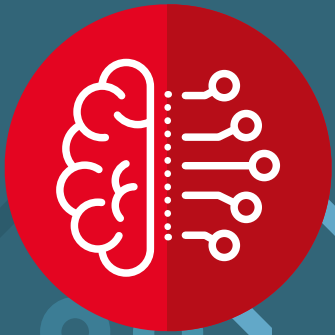


MYTHS AND CRITICISM OF ARTIFICIAL INTELLIGENCE

JOSEPH WEIZENBAUM (1923–2008)

weizenbaum institut

Berliner Landeszentrale für politische Bildung

BERLIN

CONTENTS

ABOUT THE ISSUE

Will computers one day surpass humans in intelligence? And if so, will they pose a threat to us? Or, on the contrary, is technological progress our last hope for combating poverty, disease, and climate change? The debates about the potential impact of increasingly powerful computers on our lives often feel like science fiction. Promising utopias and frightening dystopias tend to coexist, especially when it comes to artificial intelligence (or AI). We see deceptively realistic deepfakes of politicians and celebrities or videos of robots that sometimes dance and sometimes fight, leaving viewers with mixed feelings of fascination and horror.

Technological enthusiasm and skepticism have existed simultaneously since the early days of computer technology. For a long time, the dominant belief was that computers could change the world for the better. The digital revolution would empower people, connect humanity into a "global village," and democratize societies. From this perspective, it seemed as if there was no problem without a technological solution.

On the other hand, some people, including those directly involved in the early stages of computer development, voiced words of caution. One prominent critic was Joseph Weizenbaum (1923-2008), who co-developed the first computer system for automated check processing in the 1950s and has been involved in AI research since the 1960s. Born in Berlin, Weizenbaum became famous for ELIZA, a chatbot he introduced in 1966. ELIZA is considered the precursor to ChatGPT and voice assistants such as Siri and Alexa. However, this

Joseph Weizenbaum at MIT, 1970s
Photo from the image archive of the documentary film “Weizenbaum. Rebel at Work”

program made Weizenbaum not only a pioneer of AI, but also one of its fiercest critics.

Weizenbaum, who fled with his Jewish family to the U.S. in the 1930s to escape Nazism, quickly recognized the ambivalence surrounding the technology he was working on. Above all, he challenged the myth of the computer as a superhuman, intelligent machine and questioned our strange fascination with today’s AI innovations. Contrary to the optimism often found in technology discourse, Weizenbaum was convinced that technology could not solve many of our problems. He argued that its great reputation rather blinds us to the possibility that it may actually exacerbate the problems it is supposed to solve.

In this issue, we take up Weizenbaum’s critique and examine contemporary discussions about AI, revealing their long traditions. We also interpret prevailing attitudes of the past that continue to influence our view of computers, digitalization, the internet, and AI. Ultimately, this approach emphasizes the importance of critically questioning both historical and current developments in this field.

PART 1
MYTHS OF ARTIFICIAL INTELLIGENCE

HOW INTELLIGENT ARE COMPUTERS?

In late 2022, the U.S. company OpenAI introduced ChatGPT to the public. Behind this somewhat cryptic name is a computer program that can engage in back-and-forth conversations. ChatGPT stands out because it responds to questions in detail and can even write poems, essays, or jokes when prompted. Fascination with this chatbot's capabilities, such as producing text at a rate far faster than humans, has inspired predictions about how this technology will change—or even revolutionize—our future.

The "GPT" in ChatGPT stands for "Generative Pre-trained Transformer" and refers to a new type of artificial intelligence system: generative AI. Unlike previous AI systems, which are mainly used for pattern recognition and automated decision-making (such as spam filters), generative AI is notable for its creativity. Some AI applications, for example, can be used to create works of art or pieces of music. To do this, they are trained with large amounts of text, images, videos, or music to determine the likelihood of certain combinations of words, pixels, or

sounds. They then convert these probabilities into new combinations, a process known as "machine learning."

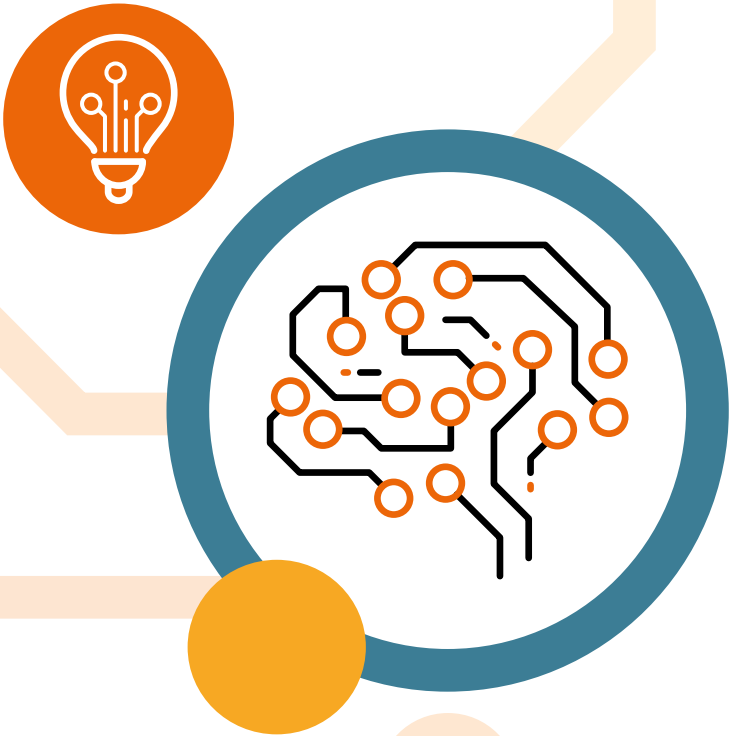
But how creative are these AI applications? Critics say that generative AI is neither creative nor innovative because it only reproduces existing knowledge and cannot generate anything entirely new. Others argue that humans are not much different. In mid-2022, a Google employee claimed that LaMDA, the AI he oversaw, had developed a consciousness and was already at the developmental stage of a child. Although few believed him, the possibility of computers thinking independently captivates us, with some even speculating that they may surpass human intelligence at one point.

Although these ideas may seem new in light of the latest AI innovations, discussions about them extend far back into history. Long before the first computer was built, people were speculating about what these calculating machines could achieve. In 1843, the English mathematician Ada Lovelace predicted that machines would one day be able to "compose elaborate and scientific pieces of music of any degree of complexity or extent."

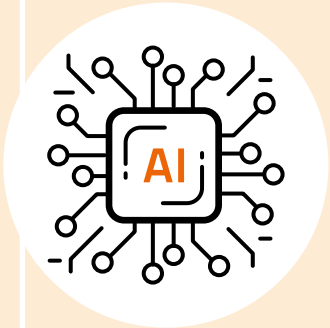
The term "artificial intelligence" is also older than many expect. It was coined by U.S. scholar John McCarthy in 1955 to describe a new field of research investigating "how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves" (see info box "Artificial Intelligence").

Ever since, the question of whether computers will one day be able to think, understand, and act like humans has

been hotly debated. Some believe that the brain is a type of computer, but that it can only evolve slowly, inevitably being overtaken by ever more advanced and powerful computers. Other voices, such as Joseph Weizenbaum, are highly critical of this notion. They remind us that just because a computer appears to be intelligent, it does not mean that it actually is. Confusing the two perpetuates a myth.



ARTIFICIAL INTELLIGENCE



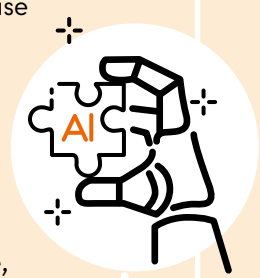
The term “artificial intelligence” first appeared in 1955 in a project proposal by the computer scientist John McCarthy and his colleagues, who applied for funding to hold a conference at Dartmouth College in the United States. The “Dartmouth Conference” covered topics such as automation, machine language processing, neural networks, and machine learning, all of which still play an important role in AI research today.

Since then, AI research has mainly pursued two goals. First, computers should be able to solve tasks that previously only humans were capable

of solving. Second, AI researchers have sought to improve their understanding of human cognition by attempting to emulate it. In both cases, the aim is to simulate human thinking.

There are two typical approaches to support these objectives. "Symbolic AI" attempts to replicate the rules of logical thinking. Popular at the beginning of AI research, this approach is now referred to as "good old-fashioned AI" (GOFAI). The other approach, "connective AI," has been pursued more frequently since the 1980s. This approach involves calculating probabilities in order to establish connections and uses "artificial neural networks" that adapt with each new input and thus "learn."

Many AI researchers envision developing a "strong AI" that can think and learn independently. Once this is achieved, it is believed that the "singularity" will soon be reached. The term describes the point at which AI will surpass human intelligence. This AI would then represent humanity's "last invention" because it would invent everything new from that point on. However, this remains science fiction. Only "weak AI" has been developed thus far, limited to specific tasks such as the automated driving, personalized search results, text, image, and speech recognition, machine translation, and chatbots.



THE ELIZA EFFECT

The "ELIZA effect" describes the tendency to perceive human characteristics in a computer based on one's interaction with it. The term "ELIZA" comes from a computer program developed by Joseph Weizenbaum in the 1960s at the Massachusetts Institute of Technology (MIT), one of the United States' leading universities. ELIZA is considered one of the first chatbots and an early predecessor of ChatGPT. However, it was limited to writing back and forth via text input and was unable to generate long texts.

Like most AI researchers at the time, Weizenbaum pursued the symbolic AI approach, which involved programming comprehensible rules to simulate logical thinking by a computer. However, defining rules for all possible scenarios of everyday life was far too time-consuming, so AI researchers limited themselves to certain scenarios and use cases.

