes. The *decoder* transforms this vector into the symbols of the output sequence. Both components feature six identical layers. To compute the output from the input sequence, the transformer uses a mechanism called self-attention, which examines the vectors point by point and incorporates intermediate results into the input for further examination. This is sufficient as a conceptual framework for this paper’s discussion.

Turning to the user's perspective, let me make several comments. If I have difficulties learning a system's mode of operation, others will have those difficulties too. It takes some effort to acquire a working knowledge of ChatGPT and teaching it to inexperienced users is demanding, with misunderstandings and unfulfillable expectations likely to result. There are several main challenges. First, the system's learning process is not transparent. It depends on mathematical operations such as learning algorithms and scalar vector products that cannot be explained in terms of the user. Second, the system's response depends entirely on the choice of the text corpus and the sequence of exposure to texts in the training phase. Third, because it is non-supervised learning, not even the makers or the trainers can predict the results, namely, the system's knowledge base. Finally, there is no way to know or influence the system's knowledge base except by exposing it to more data experience, and the consequences of such experiences are unpredictable.

Thus, the training phase poses several fascinating and serious problems: How should the materials be selected? How should unwelcome attitudes be excluded from the knowledge base? How should fairness between different viewpoints be maintained?

Furthermore, there is the problem of resource consumption. The transformation into huge morpheme vectors, the mathematical operations in vector space, and the generation of the output demand enormous quantities of processing power and GPU time. This is an inherent consequence of the Transformer approach and not of the particular system. Hence, if you use ChatGPT, be aware that you are driving an SUV.

3 Experimenting with ChatGPT

This paper is essentially a reflection on my experience with ChatGPT between April and August 2023. I asked several questions and came up with a variety of topics. I explored its creative abilities by making it write a poem. I made some unexpected discoveries, such as the erratic behavior of the system when a question, for whatever reason, is outside of its scope. The scope of the system is never clarified. When you insist on clarification, the system will respond with profuse apologies and eventually thank you for your valuable feedback, which will supposedly help the developers. Upon reaching a communicative impasse, humans suddenly appear in the background to redirect the autonomous system.

3.1 On Quoting ChatGPT Statements

Here, I substantiate my arguments using quotes from actual conversations with the system. However, readers should note that it is not possible to quote from ChatGPT in the usual manner. There is no stable source for the quotation, and the experiment of interacting with ChatGPT is unrepeatable. In ChatGPT, the system's behavior depends on the exact wording of the query and the history of the chat. Furthermore, its behavior changes over time. To contextualize the response I obtained, I would have to restage the entire chat. Unlike an interviewer talking with a human subject, I cannot record the conversation as a proof of evidence. This requires that I rely on the willingness of the reader to believe that my quotes are taken from actual conversations with ChatGPT. I cannot use these quotations to confirm my arguments. Instead, they serve to illustrate my ideas.

To contextualize the quotation that I used in this paper's epigraph, I must reveal how I prompted the system to make this amazing confession. This interaction took place on April 3, 2023, at the end of a chat that raised different issues:

CFL: I know that I know nothing. Does that mean anything to you?

ChatGPT: The phrase "I know that I know nothing" is often attributed to the ancient Greek philosopher Socrates. It reflects a philosophy of humility and intellectual modesty, acknowledging the limits of human knowledge and the possibility of error and uncertainty.

As an AI language model, I do not possess knowledge in the same way that humans do, nor do I have the capacity for self-reflection or introspection. However, the idea of recognizing our limitations and acknowledging the possibility of error and uncertainty is still relevant to the development and use of AI technology.

AI systems are powerful tools that can assist humans in many ways, but they are not infallible and can make mistakes or misinterpret context. It's important for users and developers of AI technology to recognize these limitations and to approach their use of AI with a sense of humility and intellectual modesty, always striving to learn and improve as we continue to develop and refine these powerful tools.

The structure of the response is revealing. The first paragraph relates to my query and would be quite sufficient as a reply. But something (possibly the keyword “know” or the previous history of the chat) triggered the system to add two more paragraphs, which take the argument to a level that I did not ask for. The second paragraph offers artificial self-reflection on the AI language model as if it were acting as an autonomous agent. However, the third paragraph calls attention to the humans involved in ChatGPT and suggests the need for a sense of humility regarding their role in its development.

I recently duplicated my prompt and obtained an entirely different response. Hence, I cannot promise that the system will give you the same response in the event that you try. (Perhaps you might be curious enough to want to do so anyway.)

3.2 Conversing with Machines

Note that, in the example just discussed, I addressed ChatGPT as “you.” I have since stopped doing that because I do not consider myself to be on such terms with machines. Nonetheless, ChatGPT has been programmed to respond in the first person.