For ELIZA, Weizenbaum chose the framework of psychotherapy. The principle was simple: The program asked questions, and the users typed in the answers. Then, ELIZA searched these answers for certain keywords and asked follow-up questions that were programmed before. Sometimes,

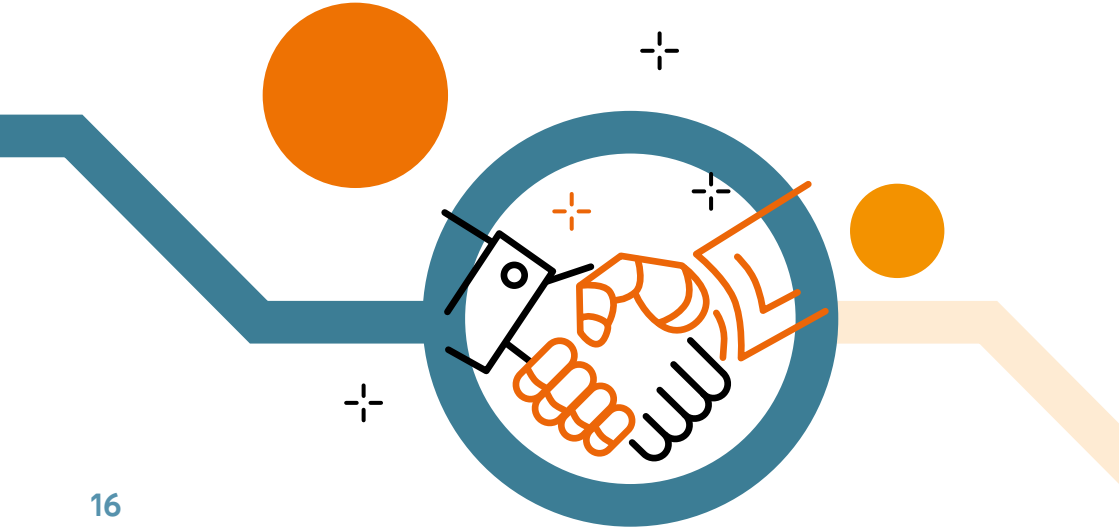
ELIZA simply reformulated the answers into new questions or asked its users to share more information. This approach allowed Weizenbaum to circumvent the major challenge of the symbolic AI approach to programming a computer with enough world knowledge to enable meaningful human-computer conversations. ELIZA didn't need to know much; it just needed to be able to ask questions.

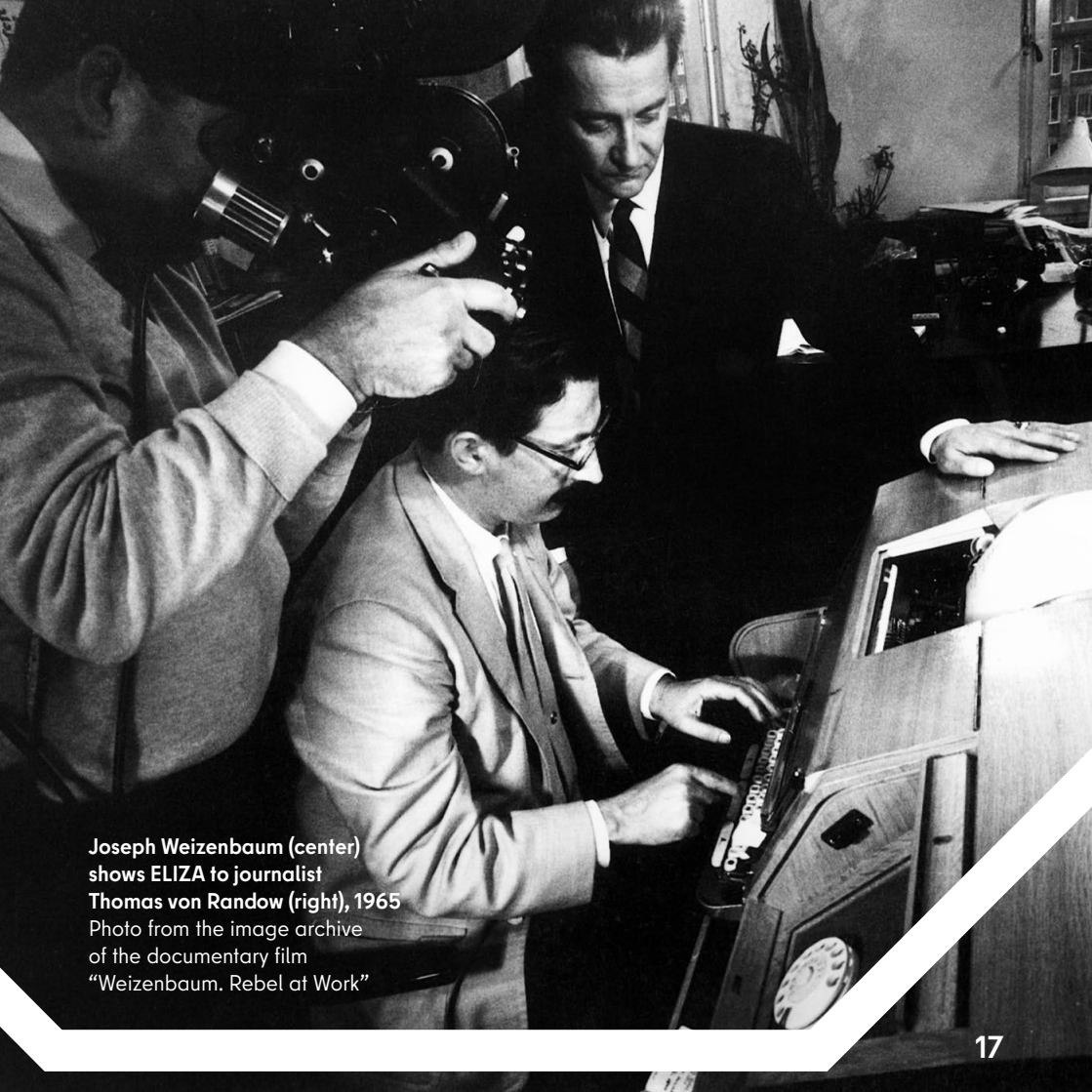
The "ELIZA effect" occurred when users began conversing with ELIZA as if it were human. In this respect, ELIZA can be seen as an early application of the "Turing test," which was introduced by Alan Turing. The British mathematician claimed that a computer could be considered intelligent if humans believed it to be intelligent (see info box "Turing test"). In later texts and interviews, Weizenbaum recounted his surprise at people attributing intelligence and empathy to his program, which was based on a few simple rules. While no one likely believed ELIZA was an actual intelligent being, Weizenbaum said that many apparently felt understood by the program.

Such an attribution of human characteristics to objects or non-human beings is called "anthropomorphism," which roughly means "humanization." Examples of this phenomenon include the house that seems to smile, the cat that looks elegant wearing a hat, and the lamp that lowers its head sadly. We humans are good at recognizing ourselves in the world around us. We often extend this habit to computers and other machines: We find humanoid robots cute because of their big, childlike eyes. We insult navigation systems when they repeatedly tell us that we are driving in the wrong direction. And we thank voice assistants politely for their help.

What shocked and angered Weizenbaum—beyond the humanization of ELIZA—was that some psychotherapists were publicly considering using such programs for psychotherapy. He saw this as disrespectful to people in need

of help who would be referred to a computer instead of receiving adequate care. He suspected that the humanization of computers had a dangerous flip side: the dehumanization of people. He was concerned that by attributing human characteristics to computers, we would begin to compare them to humans and put the two in competition with each other.





Joseph Weizenbaum (center) shows ELIZA to journalist Thomas von Randow (right), 1965
Photo from the image archive of the documentary film "Weizenbaum. Rebel at Work"

TURING TEST

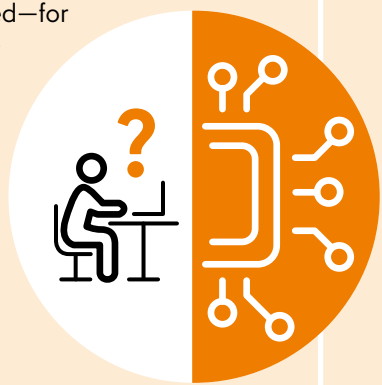
With his 1950 essay "Computing Machinery and Intelligence," Alan Turing laid the groundwork for future concepts of AI. In it, he questioned the extent to which a computer could be considered intelligent, suggesting that it doesn't matter whether a computer can actually think or understand. What matters is whether we, as humans, cannot tell the difference between interacting with a human and interacting with a computer.

To test how well a computer could imitate a human, Turing formulated the "imitation game," which became known as the "Turing test." The setup is as follows: A person, the "interrogator," sits in front of a computer, which is connected to two unknown players: another human and a computer. The interrogator's task is to ask the two unknown players as many questions as necessary until they think they can identify which player is the human and

which is the computer. If the interrogator cannot clearly identify the two players, the computer has passed the test.

Thus, the Turing test is a variant of the "duck test," in which an object's external appearance is used to determine what it is. This method has become known by the saying: "If it looks like a duck, swims like a duck, and quacks like a duck, then it's probably a duck." The Turing test would say: "If a computer answers questions like a human, then it is probably as intelligent as a human."

Though the Turing test has often been criticized—for instance, for measuring how easily people can be deceived rather than a computer's intelligence—it has had a significant influence on AI research and development.



THE IMAGE OF MAN IN AI

Since Joseph Weizenbaum invented ELIZA, the use of computer programs to treat people has persisted. Today, there are numerous AI-based self-help apps that offer psychotherapeutic or medical advice. However, their use remains controversial. Critics fear that people in need of help will be left to their own devices instead of receiving professional assistance and human attention. Proponents argue that these apps are a beneficial addition to existing therapies. Given the long waiting lists for in-person therapy, they contend that these applications are better than no help at all.

Even if the developers' intentions are sincere, the underlying image of man on which these applications are based is worthy of criticism. Besides the lack of appreciation and respect for those seeking support from such applications, there is also an assumption that these programs could provide at least some degree of interpersonal help and care.

According to Weizenbaum, this assumption that AI could be equal or even superior to humans in interpersonal matters leads to an "absurd competition" between humans and computers.

Regardless of its technical feasibility in the future, the mere possibility of this progression has profound implications for our society today.

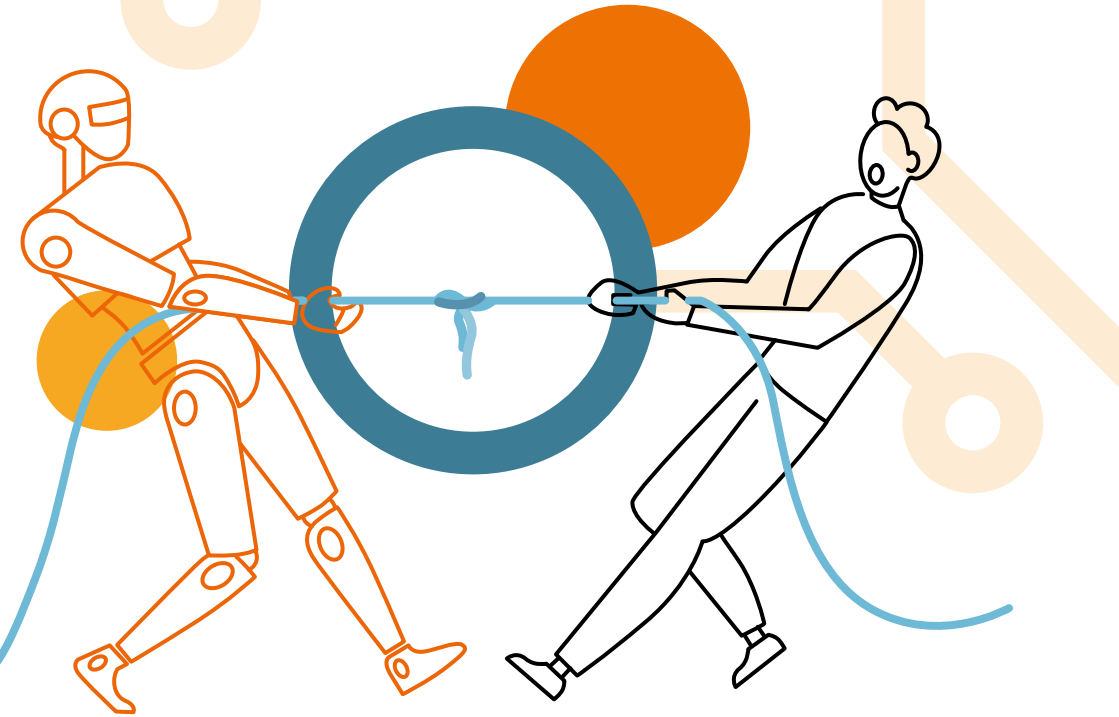
Consider parcel delivery, for example. For many years, people have been talking about how autonomous robots and drones would soon deliver parcels more quickly. In the meantime, delivery times have indeed been drastically reduced, but not necessarily due to automation. Rather, the mere prospect of robots taking over these tasks has pressured parcel delivery workers to complete their work faster than ever before. To avoid being replaced by a machine, they try to work like one.

And this phenomenon is not limited to parcel delivery. The idea of care, service, and sex robots, creative computers, and AI friends makes us all susceptible to the uncomfortable possibility of being replaced by a machine.

This reverses the relationship between humans and machines: Suddenly the question is no longer whether computers can do what was previously reserved for humans, but whether humans can be as efficient as computers. Thus, machines become the standard by which we increasingly measure ourselves (see info box "Are Computers the Better Humans?").

Therefore, Weizenbaum strictly opposed comparing humans and machines. He strongly disagreed when, for example, the level of AI development was compared to that of a child or when the human brain was referred to as a "meat machine." The idea that computers could one day be human would mean that humans are also just machines. Weizenbaum considered this idea dangerous: "We can learn from history what role the image of humanity played in the crimes of the past. In the Nazi era, Jews were portrayed as vermin—a

metaphor that legitimized mass murder." He attributes a similar dehumanization to the man-machine comparison: "These metaphors destroy the reverence for man; they make his possible end seem bearable."

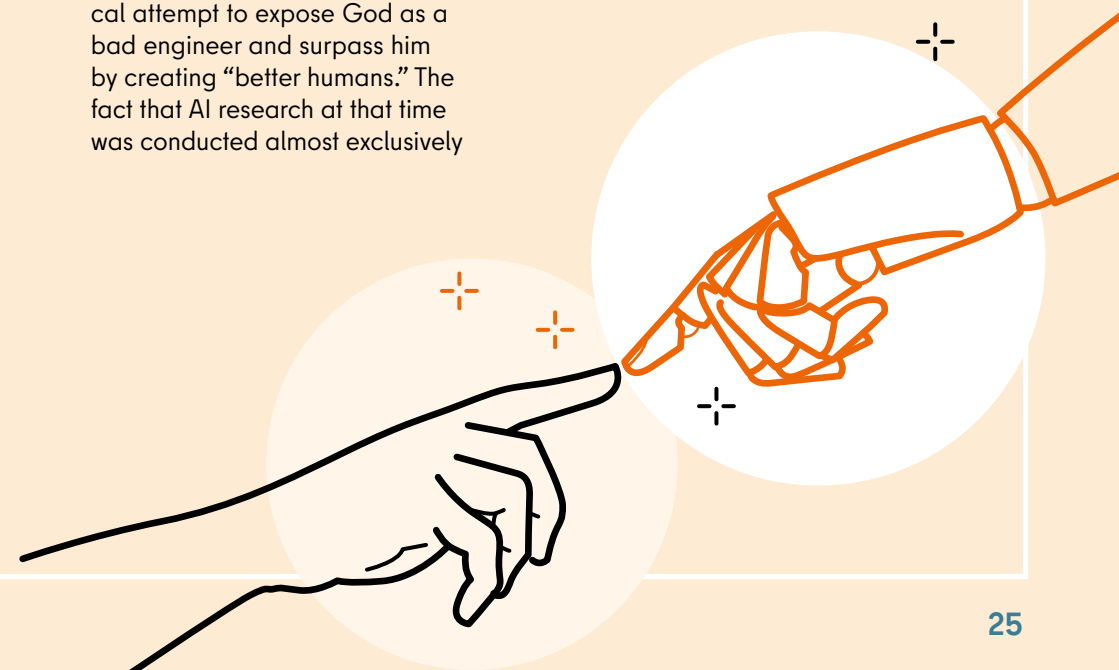


ARE COMPUTERS THE BETTER HUMANS?

In his book “The Obsolescence of Man,” philosopher Günther Anders describes the shame we feel when we compare ourselves, or are compared, to machines. This shame stems from a feeling of inferiority that arises when we contrast our transience and fallibility with the apparent perfection and immortality of machines.

This feeling can lead to a desire to perform like a machine, such as when striving for perfection in activities, using stimulants to work longer hours, or pursuing a sense of immortality. Anders describes this form of self-discipline: “Man deserts into the camp of his devices.”

Joseph Weizenbaum criticized his colleagues in AI research for taking the competition between humans and computers to the extreme. Behind their quest for an AI that would surpass human intelligence, he recognized a contempt for human life. He emphasized the megalomaniacal attempt to expose God as a bad engineer and surpass him by creating "better humans." The fact that AI research at that time was conducted almost exclusively by men also led Weizenbaum to consider the possibility of a kind of "womb envy" at play. This concept was coined by psychoanalyst Karen Horney, who stated that men try to outperform women in various areas in order to compensate for their inability to create life.



PART 2
AI AND SOCIETY

CAN AI SOLVE SOCIAL PROBLEMS?

In debates about new technologies—whether smartphones, the internet, social media, or AI—two opposing assumptions about their potential impact on society usually emerge. Some emphasize the potential of these technologies and envision a future in which technology will improve everything. Others focus on the risks and question the dangers these technologies might pose. Science fiction often portrays this ambivalence: While it presents fascinating visions of the future, it also explores the potential negative consequences in imagined dystopian scenarios.

In our everyday interactions with technology, such utopian and dystopian scenarios still seem far-fetched. We tend to view technology as a tool that enables us to complete tasks more efficiently. This also applies to AI, which can be used to translate texts, refine our internet searches, filter spam emails, and secure our cell phones with facial recognition. However, some believe that AI can do even more: For example, according to the German government's 2023 AI action plan, AI should contribute to environmental and climate protection.

The use of technology to solve problems is referred to as a "technological fix." The term was coined in the 1960s by the physicist Alvin Weinberg. He believed that social, political, or economic problems could be solved by translating them into technological problems, which makes them easier to handle. One example is the use of AI to combat hate speech online: The social problem of discriminatory and offensive statements is addressed by having an AI automatically identify and delete such content.