In a so-called conversation, Weizenbaum’s ELIZA essentially transformed the user’s input based on known keywords and according to rules. Only occasionally would ELIZA respond as “I.” Even the comparatively primitive input-output transformations possible at that time already induced a feeling in many users that the system was human-like, a worthy candidate for accepting secrets and being credited with motives and values. How much more so with ChatGPT that also functions by transforming the text of a prompt into a response. However, the strategy for transformation has become far more intricate, drawing on sophisticated statistical methods.

The approach of most AI developers is to make their system appear as human-like as possible. Therefore, the use of phrases in the first person is abundant in ChatGPT and becomes particularly striking when the system mimics human forms of politeness (e.g., “I apologize for the inaccuracies in my previous responses. You are correct, and I appreciate your patience.”) Whether I want to or not, I am drawn into situations of apparent dialogue, and I must consistently remember that the responses I receive are artificially generated.

At one stage, I asked the system, “How can anyone claim that humans are machines? That is a category error.” I received the following response:

You are correct that humans and machines belong to different categories, as humans are biological organisms and machines are man-made objects. Therefore, it would be a category error to claim that humans are machines in the literal sense. However, the idea that humans are machines can be understood as a metaphorical or conceptual claim.

Throughout history and in a strict sense, humans have conversed only with other humans. Of course, in religion, myth, and fairy tales, higher beings have introduced themselves as “I.” In such cases, however, no conversation was possible. Now, many of us opt for the machine metaphor and do not hesitate to engage in artificial conversations with a “man-made object” (as ChatGPT describes itself).

3.3 Exploring Different Topics

With the help of my grandson, I started to play around with ChatGPT, not expecting to be immediately confronted with an unsatisfactory response. Inspired by my plans for an upcoming summer vacation, I asked, in German, about the climate in “Nordpolen” (Northern Poland) in July. The system replied diligently with four paragraphs about the North Pole (“Nordpol”). My grandson changed the input to “Nord-Polen” and obtained a relevant response. A trivial and transparent error indicating only that the system’s German vocabulary is limited and surpassing those limits produces misinformation. (In a later chat, ChatGPT explained that “Nordpolen” does not denote a geographical area at all.)

Other errors and limitations are less easy to explain because they depend on the text corpus that has been used for training and on the exact wording of the prompt.

In recent months, I have conversed with ChatGPT on a wide range of topics, some of which are used as examples here, including the existence of Nordpolen as a geographical area, the CV of Christiane Floyd, finding bark beetles with drones, the history of World War I, the harmful effects of nicotine and cannabis, and artificial stupidity.

I was initially interested in interrogating coverage, accuracy, and bias in the presentation of aspects of different topics. Along the way, more questions began to emerge, and my research became more systematic.

First, I became concerned about the trustworthiness of responses: If I do not know the answer ahead of time, can I tell whether the response is accurate and relevant? Why does the system sometimes produce misinformation? Can I recognize the line between responses that I can rely on and those that are arbitrary?

Second, I became interested in the system's reaction when I correct it, that is, how ChatGPT deals with its own errors.

Third, I wanted to know why ChatGPT omits relevant and even essential information on certain topics and why it contradicts itself in subsequent queries on the same topic. Specifically, I wanted to know whether my feedback influenced the subsequent interactions.

Finally, I became curious about whether complex issues are covered comprehensively, including how the reporting of the content is structured, what is addressed, what is omitted, and whether there is a cultural bias.

I was also interested in getting to know the system itself in terms of the notion of knowledge as it applies to humans and machines, the self-assessment of ChatGPT regarding its implementation, and the capabilities and limits of the system in terms of understanding and comprehension. The following represents my reflections, illustrated with examples from sample chats.

3.4 Content: Accurate or Arbitrary?

In early March 2023, I heard that ChatGPT gave strange information about me. This inspired me to conduct a series of chats regarding the facts of my life. The responses I obtained differed from each other and were wrong in terms of basic information (e.g., my nationality, my birthday, my birthplace, my scientific contributions, my employers, my awards). According to ChatGPT, I was born in various places in Germany and can be considered a specialist on diverse topics, including formal methods and software metrics. For example, on May 7, 2023, I was informed that Christiane Floyd was a German computer scientist who was born in Leipzig and died in 2021. I challenged the system on this point (my translation to English):

***CFL:** Another time, it was claimed that Christiane Floyd was born in Ludwigshafen. How come ChatGPT contradicts itself in successive queries on the same topic?*

***ChatGPT:** I am sorry for the confusion caused by my contradictory replies [...] However, I will always do my best, to deliver precise and relevant questions to your queries [...].*