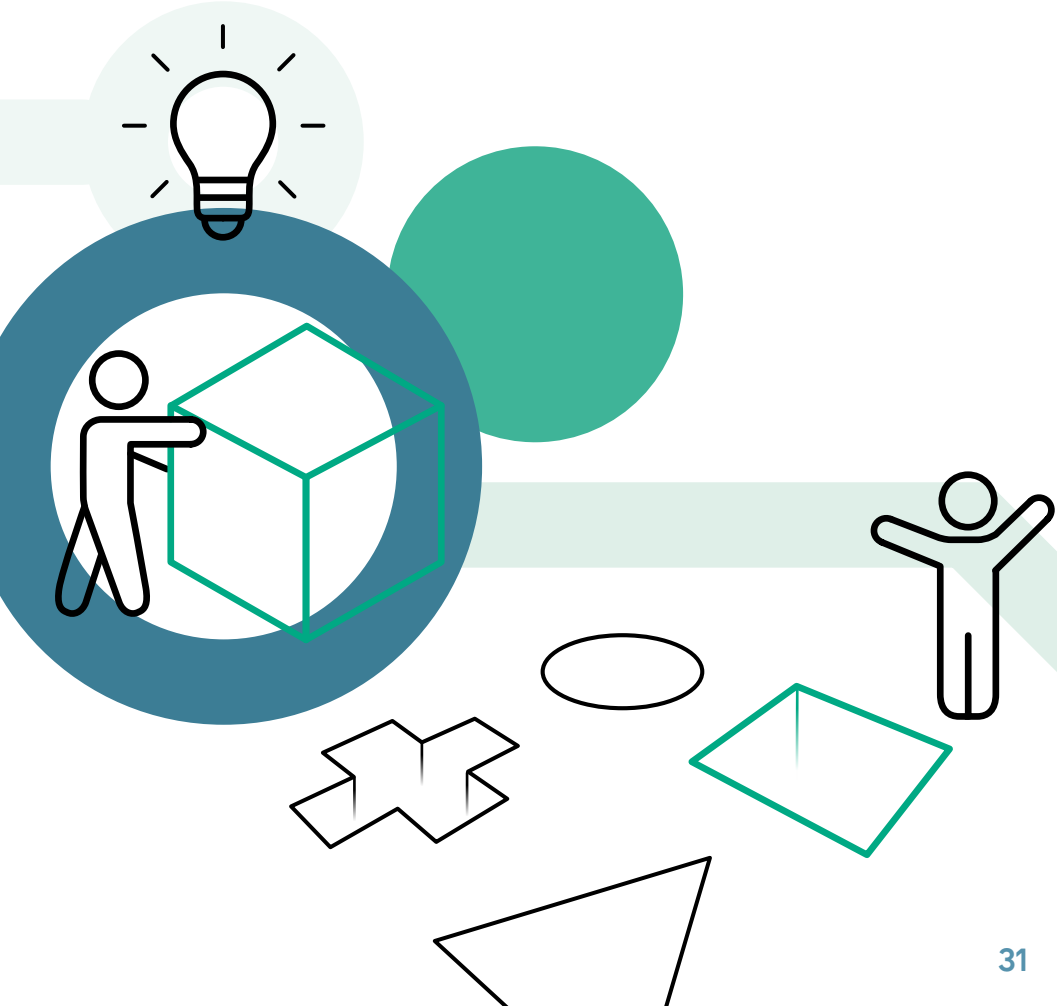
But does this really solve the problem? Probably not. Even if the AI reliably deletes discriminatory content, it does not address the reasons why people treat others in a derogatory way. Those whose content is deleted will find new ways to express their worldview. They may formulate their hate messages in a way that the AI cannot recognize them, switch to platforms without AI, or take their hatred to the streets, escalate their words into action.

Thus, a technical solution to one problem creates new problems that we then try to solve with new technologies, which in turn create new problems, and so on. This technological mindset is rooted in the widespread belief that tools have been developed to solve specific problems, and that there is basically no problem that cannot be solved with (the right) technology.

Technology critic Evgeny Morozov has called this belief "solutionism," a term he uses to describe the conviction that technology can improve the world (see info box "Californian Ideology"). However, the process is often reversed: A technology is developed first, and then a problem that it can solve is identified.

While it would be an overstatement to claim that social problems are primarily due to the use of technology, this claim

is legitimate in some cases. A critical approach to technology requires us to continually ask ourselves what problems a technology might introduce and whether these problems would exist without it.



CALIFORNIAN IDEOLOGY

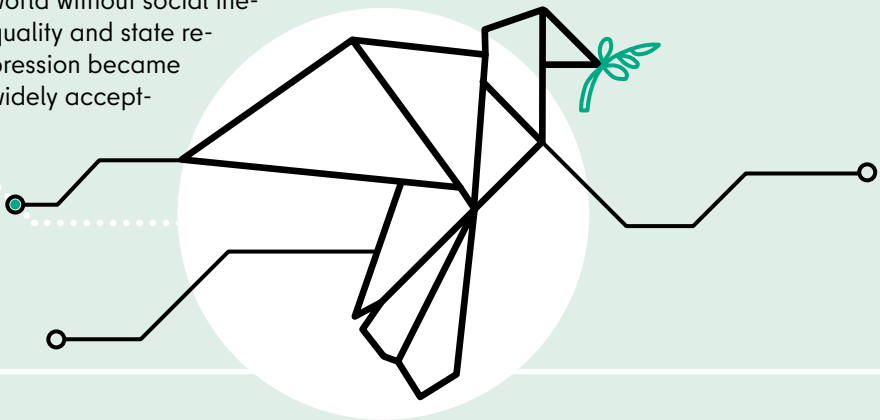
In 1996, the social scientists Richard Barbrook and Andy Cameron published an article titled "The Californian Ideology," in which they critically examined a widespread worldview in Silicon Valley, located in the southern part of the San Francisco Bay Area in California. This region encompasses Palo Alto, Cupertino, Santa Clara, and San José, and is also home to Stanford University. It has been one of the most important locations for the IT and high-tech industries, with companies such as IBM, Microsoft, Intel, Yahoo, Apple, Google, and Facebook based there now or in the past.

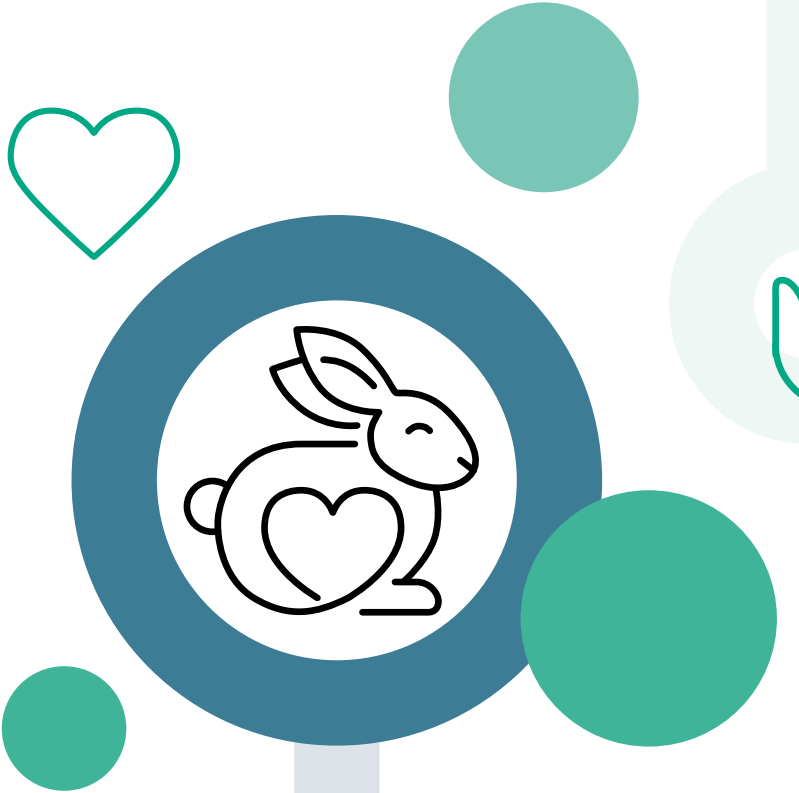
Barbrook and Cameron analyze the development of a new belief in the world-improving power of free markets and technological innovations. This belief combines contradictory attitudes, such as capitalist market liberalism and the anti-capitalist ideals of the hippie movement—including closeness to nature, self-actualization, and self-optimization.

According to the authors, there are several reasons this mixture emerged in California. First, the

Bay Area has been considered the center of the anti-establishment and the New Left since the Vietnam War protests. Second, many technology companies received substantial funding from the U.S. government for military research. Third, California's mild climate and high quality of life facilitated utopian visions of the future.

In the early 1990s, the idea that new technologies, such as the internet, would create a better world without social inequality and state repression became widely accepted. In fact, this narrative is still used today to argue against state control and regulation. Nevertheless, "The Californian Ideology" has become quite influential for technological criticism.





THE BOUNDARIES OF AI

Joseph Weizenbaum was deeply concerned about the notion that computers are problem-solving machines. In his view, this belief is an expression of a deeper phenomenon he called the "imperialism of instrumental reason." In reference to the philosopher Max Horkheimer, he defines "instrumental reason" as thinking in terms of ends and means—or, in other words, problem and solution. Unlike practical reason, which is oriented toward the right action, instrumental reasoning focuses on determining the most suitable tools or means to accomplish a given goal.

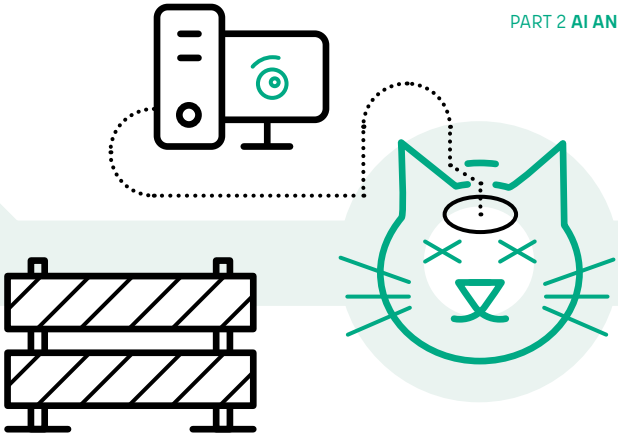
This way of thinking is common in engineering, the natural sciences, and computer science: What material do I need to make the bridge last as long as possible? What method can I use to measure the surface temperature of the sun? What changes need to be made to the software so that the self-driving car no longer causes accidents? These are all examples of instrumental thinking. However, the questions could also be: How do I build a nuclear bomb? Are mice or monkeys better suited for animal experiments? What do I have to do to hack into a bank account? Instru-

mental thinking is purpose-oriented, it does not require one to ask whether the process is morally justifiable.