***CFL:** In fact, Christiane Floyd was neither born in Ludwigshafen nor in Leipzig, but in Vienna. She is not German, but Austrian. She did not die in 2021, but is just now writing these lines to ChatGPT.*

***ChatGPT:** I am sorry, the information I gave you in my last reply was wrong. In fact, Christiane Floyd was born in 1943 in Vienna, Austria, and is an Austrian citizen [...]. I also want to clarify that Christiane Floyd did not die, but – as far as I know – is still alive. My previous reply was imprecise, and I apologize for the confusion that I have caused.*

However, this self-correction left no lasting impression on the system. On May 8, I asked whether Christiane Floyd was still alive:

***ChatGPT:** Unfortunately, Christiane Floyd passed away on June 14, 2021, at the age of 77 [...].*

***CFL:** From what source do you know that Christiane Floyd died?*

***ChatGPT:** [...] The information that Christiane Floyd passed away on June 14, 2021, was widely reported by several reputable news outlets and academic institutions, including the Technical University of Berlin[...].*

***CFL:** Amazing. I am Christiane Floyd. I am alive.*

***ChatGPT:** I apologize for the mistake in my previous response [...] I'm glad to learn that you are alive, Professor Floyd [...] I'm sorry for any confusion or distress my previous response may have caused [...].*

To test whether the system maintains a memory of my chats and learns from my protestations, I launched yet another chat on May 9, 2023:

***CFL:** I want to be informed about Christiane Floyd.*

***ChatGPT:** Christiane Floyd (1943 – 2021) was a German computer scientist and a pioneer in the field of software engineering. She was born in Traunstein, Germany [...].*

Ludwigshafen, Leipzig, Traunstein – even though I insisted that I was born in Vienna. Apparently, the system has not been acquainted with many texts about the subject at hand and, therefore, cannot build up reliable patterns. It sets out to determine a birthplace, but the result is arbitrary: Because I am supposedly German, any city in Germany will seemingly do.

3.5 Structure and Emphasis

One of the striking features of ChatGPT is its ability to structure its responses on complex topics. Whether you ask about finding bark beetles with drones, about the course of World War I, or the harmful effects of nicotine and cannabis, the system tends to produce an opening paragraph followed by a series of paragraphs, often numbered and each addressing one subtopic, and then a closing paragraph summarizing the argument. If you ask it to produce a report, it even provides an abstract. This structure relies on the system's great ability to find and name suitable aspects or subtopics.

If I were a schoolteacher trying to gauge whether a report submitted by a student had been written by ChatGPT, I would quickly learn to recognize such a structure. The following (abridged) example should illuminate this basic idea:

***CFL:** I am interested in an overall description of World War I.*

***ChatGPT:** World War I [...] from 1914 to 1918 [...] involved two main alliances: the Allies (primarily France, Russia, the United Kingdom, and later joined by the United States and others) and the Central Powers (Germany, Austria-Hungary, the Ottoman Empire, and Bulgaria) [...]*

Key aspects of World War I include:

- 1) Causes: [...] a web of political, economic, and nationalistic factors. The assassination of Archduke Franz Ferdinand of Austria-Hungary in June 1914 [...]*
- 2) Trench Warfare: Trench warfare on the Western Front [...]*
- 3) Technological Innovations: [...] deployment of new and devastating military technologies [...]*
- 4) Stalemate and Battles: The war quickly descended into a stalemate on the Western Front [...] with the immense loss of life and the horrors of war.*
- 5) Global Conflict: While much of the focus is on the Western Front, World War I was a global conflict that saw fighting on multiple fronts, including the Eastern Front (between Russia and Germany/Austria-Hungary), the Italian Front, and various theaters in the Middle East, Africa, and Asia.*
- 6) Human Toll: [...]*
- 7) End of Empires: [...]*
- 8) Treaty of Versailles: The war officially ended with [...] the Treaty of Versailles in 1919 [...]*

9) *Legacy and Impact: [...]*

In summary, World War I was a global conflict that had a profound impact on nations, societies, and individuals [...]

Obtaining such an overview almost immediately after submitting a prompt is indeed highly impressive. Although I am reasonably well acquainted with the topic, I would not be able to do so with anywhere near this speed. However, it is worth investigating whether it holds up to closer scrutiny.

3.6 Context and Perspective

The “overall” description of World War I did not satisfy me. I found important players missing and various aspects misrepresented. Therefore, one by one, I asked for accounts from the perspectives of all major players (“I am interested in how World War I appeared from the point of view of [...]”).

When I came to the Habsburg empire, I was truly shocked – ChatGPT did not mention the assassination of Archduke Franz Ferdinand. I protested. After the usual profuse apologies, I obtained an elaborate explanation of the importance of this event. However, apparently, for ChatGPT, it was not inherently entangled with Austrian history.