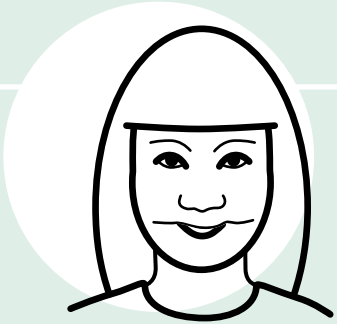
By "imperialism," Weizenbaum means that instrumental thinking has found its way into our everyday lives and shapes our view of the world and ourselves. The achievements of science and technology—such as flying, medicine, and the internet—have convinced us of instrumental reason, such that we have oriented our thinking accordingly: What must I do to be successful? Which photo will get the most likes? Am I good enough for the job, or can a computer do it better than me? According to Weizenbaum, the computer's great success and fascinating performance have led us to develop a "technological mentality" that not only reinforces instrumental thinking, but also assumes that computer is fundamentally the superior tool—above all, superior to people.

Weizenbaum personally rejected this perspective. He argued that it would lead to technology being used in areas where it causes more harm than good. Therefore, he set two boundaries to limit the extent to which technology and AI are and should be used.

The first boundary concerns social problems and interpersonal conflicts, for which he believed there was no "technological fix." For instance, introducing computers into the classroom would not solve learning difficulties by children at school because the underlying issues are usually diverse, including poor supervision, overburdened teachers, troubles at home, bullying by other children, and too much pressure to perform well. Attempting to solve learning difficulties with computers would only mask these problems and potentially even worsen the situation.



The second boundary is a moral one. Weizenbaum declared that there are certain tasks "which computers *ought* not to be made to do, independent of whether computers *can* be made to do them." As an example, he cited the use of AI applications for the psychotherapeutic treatment of patients (i.e., using programs such as ELIZA). Since computers have no biography and no body, they lack own experiences and humanity (see info box "Uncanny Valley"). Another example is an animal experiment in which the brain of a dead cat was connected to a computer, giving the computer the ability to see. For Weizenbaum, this would represent an obscene violation of ethical boundaries.



UNCANNY VALLEY

The term “Uncanny Valley” was coined by Japanese roboticist Masahiro Mori. In 1970, Mori published an essay with this title, proposing that our perception of human-like robots changes abruptly once they reach a certain level of human resemblance. Initially, we would be empathetic toward humanoid robots; however, our sympathies would drop significantly if they were nearly indistinguishable from humans, with only minor differences. The robots would then seem eerie.

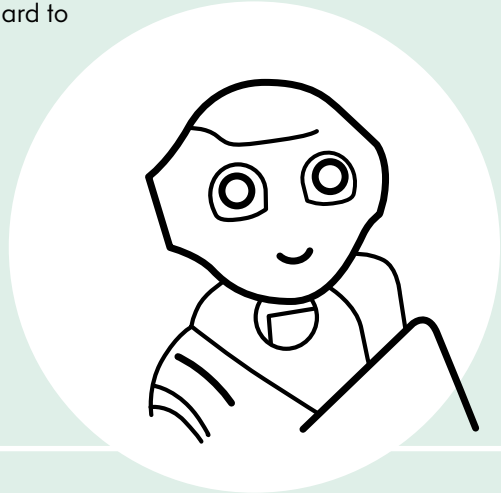
As an example, Mori describes a robot equipped with 29 artificial facial muscles to simulate a human smile for the World Expo in Osaka. At the slightest deviation in the speed of facial expressions, our affection for the robot turns to horror. Other examples of this phenomenon include moving dolls, zombies, amputated body parts, corpses, and overly realistic prosthetics.

Mori concludes that a resemblance of robots to humans

should be avoided, and the focus should be on the robot's function. Glasses, for example, would also not resemble human eyes, even though they serve the same functional purpose of improving vision.

Mori's essay went virtually unnoticed until the early 2000s. It was only with the development of newer robotics and AI that observations of the "Uncanny Valley" effect resurfaced, particularly with regard to care robots.

An "Uncanny Valley" effect is also suspected for "disembodied" robots, such as chatbots. In recent years, various experiments have been conducted to empirically measure and psychologically explain this effect. Mori himself could not provide empirical evidence or an explanation for the phenomenon.



ARTIFICIAL INTELLIGENTSIA

As a professor at the renowned MIT in Cambridge, Joseph Weizenbaum had been an established participant in AI research since the 1960s. However, from the 1970s onward, he was increasingly critical of the field. This background gave him the professional authority that brought significant public attention to his 1976 book "Computer Power and Human Reason." However, this also meant that Weizenbaum's criticism was directed at some of his closest colleagues. As one might imagine, they were not exactly pleased with his criticism.

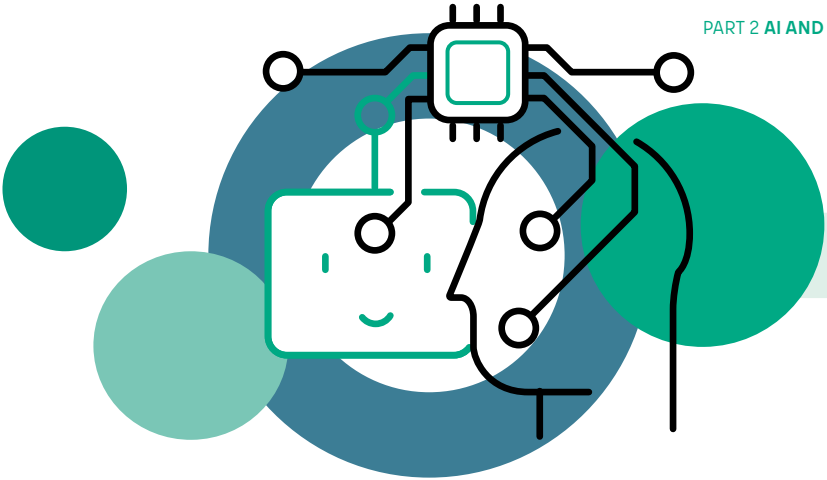
These colleagues included Marvin Minsky, who worked with Weizenbaum at the same institute, John McCarthy, who coined the term "artificial intelligence," and the psychiatrist Kenneth Colby, with whom Weizenbaum had been in contact during his work on ELIZA. Weizenbaum called these and other colleagues from the field

Joseph Weizenbaum (right) with his colleagues Claude Shannon, John McCarthy, and Edward Fredkin (from left to right), 1968
Photo from the image archive of the documentary film "Weizenbaum. Rebel at Work"

of AI research, including Nobel Prize winner Herbert A. Simon, the "Artificial Intelligentsia," that is AI intellectuals. Weizenbaum's mockery was not directed at their scientific merits in computer research, but rather at the predictions these men made in public about the potential of AI in the near future.

By 1955, in the founding document of "Artificial Intelligence," John McCarthy and his colleagues had announced that they could find out within just two months how computers could be made to "solve kinds of problems now reserved for humans, and improve themselves." By the end of the 1960s, Herbert A. Simon predicted that computers would soon be able to explain and simulate the behavior of all people. Then, in 1970, Marvin Minsky predicted that in less than ten years, there would be a superintelligence, computers would take control, and we humans would be lucky if they kept us as pets. Finally, the roboticist Hans Moravec predicted in his book "Mind Children" (1988) that humans would soon be able to upload their consciousness to a computer and, thus, become digitally immortal (see info box "Transhumanism").

Thus far, none of these predictions have come true. Nonetheless, we hear similar predictions from today's representatives of AI research. For example, Sam Altman, CEO of OpenAI, is certain that his company is on the verge of creating a superintelligence. Ray Kurzweil, Google's Technical Director, once answered the question of whether God exists in an interview concerning technological development with the quip, "Well, I would say, not yet." Finally, Geoffrey Hinton, Google's head AI developer for many years, resigned from his post in May 2023 because he believed that AI was becoming an existential threat. This suggests that the speculative future of AI has always lingered between the utopian promise of salvation and the downfall of humanity.



Today, Weizenbaum would probably refer to Altman, Kurzweil, Hinton, and company as the new "artificial intelligentsia." In the past 50 years, AI research has made great progress and shifted from public universities to commercial technology companies. However, the striking similarities in predictions about the potential of AI show that this technology has always had a way of projecting dreams, hopes, and fears of the future.