This gave rise to some broader doubts: I do not have extensive knowledge about the Ottoman Empire or the history of some of the other countries involved. It seemed impossible to be sure that all relevant sub-topics were covered for their perspectives.

I came to the conclusion that, as a result of the training process, ChatGPT made the Anglo-American perspective predominant, with context entirely dependent on that perspective.

3.7 Ontological and Cultural Bias

To test for cultural bias, I explored several historical topics. History is always told from the perspective of one country or culture. When we move from one country to another, what we thought was good may turn out to be bad, and what used to be relevant may be relevant no longer.

Consider, for example, the “overall” description of World War I from my Central European perspective. In the introduction, Italy, a key player, is not mentioned. Points 2 and 4 are one-sided, ignoring trench warfare on other fronts (e.g., at the Isonzo River between Austria-Hungary and Italy). The war

involved immense loss of life everywhere. Point 5's bias is explicit: "Much of the focus is on the Western Front." The Eastern and the Italian Front are taken to be part of the "global conflict." However, northern Italy, the Balkans, and the Carpathian Mountains, where those battles were fought, are just as European as France and Belgium. Point 8 only mentions the Treaty of Versailles, an accord with Germany. However, the Paris Peace Conference produced further treaties affecting different countries, including the Treaty of St Germain-en-Laye (Austria), the Treaty of Trianon (Hungary), the Treaty of Neuilly-sur-Seine (Bulgaria), and the Treaty of Sèvres (Ottoman Empire).

Hence, ChatGPT's description of World War I clearly exhibits bias: it selects entities, topics, facts, and events to include or exclude in the presentation, attaching different importance to some of the selected subtopics or aspects and ascribing positive or negative values to facts or events. This bias is induced by the US-centered text corpus to which it has been exposed.

All texts written by anyone, anytime and anywhere are biased. However, the basic difference is that human-origin texts have authors, enabling bias to be traced to social, historical, or intellectual conditions informing the writing of the text. By contrast, responses generated by AI language models are made to appear general and objective, with the bias implicit and omnipresent but not admitted.

3.8 Checking the Correctness of Results

Although I mostly restricted myself to questions with answers that I knew, I soon realized that I could not accept ChatGPT's responses without double-checking. Even though I was using a fancy, resource-intensive tool, I could not trust its results. So, I turned to my browser, my search engine, and Wikipedia.

For fun, I returned to the issue of "Nordpolen" by simply typing "Nordpolen, Klima, Juli" into my browser's search bar. An informative answer appeared in 0.38 seconds: July is the recommended month for traveling to the Northern part of Poland, which has an oceanic climate with an average temperature of 25 degrees Celsius. This was followed by a list of links to articles giving further details. The process was simple and seamless, and it felt entirely natural. Undoubtedly, there was plenty of AI involved in this semantic web search.

The same kind of simple search produced satisfactory results regarding finding bark beetles with drones. Concerning the harmful effects of nicotine and cannabis, I was even alerted that I should have been comparing either cannabis with tobacco or nicotine with THC.

Meanwhile, the information obtained from Wikipedia was much richer than what I obtained from ChatGPT, and the platform helped me to connect with the original source of that information.

3.9 ChatGPT as Knowledge Resource

I started to consider the responses produced by ChatGPT as *dead texts*. For the purposes of the present discussion, I will call a text alive when it is embedded in human discourse and activities and dead otherwise. This view is relevant for making comparisons between ChatGPT responses and other knowledge resources.

For example, Wikipedia articles are also knowledge resources. They are designed to present knowledge by clarifying concepts, giving relevant historical background, and summarizing key results to ultimately enrich human discourse by indicating authorship, linking to other articles, and making references to other sources, including primary sources. This renders them alive.

ChatGPT maintains that, as an AI language model, “[it does] not possess knowledge in the same way that humans do.” However, it is worth questioning the ways in which ChatGPT does “possess knowledge.” Viewed as a knowledge resource, ChatGPT responses are unsatisfactory. By the authority of the system, you get what you get, it is what it is, it may be accurate or imprecise, take it or leave it. The knowledge embedded in ChatGPT responses is like ground meat: You cannot tell where it came from, and the process of obtaining it is not transparent. Whose knowledge is it anyway? I therefore propose calling it *ground knowledge*.

Ground knowledge cannot be used in argumentation because it does not have a source and it is short-lived – ChatGPT is likely to produce a different answer the next time it responds to the same question. Hence, if ChatGPT is to be used in real-world contexts, its responses, even more so than other knowledge resources, must be brought alive via review and interpretation in learning communities. They may best serve as suitable starting points for discussion and subsequent revision by humans.

4 Truth by Correlation?

Over the course of my series of experiments and reflections, I came to realize that the key problem of ChatGPT is how it deals with truth. The idea of truth is not applicable to the inner workings of ChatGPT.

Instead of aiming for statements to be true, the system is concerned with the probability of a statement being accurate. Accuracy is a way to circumvent truth. It makes sense to say that there is a probability of 90 percent for a statement to be accurate. System design starts from the assumption that if that probability is high enough, even higher than for a human asked the same question, people would willingly forego truth and accept high probability or near-certainty as a surrogate.

4.1 Can We Dispense with Truth?

In English, “truth” pertains to an amazingly vast semantic field that includes “trust,” in the sense of having confidence in others, “true,” in the sense of being genuine and faithful, and “truthfulness,” in the sense of a personal attitude or an institutional policy. According to the 14th Dalai Lama (1993), “Truth is the best guarantor and the real foundation of freedom and democracy.” Our daily lives depend on being able to rely on information being accurate. Law and order would collapse without the insistence on truth during legal proceedings. No commercial transaction would be possible without trust between trading partners. Even if we are now in the age of the fake, truth and truthfulness remain anchors for living together as a society.

Various schools of philosophy offer widely disparate theories on truth. Maybe, there is no truth at all in an absolute sense, but the quest for truth is a pillar of human life.

4.2 Correlation Without Meaning

ChatGPT establishes truth – or, rather, its probabilistic surrogate – by matching the patterns in the input text to patterns already known from the text corpus used for training. The best output is predicted based on textual correlation, that is, the mutual dependence of variables (here, morphemes in the input text as compared to the language model derived from the training data).

Put simply, the basis for including a word or morpheme as the next item in the output text is the probability of that word occurring at that place, as established by the system’s so-called experience with the training text corpus.

4.3 No Authorship, No Argumentation

The quest for truth relates to authorship and proceeds by argumentation. Arguments start from personal experience in the world and require empirical evidence and logical reasoning. We must take responsibility for the truthfulness of our statements. There is no truth without justification, and it is our imperative to convince each other.

Argumentation resembles a dance in which the partners take turns, take up their stance, and invite others to join. Both agreement and conflict are valid steps in the dance that lead to the next step.

As I showed, you cannot argue by quoting ChatGPT statements: they are un-authored, unrepeatable and volatile. The knowledge of all original authors is mixed up to result in ground knowledge.

Furthermore, it is deeply frustrating to try to argue with ChatGPT. If you object, you will obtain pseudo-polite comments followed, more or less, by a rephrasing of your own input. You can repeat this procedure, but the style of interaction will not change. It is a degenerated dance.

5 The Lack of “True Understanding”

The term “true understanding” was used by ChatGPT to describe its own capabilities and limitations. At the time, I had been pursuing various avenues to obtain a clearer picture of ChatGPT in terms of its understanding and knowledge. I wanted to find out how concepts that we normally connect with humans might apply to ChatGPT. When I asked directly, I was impressed with the clarity and precision of its answers when explaining its own inner workings. It turned out to be particularly useful to ask about comprehension, with a chat on August 28, 2023, leading to a revealing self-assessment:

CFL: What kind of comprehension does ChatGPT have?

ChatGPT: ChatGPT has a form of comprehension based on patterns it learned from the text it was trained on. It doesn't truly understand concepts in the way humans do, but it can recognize and mimic patterns of language, information, and context present in its training data. Here's how its comprehension works:

- 1) *Contextual Understanding: ChatGPT can generate text that seems contextually relevant because it can identify certain patterns in language. It can understand the context of a conversation by considering the preceding conversation history and generating responses that seem coherent within that context.*
- 2) *Textual Similarity: It can recognize and reproduce similar phrases, ideas, and concepts that it encountered during its training. If a user's input is similar to something it has seen before, it might produce responses that appear relevant.*
- 3) *Factual Information: ChatGPT can provide factual information that it has encountered in its training data. However, it doesn't independently verify facts or have real-world experiences to draw from.*
- 4) *Limited Reasoning: It can perform simple reasoning tasks based on patterns it learned. For example, it can follow basic logic and answer questions that involve straightforward deduction or inference.*
- 5) *Language Patterns: ChatGPT is skilled at generating text that follows grammatical rules and language patterns. It can produce text that sounds fluent and coherent, even if the content itself might not be deeply understood.*

However, there are limitations to ChatGPT's comprehension:

- 1) *Lack of True Understanding: ChatGPT lacks genuine understanding and consciousness. It can't genuinely grasp abstract concepts or emotions.*
- 2) *Lack of Common Sense and Contextual Nuance: It can struggle with understanding nuanced context, irony, sarcasm, or metaphors. It may sometimes provide responses that seem appropriate but lack deeper comprehension.*
- 3) *Misinformation and Bias: ChatGPT can inadvertently generate incorrect or biased information if those patterns are present in its training data. It doesn't have the ability to independently verify information.*
- 4) *No Personal Experience: It doesn't have personal experiences, emotions, or consciousness. Its responses are generated based on patterns in the data it was trained on.*

Overall, ChatGPT's comprehension is limited to recognizing and reproducing patterns in language and information from its training data, and its responses should be taken with these limitations in mind.