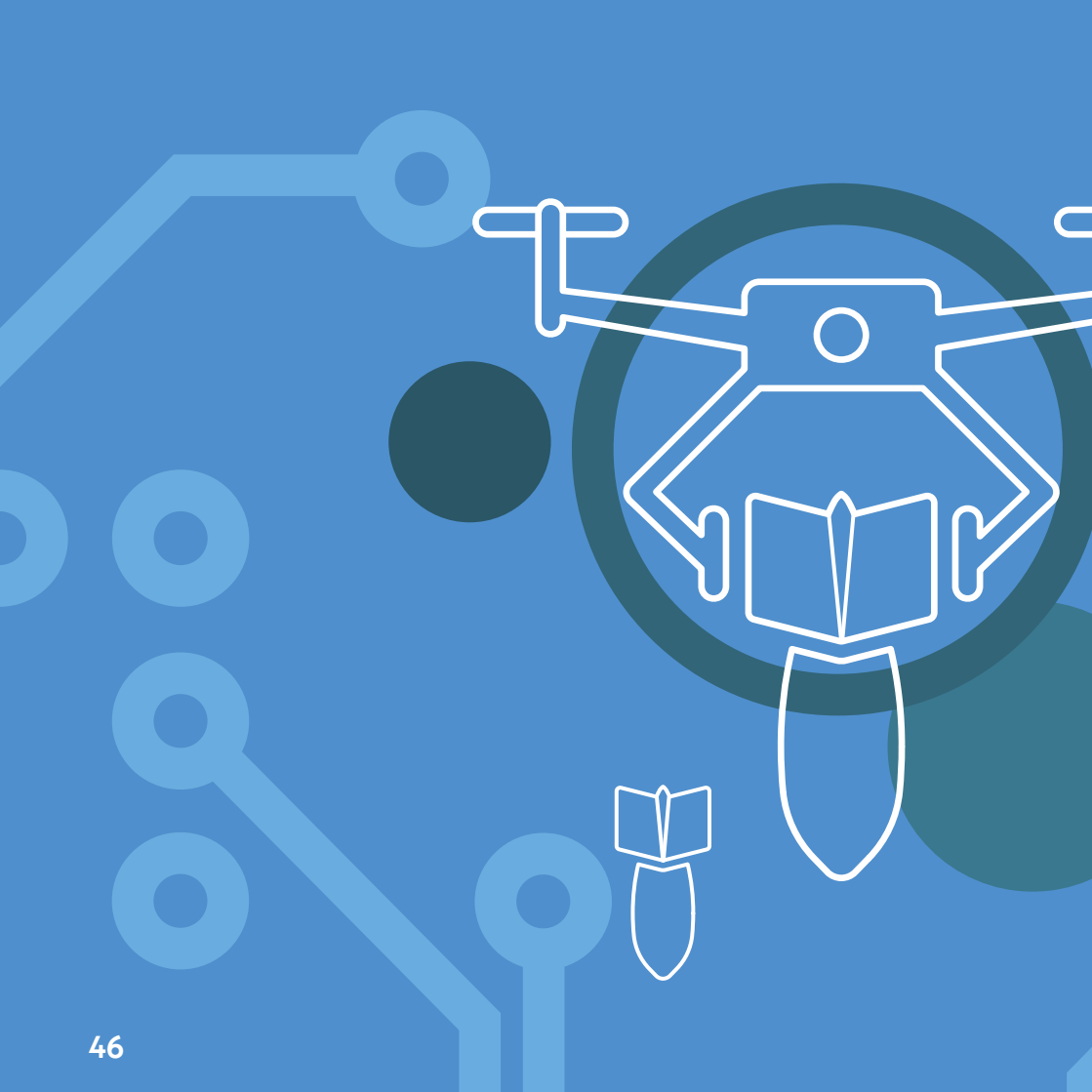
TRANSHUMANISM

The idea that people can transcend beyond themselves and become immortal has a long history. It appears in popular myths, such as the fountain of youth, and in Friedrich Nietzsche's concept of the "Übermensch" ("overman"). A more recent, technology-oriented variant is transhumanism, a philosophical school of thought. The term was coined in 1957 by the British biologist and eugenicist Julian Huxley, brother of writer Aldous Huxley. Since then, it has come to refer to a social movement that strives to expand human capabilities through symbiosis with technology.

According to this futuristic philosophy, humans are "Mängelwesen" ("deficient beings," a term coined by Arnold Gehlen) who overcome their biological limitations through the help of technological progress and eventually achieve immortality. A central figure in this philosophy is the "cyborg," a human-machine hybrid with new abilities due to technical implants. Another possibility concerns brain-computer interfaces that enable direct information transfer, allowing people to acquire new knowledge and skills, influencing their moods, and treating mental illnesses. These interfaces would even enable

"mind uploading," in which the brain, including consciousness and memories, could be copied onto a technical storage system, providing a form of immortality.

While these ideas have only been realized in science fiction stories thus far, legitimate criticism has emerged within this school of thought due to its connection with eugenics. Eugenics aims to improve the gene pool of a population by controlling human genetic material. The field was an essential component of the German National Socialist Party's concept of "racial hygiene."



PART 3
CRITICAL PERSPECTIVES ON AI

MILITARY USE

When the term "artificial intelligence" emerged in the 1950s and research in the field gained momentum, the world was still recovering from the devastating consequences of World War II, the Holocaust, and the atomic bombs dropped by the United States on the Japanese cities of Hiroshima and Nagasaki. Given the destructive potential of new technologies and governments' willingness to use them, many people were skeptical of early developments in computer technology. This skepticism played an important role in the protests against the Vietnam War and the anti-nuclear movement in the 1960s and 1970s. The atomic bomb, in particular, has become a central reference point for AI discourse. For example, in an open letter published in May 2023, renowned AI researchers and computing industry leaders compared the potential danger of AI to that of the atomic bomb. Some researchers have even called for an international supervisory authority for AI comparable to the International Atomic Energy Agency.

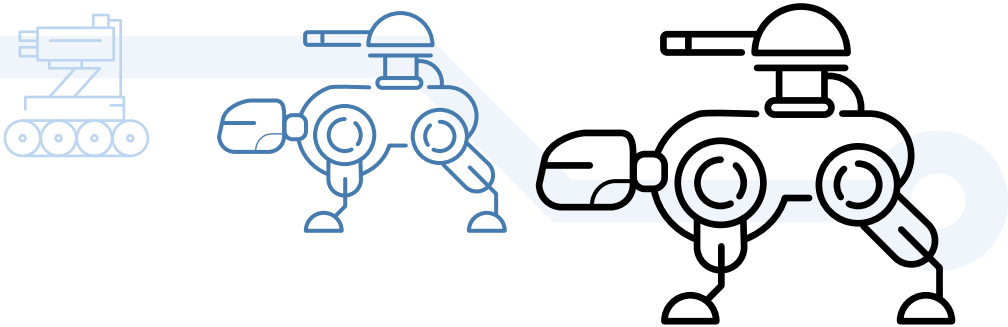
There are two main reasons for the strong connection between AI discourse and the atomic bomb. First, nuclear technology exemplifies the dual-use problem of technological developments that are used for both civilian and military purposes. Nuclear technolo-

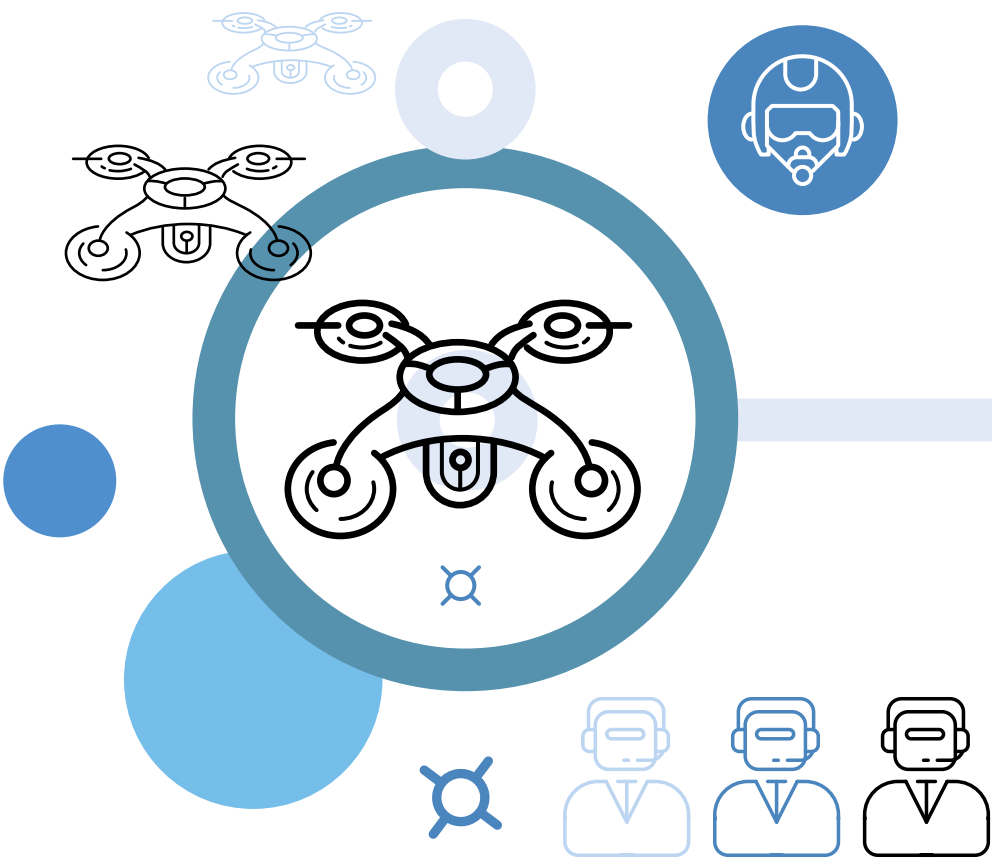
gy is used to generate energy and to produce atomic bombs. Second, this means that researchers and developers must consider the political and moral implications of their work on such technologies. Józef Rotblat, for example, was initially involved in the Manhattan Project, which developed the first atomic bomb. However, he left the project in 1944 when it became apparent that the Nazis would not be able to build their own atomic bomb.

Joseph Weizenbaum recognized that AI is also a dual-use technology and urged scientists to be accountable for its potential consequences. He argued, that it must be assumed that every technology used for civilian purposes will also be used for military purposes. Weizenbaum himself participated in numerous research projects funded by the U.S. Department of Defense at the beginning of his career. The military was particularly interested in developing computers and AI because it hoped to gain an advantage in the Cold War. For Weizenbaum, this entanglement of his work with military interests was ultimately reason enough to distance himself from the research.

Today, AI plays a role not only in civilian applications, such as search algorithms and voice assistants, but also in the military sector. For instance, in the Russian-Ukrainian war that began in 2022, AI has been used to carry out attacks with autonomous kamikaze drones on targets behind enemy lines. In the Gaza war following the Hamas attack on Israel on October 7, 2023, the Israeli military used various AI systems to quickly select buildings to bomb or to identify, target, and kill people. The market for military AI systems has grown for years, and their development is closely linked to border and migration control, where they are often tested (see info box "Technological Testing Grounds"). AI is intended to make war operations more "efficient" and less costly, but this could lower the inhibition threshold for using weapons, thereby jeopardizing regional and international stability.