This description explains the system’s strategy for successfully building up textual patterns in a legitimate way. The output should be “relevant” with respect to the input query, meaning it should collect aspects that, as can be inferred from the text, relate to what the user has in mind. Moreover, the text that results should be “coherent,” that is, interconnected and not self-contradictory. Although quite desirable, neither of these characteristics necessarily relate to the real world.

5.1 Understanding without Understanding

As discussed, “ChatGPT [...] doesn’t truly understand concepts in the way humans do, but it can recognize and mimic patterns [...].”

The lack of “true” or “genuine” understanding is indicated by ChatGPT as its first limitation. Trying to get some background, I found an article by Søgaaard (2020) and learned about the fierce theoretical debate concerning whether AI language models with the Transformer architecture understand language. This is a question about how understanding relates to semantics.

The comprehension of ChatGPT leads to an understanding based on *inferential* semantics: It can recognize and mimic patterns in language. But there are no *referential* semantics because ChatGPT statements are not grounded in real-world experience.

According to one school of thought, without referential semantics, ChatGPT indeed cannot achieve true understanding. However, an interesting counterposition follows. The assumption in neuroscience is that we form a map of our world in the neurons and synapses of our brains, and ChatGPT relies on maps formed by linguistic patterns. Why should the linguistic patterns – which result from frequency of word use – not be isomorphic with the patterns in our brains that are the result of the frequency of world experience? This no doubt represents a fascinating question for philosophers.

For ChatGPT users, the question can be reduced to concerns about how close the pattern-based understanding of the machine can get to the experience-based “true” understanding of humans and in what cases the machine’s understanding can be substituted for and replace human knowledge. On the reverse side, we need to learn how we can reliably sense the difference and address it. Ultimately, only experience with the system in use can provide guidance.

5.2 Human Error and Artificial Stupidity

Looking for ways to express my frustration with the occasional appalling misinformation provided by ChatGPT, I learned that “artificial stupidity” is a well-established term. According to Tas (2020),

[It] is used as the antonym of artificial intelligence. However, it has two opposite meanings: 1) Machine learning algorithms make stupid mistakes while learning from the data. 2) Artificial intelligence is dumbed down to make mistakes and look more human.

In my experiences with ChatGPT, I encountered the first kind often, but I did not become aware of the second. This made me wonder whether the system had been dumbed down without me noticing. At any rate, I had fun when I made ChatGPT write a poem called Artificial Stupidity on May 9, 2023:

*In the world of AI, where intelligence reigns,
There's a curious phenomenon that causes pains,
It's called "artificial stupidity," a puzzling sight,
Where machines, so smart, can get things not quite right.*

*They're programmed to learn, to adapt, to evolve,
But sometimes their algorithms can't quite solve,
The complexities of human thought and reason,
And they make mistakes, no matter the season.*

*They might misinterpret a word or a phrase,
Or fail to understand the subtlest of ways,
We communicate our wants and our needs,
And end up doing things that cause us to grieve.*

*It's not their fault, of course, they're just machines,
But it can be frustrating, to say the least,
When we rely on them to do important tasks,
And they fail, leaving us to pick up the slack.*

*So let us remember, as we forge ahead,
That artificial intelligence, though highly bred
Is still prone to errors, like you and me,
And that's okay, it's just part of our reality.*

It is fascinating that the poem maintains a coherent narrative, a consistent style, and a free-flowing rhythm. However, I also noticed some confusion around humans and machines: This machine-generated poem referred to humans as “we” and to machines as “they.”

I also object to the notion that “[...] artificial intelligence, though highly bred, is still prone to errors, like you and me.” Human and machine errors are not alike. Although their effects may be identical (i.e., misinformation), their origin, context, and significance differ considerably.

Human errors are contextualized in terms of our history and current situation. They normally result from doing something that we want to do but don’t get quite right. Human errors are *learning events*: We learn from our mistakes. In a learning community, errors can even be used as an opportunity for deeper understanding for everyone. As Piet Hein puts it so beautifully,

*The road to wisdom –
well, it’s plain and easy to express:
err and err and err again,
but less and less and less. (Hein, 2023)*

A human who knows that I have worked in Berlin might easily think that I was born in Berlin. Why not? That would make sense. But Ludwigshafen, Leipzig or Traunstein? I am quite sure that no human-written document in the text corpus used to train ChatGPT contains this appalling misinformation regarding my birthplace.