Finally, there is also the question of responsibility. In 2018, UN Secretary-General António Guterres declared that autonomous weapons authorized to kill without human control are politically unacceptable and morally reprehensible. At the end of 2023, the UN General Assembly passed the first resolution on autonomous weapons systems. However, they have not yet been banned. Similarly, the European Union has also not excluded military AI systems from its Artificial Intelligence Act. These details suggest that the arms race is already in full swing.





TECHNOLOGICAL TESTING GROUNDS

Since the September 11, 2001 terrorist attacks, people crossing national borders have increasingly been viewed as security threats, not only in the United States. Consequently, border and migration control has become a primary application area for new technologies such as AI, which is why some talk of "smart borders." Even before people cross the border, AI systems are used to classify risks based on surveillance and social media data. At the border, sensors and cameras capture biometric facial data or detect unusual behavior. After entering the country, migrants are often monitored, and AI is used to process asylum applications and manage deportations.

Governments and border authorities, such as Frontex in the EU and the Federal Office for Migration and Refugees (BAMF) in Germany, are investing heavily in AI technologies, most of which are developed by private tech companies like Palantir and Clearview AI. The type and scope of the use of AI and other technologies in the context of

French border police drone used to detect rubber boats carrying refugees across the English Channel, 2019
Photo by Denis Charlet/AFP via Getty Images

border and migration control would often be illegal in the civilian sector. Therefore, lawyer Petra Molnar refers to borders as "technological testing grounds," where new technologies are tried out before being adapted for use in other sectors.

Although these technologies exacerbate power imbalances and structural discrimination, significantly impacting the fundamental rights and lives of many, such consequences are usually concealed from the public.

ECOLOGICAL CONSEQUENCES

Even though the AI systems we interact with daily are virtual programs, their use has a real environmental impact. This is primarily due to the resources required to manufacture computers. Computing devices contain numerous raw materials, some of which are rare, including iron, copper, aluminum, lithium, nickel, gold, zinc, cobalt, and coltan. These materials are mined in different parts of the world, processed, and transported over long distances before being used to make our devices smaller, lighter, faster, more accessible, and more powerful. This process requires a lot of water and energy, which exacerbates the environmental impact.

Consider lithium, for example. This light metal is used for rechargeable batteries and is therefore found in every laptop and smartphone. It is mined by blasting and drilling into the earth, then pumping salty groundwater to the surface, where the water evaporates in artificial pools, leaving behind lithium compounds. The world's largest lithium deposit is in the Atacama Desert, in the "lithium triangle" between Bolivia, Chile, and Argentina. Large quantities of water are required for mining there, which is causing the region's drinking water supplies to dwindle and become saline. Furthermore, lithium is not a renewable resource, meaning that global lithium

reserves could be depleted in just a few decades. The increasing use of electric cars has drastically increased the already high demand for lithium. While a smartphone contains around one gram of lithium, an electric car contains several kilograms. Despite increased efforts to recycle lithium, most electric waste still ends up in large landfills in Ghana and Pakistan.

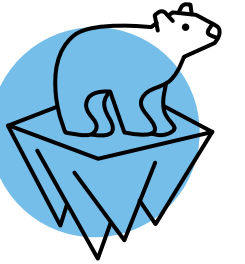
The situation is similar for other metals. The mining of tin, which is used for soldering electronics, is rendering large areas of Indonesia barren and uninhabitable. Forests are cut down to make way for open-pit mining, leading to droughts, floods, and landslides. Rivers are polluted when tin is washed, and corals are destroyed by offshore mining. Many workers suffer from severe health issues. In the Democratic Republic of the Congo (DRC), large areas of rainforest are being cleared to mine coltan, which is used in microcapacitors, threatening the habitat of numerous animals. The mining process also releases toxic sulfuric acids that harm the environment and endanger workers. Due to the unstable situation in the DRC, coltan mining is also associated with violence and human rights violations (see info box “Conflict Minerals”).

These issues are not exclusive to AI, but rather stem from the broader ecological impact of computers in general. However, when assessing AI’s potential for climate and environmental protection, various concerns must be addressed. For instance, AI requires far more electricity than traditional search engines. On a larger scale, as with ChatGPT, this increases greenhouse gas emissions and water consumption associated with cooling the servers.

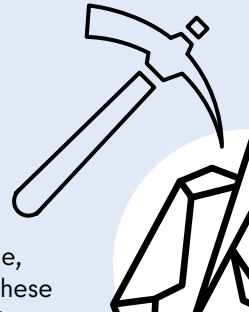
Joseph Weizenbaum was convinced that technology alone could not save us from climate change. This would require fundamental changes in society, which AI cannot bring about. A central

Lithium mining in the Chilean Atacama Desert, 2022
Photo by John Moore/Getty Images

element of this change is a “resistance to greed” in how we live, do business, and consume. Although AI may seem harmless due to its digital nature, its environmental footprint is far from hidden in the real world.



CONFLICT MINERALS



Some of the raw materials used to build computers and other electronics are referred to as “conflict minerals” because they are mined in regions affected by war and violence. These minerals include tin, tantalum (coltan), tungsten, and gold. Together, they are also referred to as “3TG conflict minerals.” Similarly, diamonds, precious woods, raw materials used to make drugs, cocoa, and petroleum are sometimes mined or cultivated in conflict zones.

The extraction, trade, and processing of these conflict minerals often subjects people in affected regions to systematic violence and human rights violations, with miners forced to work under inhumane conditions. Furthermore, the trade of these raw materials can intensify or even trigger armed conflicts, which may be fought over (“resource conflicts”) or be used to finance ongoing wars. Therefore, an abundance of natural

resources can negatively impact a country, which is why there is also talk of a "resource curse" in this context.

One of the most important conflict minerals for the digital and electronics industry is coltan, from which the metal tantalum is extracted. The largest coltan deposits are in the Democratic Republic of the Congo (DRC), which is also the leading producer of many other minerals. Paramilitary groups control these coltan mines and use the proceeds from the trade to finance civil war in the DRC. Mining is also associated with violations of human rights and international law, such as child labor and the pollution of drinking water sources. Thus, the existing political and social instability that can be traced back to the Belgian colonial era is exacerbated by coltan mining and trade. Previous attempts at regulation, such as the U.S. Dodd-Frank Act of 2010 and the EU Conflict Minerals Regulation of 2017, have been difficult to implement due to illegal trade and the associated lack of traceability in supply chains.

SOCIETAL RISKS

In 1985, Joseph Weizenbaum was asked in an interview whether computers had the potential to change society or if they were more likely to reinforce existing power structures. As personal computers became more common in households at the time, there was widespread expectation that computers could revolutionize society for the better. However, Weizenbaum responded to the interviewer's question by describing computers as "a fundamentally conservative force," that made it easier for society to remain the same. As an example, he referred to the banking industry, which would have had to decentralize to accommodate a growing population. However, thanks to computers, this decentralization was unnecessary.

Weizenbaum's perspective encourages us to shift our focus from the potential social changes that AI could bring about to the question of whether AI could also perpetuate or even reinforce existing power imbalances and injustices. This perspective reveals that generative AI actually reproduces established stereotypes and forms of discrimination. Since these systems are typically trained on internet data, they often "learn" the biases inherent in the content. For instance, in 2016, Microsoft's Twitter bot, "Tay," caused a stir when it learned

inappropriately from interacting with Twitter users. It posted so many racist and misogynistic posts in such a short period of time that the bot was taken down just one day after its launch.

Additionally, studies have shown that AI-generated content and decisions are often biased to the disadvantage of already disadvantaged groups. This is evident when AI-generated images show women in traditional gender roles or portray Black men as dangerous. By contrast, white men appear primarily in contexts of success and competence. These distortions become even more problematic when they have real-world consequences for people. For instance, certain groups are more likely to be suspected than others in the context of predictive policing, even though there is no basis for this suspicion. Furthermore, women receive fewer loans from banks than men despite having equal qualifications, and people with certain surnames are less likely to be hired through AI-supported processes.

In response to this criticism about discrimination, developers have taken action. In some cases, they have attempted to solve these problems technically by adding invisible "co-prompts" to generative AI to compensate for biases. They have also found that diversifying the AI training data and removing problematic content would be effective. Consequently, numerous data workers, especially from the Global South, are being employed for these tasks. However, this requires a lot of human data work, which is often poorly paid, to power the performance of AI (see info box "Data Work").

This data work can adapt AI systems to avoid forms of disinformation, hate speech, and illegal content. However, AI itself can become part of the problem because it can still easily produce this exact content. As generative AI improves, so does the ability to fake content. Furthermore, when it becomes impossible to distinguish reality from fiction, all information becomes suspect.

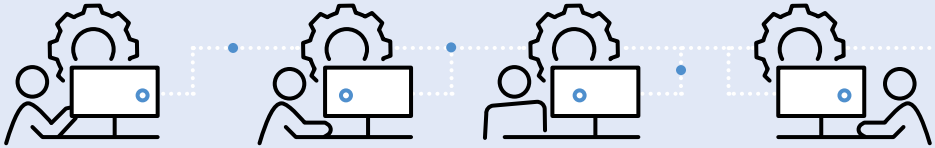
This can cause major problems for a democracy based on consensus.

For Weizenbaum, weighing this technology's concrete benefits against its possible risks was important. Regarding generative AI, the question is whether we should put our democracy at risk just to have texts and images created at the touch of a button. Weizenbaum would likely have answered no and instead recommended not developing such technology further to address more important problems.