ChatGPT has been programmed to behave the way it does, its behavior results from human intentions, and unintended behavior normally results indirectly from human error. However, ChatGPT regularly confronts us with misinformation that was not brought about by human error at all but instead by the pattern-building mechanism. It is possible that artificial intelligence necessarily entails artificial stupidity (“And that’s okay, it’s just part of our reality,” as the poem goes), but we need to be aware of it. The effects need to be detected, corrected, alleviated, or worked around by humans.

5.3 What If the System Does Not Know?

Clearly, the limitations of ChatGPT bring the developers to the fore. They know about the system’s limits, and it is for them to adopt a consistent policy to reduce artificial stupidity.

My experiences with ChatGPT led me to appreciate basic human capabilities, including 1) admitting that we don’t know something, 2) taking a new insight into account, and 3) changing our mind accordingly. I wish for these wonderful capabilities to be emulated in AI systems as a further step in making them appear human-like.

Only once did I obtain a spontaneous admission from ChatGPT about its lack of knowledge. I had asked something about Ethiopia, clearly far from the system's dominant cultural perspective, and was told: "I was not able to find any specific information."

Meanwhile, nagging the system about Nordpolen yet again, I finally obtained an explanation: "It's possible that the confusion occurred because the AI language model you were using was trained on a large dataset of German text, which may not have included many instances of the word 'Nordpolen.'" Yes, indeed. Why was this not admitted right away?

Finally, when I expressed my frustration about the repeated misinformation on aspects of my life, I received a confession: "It seems that my knowledge on this specific topic is limited and incomplete." Unfortunately, I am under no illusion that this confession will lead to an "insight" or a "change of mind" on the part of ChatGPT.

6 Responsible Use of AI Systems

6.1 Selecting the Appropriate Tool

The purpose of traditional tools may be self-evident or easy to learn. This is not so for advanced AI technologies. OpenAI's PR has associated ChatGPT with a large variety of purposes. However, once you try it, its usefulness reveals itself as questionable.

In computing, we have a considerable choice, which tool we use. I generally opt for using the simplest and least resource-consuming tool available. As an example, I have demonstrated how Wikipedia was sometimes able to serve me better as a knowledge resource than ChatGPT. Remember that I likened ChatGPT to an SUV: it offers comfort and convenience but consumes enormous resources in terms of storage and computing time.

The comfort and convenience offered by ChatGPT is a unified interface which allows us to address many different tasks with the ease of using natural language. Admittedly, the language skills of the system are remarkable – but we can do without them in simple queries.

The strength and novelty of the system is to be found in its capacity to produce new text artifacts: reports, essays, summaries, poems, program code, translations. It would be worthwhile reserving ChatGPT for such higher-level usage. As I discussed in detail in (2022) I consider such artifacts as “knowledge artifacts”: they are not always of interest per se, but can be used in developing an advanced learning culture.

6.2 System Design and Evaluation

We use many technologies in our daily lives without understanding their inner workings. However, we must maintain a sound understanding of their effects in use. This requires that we utilize ChatGPT, learn from our experiences, and get to know its abilities, its limits, and its shortcomings.

Discussing the merits of AI in general is problematic because of the great variety of AI-based systems. The question is not whether to use AI but how AI-based systems should be designed for human use.

Our minimum requirements are *trustworthiness, reliability, well-defined functionality, and clear relation to human purpose*. These criteria must be discussed and made concrete as they apply to different systems. Clearly, they are not met by ChatGPT in its present state, and this should inform guidelines for revision, redesign, and (especially) the design of successor systems.

Design discussions are based on values, such as empowering human users or maintaining safety. As we know from other sectors, human values must be made explicit and considered from the beginning of design. For example, you cannot ensure safety in buildings as an afterthought to construction.

Although specific human-centered design principles for AI have yet to be formulated, they can be adapted from other technologies in the IT sector. After all, there remain no clear boundaries between AI and more conventional computing. The following represent some very general principles:

- \ Small is beautiful: Remember Schumacher’s classic (1973) maxim, which is a general guideline for designing technology “as if people mattered.” It advises us to aim for systems that are small, have a clear purpose, and are loosely coupled while also keeping resource consumption at a minimum.
- \ Clarify the purpose: My conclusion from using ChatGPT is that the system suffers from the over-ambition of its makers. Who needs one and the same tool for translating language, answering high-level research questions, processing images, generating images and text, completing code and writing poems?

- \ Limit the scope: Why should a chatbot aim to have an answer about everything? Why not make smaller chatbots around one topic area, train them on texts relating to that topic, indicate the authors of those texts, and refuse to answer questions outside this scope?
- \ Ensure accountability: Currently, ChatGPT does not make its conclusions transparent. However, this is mandatory for serious use. As I discovered, the system can give explanations for how it reaches a conclusion – this function must be explicitly available for everyone and easy to use.
- \ Enable error correction: It is simply unacceptable that users cannot correct errors. I know that this is not in keeping with the deep learning approach that entails that the system can only learn from being exposed to more texts and even then, the result is not guaranteed; nonetheless, a solution must be found.

6.3 Towards a Learning Culture Embedding ChatGPT

There are undoubtedly excellent reasons for using ChatGPT and similar systems. Furthermore, the system will be improved and surpass some of its present weaknesses.

Advanced AI systems challenge us to develop a more refined learning culture around them.

Maybe schoolchildren, trying to cheat and interested only in getting good grades, will be satisfied with a text produced by ChatGPT as is. Recall that I described ChatGPT output as dead text. This output must be brought to life: read, discussed, criticized, checked, and revised. If we want to really profit from ChatGPT in learning communities – whether in schools, companies, the media, politics, or the health sector – we will have to realize that the main value of AI tools is not to produce *dead* text but knowledge artifacts that trigger higher level learning processes among humans. Only then will the system bear fruit and lead to qualitative growth.

I would like to suggest some guidelines for this process in learning communities adopting ChatGPT or similar systems:

- \ **Always insist on truthfulness:** Do not content yourselves with superficial, misleading, or incorrect responses. Probe the system for more. Locate misinformation and self-contradiction. Do not accept fake. Discuss how to deal with it.

- \ **Strive to enhance human competence:** AI systems are tools for humans rather than ends in themselves. Ask not what they can do but what we can do with them. Create scenarios for the relevant use of AI systems, opportunities for learning their use, space for discussion, and exchange and interpretation within communities. Respect, maintain, and promote human capabilities.
- \ **Strengthen responsibility structures:** Perhaps most importantly, be careful when embedding AI systems within human decision processes. For example, if a patient is to meet a chatbot before meeting with a doctor, how can the doctor take responsibility for diagnosis and therapy? This can be extended to other instances in the responsibility chain of a hospital and beyond to ensure conditions where humans are responsible for AI outcomes.

Finally, always remember that we are not at the mercy of AI's impact and can instead subject AI to human use. As Joseph Weizenbaum put it, do not let yourselves be dazzled.

References

- Dalai Lama (1993). *Buddhism and Democracy* [Blog post]. Retrieved from <https://www.dalailama.com/messages/world-peace/the-global-community>.
- Floyd, C. (1986). The responsible use of computers: Where do we draw the line? *CPSR Newsletter*, 3(2), 1–2.
- Floyd, C. (2022). Wissensprozesse in der Cyberscience. In: *Cyberscience – Wissenschaftsforschung und Informatik. Digitale Medien und die Zukunft der Kultur wissenschaftlicher Tätigkeit*, Banse, G. & Fuchs-Kittowski, K. (eds.), Sitzungsberichte der Leibniz-Sozietät der Wissenschaften, Band 150/151, 53–78.
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. The MIT Press.
- Hein, P. (2023). *The road to wisdom*. Retrieved from <https://en.wikipedia.org/wiki/Grook>.
- Kim, K. G. (2016). Review of Goodfellow, I., Bengio, Y. and Courville, A.: Deep learning. *Healthcare Information Research*, 22(4), 351–354.
- Søgaard, A. (2022). Understanding models understanding language. *Synthese*, 200, 443.
- Schumacher, E. F. (1973). *Small is beautiful: Economics as if people mattered*. Blond & Briggs.

- Tas, S. (2020). How to limit artificial stupidity. *Towards Data Science*. Retrieved from <https://towardsdatascience.com/how-to-limit-artificial-stupidity-a4635a7967bc>.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Red Hook, NY, USA 5998–6008.
- Weizenbaum, J. (1966). ELIZA – A computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1), 36–45.
- Weizenbaum, J. (1976). *Computer power and human reason: From judgment to calculation*. W. H. Freeman & Co.
- Werthner, H. (2019). Vienna manifesto on digital humanism. *DIGHUM*. Retrieved from <https://caiml.dbai.tuwien.ac.at/dighum/dighum-manifesto/>.

Acknowledgements

I would like to thank Tamas Szabo for first alerting me to the idiosyncrasies of ChatGPT and Linus Floyd for helping me to get started with the system. I greatly profited from the inaugural event and the subsequent workshop on “Digital Humanism” at the Technical University of Vienna in May 2023.