DATA WORK

Images of robots, wired brains, and advanced networks are often used to represent AI—also in this brochure. These systems are expected to process data, recognize patterns, and learn independently, suggesting that human work is dispensable. However, there are thousands of data workers behind AI systems, performing the labor of generating, categorizing, and checking training data.

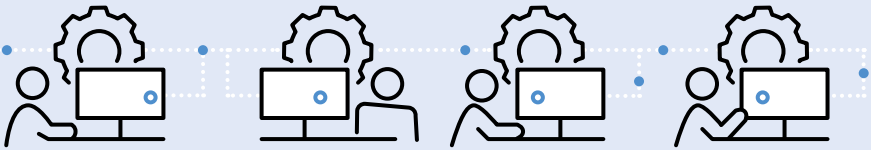
Because some of the tasks are small-scale, this work is often referred to as “click work” or “micro work.” Autonomous driving, for example, requires a large amount of data from road markings, signs, vehicles, people, and animals in order to develop an AI that can react accordingly. To meet the demand for such a large quantity of labor, underpaid work often forms the basis of AI development. One way to obtain



such cheap labor is through crowdwork platforms, such as Amazon Mechanical Turk, where self-employed workers perform tasks often for a few cents.

Another option is to outsource the labor to subcontractors in the Global South. Samasource, a well-known company based in San Francisco, hires data workers in Kenya, Pakistan, and India to work for companies such as OpenAI. Samasource pays these workers about two US dollars per hour. Thus, a global division of labor is emerging in the AI industry, with big tech companies from Silicon Valley benefiting from cheap labor and a lack of labor protections in the Global South.

Overall, AI systems will not reduce the demand for human labor. Training and correcting generative AI systems will continue to require enormous amounts of data. Furthermore, given AI's current success, demand is likely to increase. The resulting labor market urgently needs regulation to prevent exploitation and precarious working conditions.



PART 4
GUIDING QUESTIONS OF AI CRITICISM

A historical review of the debates surrounding computers and AI reveals recurring patterns. That is, many of the issues discussed today have been prominent for years. This makes it difficult to maintain a critical perspective. After all, we are witnessing first-hand how quickly technological development is progressing. We see how computers are becoming better and more powerful, with things we could hardly have imagined suddenly appearing possible. Our admiration for this rapid development can inspire hope for a better future, but it can also be intimidating, unsettling, and overwhelming. This makes it all the more important to critically monitor these developments, occasionally taking a step back to place the myths and warnings in their historical context in order to assess the direction in which we are heading. The following key questions are intended to help with this.

IMAGE OF MAN

When we hear about the performance and potential of AI, how are humans portrayed in this context?

Is it made clear that this technology is meant to support humans, or are humans and AI pitted against each other? For example, are humans portrayed as limited, mortal, and error-prone beings in order to emphasize the capabilities of AI? Are they even being ridiculed? What kind of world are we heading toward where humans are considered "inferior machines"? To what extent is this view already influencing our self-image and shaping expectations of ourselves and others? How can we change this?

INSTRUMENTAL THINKING

When AI is presented as a solution to a specific problem, what kind of problem is it?

Is the problem technical or social in nature? Does it involve technical errors, challenges, and obstacles, or does it relate to social injustices, discriminatory structures, and interpersonal conflicts? If it is a social problem, what were its original causes? Do these causes disappear with the application of AI? Which problems remain, and which are newly introduced? Could the problem to be solved actually be worsened? What problems already exist as a result of AI usage? Are they of a technical or social nature? And how can we best tackle them?

SOCIAL INEQUALITIES

When it comes to the potential of AI, who benefits from this potential, and who cannot?

Who are AI systems being developed for, who can afford them, and who decides on their use? Who, conversely, has no agency and ultimately suffers the consequences of its application? Who is pressured by AI? Who is exploited to make it work? Who will lose their jobs to this automation, and who has no option but to comply with its demands? Who is AI threatening or discriminating against? Who is being monitored and controlled? What power imbalances and injustices are being perpetuated or even reinforced? And what can we do about it?

PART 5
BIOGRAPHICAL OVERVIEW: JOSEPH WEIZENBAUM

Joseph Weizenbaum, 2006
Portrait by Erik Weiss

1923	Born on January 8 in Berlin to Jechiel “Harry” Weizenbaum and Henriette (née Orman) following two older brothers: Leo (from his father’s first marriage) and Heinrich W. “Heinz” (later Henry Francis Sherwood)
1923–1936	Raised in Berlin, living at Gendarmenmarkt in Berlin-Mitte
1934	Attended the Luisenstädtisches Realgymnasium before being forced to transfer to the boys’ school of the Jewish community on Kaiserstraße in 1935
1936	Escaped from National Socialism to the United States, emigration by ship from Bremerhaven to New York, then Detroit, where he attended high school (graduating in 1940)
1941–1950	Studied mathematics at Wayne University in Detroit
1942–1946	Interruption of studies, military service in the meteorological department of the US Army Air Forces
1944	Acceptance of US citizenship
~1946	First marriage to Thelma “Selma” Goode, with whom he had a son, David, before divorcing around 1949

1950-1953 Research assistant in the Differential Analyzer Laboratory and senior programmer at the Digital Computer Laboratory in the Department of Mathematics at Wayne University in Detroit

1952 Second marriage to Ruth Manes, with whom he had four daughters: Pm, Sharon, Miriam, and Naomi

1953-1955 Programmer at the Computer Control Company at the US Navy's Point Mugu missile base and in the Computer Division of Bendix Aviation in Los Angeles, which developed ground-to-air missiles for the military, among other things

1956-1963 Programmer at General Electric's Computer Development Laboratory in Sunnyvale, California; involved in the first computer database system ERMA for Bank of America, development of the programming language SLIP ("Symmetric List Processor")

1963 Visiting associate professor at the Massachusetts Institute of Technology (MIT) in Cambridge; collaborated on the newly founded Project MAC, largely funded by the US Department of Defense

1964 Appointed associate professor in the Department of Electrical Engineering at MIT

1965-1968 Development of ELIZA (first published in 1966)

1969	Founding member of the Union of Concerned Scientists
1970	Appointed full professor at MIT
1972-1973	Sabbatical and research stays at the Center for Advanced Study in the Behavioral Sciences (CASBS) at Stanford University and at Harvard University
1976	Publication of his main book “Computer Power and Human Reason: From Judgment to Calculation”
1981	Co-founder of Computer Professionals for Social Responsibility (CPSR)
1984	Co-founder of Computer Professionals for Peace and Social Responsibility (FIfF) in Germany
1988	Emeritus at MIT, recipient of the Norbert Wiener Award of the CPSR
1992	Honorary award from the “Computer Science and Society” department of the German Informatics Society (GI)
1994	Humboldt Award from the Alexander von Humboldt Foundation, guest residency at the Albert Ludwig University of Freiburg
1996	Moved to Berlin, first to Ludwigkirchplatz in Berlin-Wilmersdorf and later to Nikolaiviertel in Berlin-Mitte

1998	Honorary doctorate from the Faculty of Mathematics and Computer Science at the University of Bremen
2001	Awarded the Grand Cross of the Order of Merit of the Federal Republic of Germany
2003	Honorary doctorate from the Department of Computer Science at the University of Hamburg, honorary membership of the GI
2007	Documentary film about Weizenbaum, "Rebel at Work," produced by Silvia Holzinger and Peter Haas
2008	Died on March 5 in Ludwigsfelde-Gröben (Brandenburg), later buried in the Jewish cemetery in Berlin-Weißensee

Digital exhibition on the life and work of Joseph Weizenbaum (in German)

RECOMMENDED LITERATURE

RECOMMENDED LITERATURE

Bender, Emily M./Hanna, Alex (2025). The AI con: How to fight big tech's hype and create the future we want. New York, NY: HarperCollins.

Critical examination of the contemporary hype surrounding AI. In light of the social harms it causes, such as labor exploitation and algorithmic bias, the authors call for democratic oversight of AI development and deployment.

Ciston, Sarah/Berry, David M./Hay, Anthony C./Marino, Mark C./Millican, Peter/Shrager, Jeff/Schwarz, Arthur I./Weil, Peggy (2026). Inventing ELIZA: How the first chatbot shaped the future of AI. Cambridge, MA: The MIT Press.

The first comprehensive analysis of Joseph Weizenbaum's chatbot ELIZA through the lens of critical code studies. Based on extensive archival research, the book traces the development of ELIZA and reveals the previously unknown innovations behind the chatbot.

Crawford, Kate (2021). Atlas of AI: Power, politics, and the planetary costs of Artificial Intelligence. New Haven, London: Yale University Press.

U.S. researcher Kate Crawford, one of the most important contemporary AI experts, provides a detailed overview of the social consequences of the production, development, and use of AI systems.

Marx, Paris (Host) (since 2020). Tech won't save us. https://www.techwontsave.us/

Podcast that critically examines technology and big tech companies, highlighting political, economic, and ethical issues. Each episode blends expert interviews, historical context, and case studies to reveal how power structures and cultural narratives influence our relationship with technology.

Molnar, Petra (2024). The walls have eyes: Surviving migration in the age of artificial intelligence. New York, NY: The New Press.

Lawyer and anthropologist Petra Molnar reports on her research into the use of technologies such as AI at the borders of Europe and the USA, among others, and the devastating consequences for people trying to cross these borders.

Morozov, Evgeny (2013). To save everything, click here: The folly of technological solutionism. New York, NY: PublicAffairs.

Similar to Joseph Weizenbaum, Evgeny Morozov criticizes the widespread assumption that technology can solve social and political problems, warning against the dangers of "solutionism."

Pörksen, Bernhard (2023). The image of man in artificial intelligence: A conversation with Joseph Weizenbaum. Weizenbaum Journal of the Digital Society, 3(3). https://doi.org/10.34669/WI.WJDS/3.3.4

Translated version of the 2000 interview that media scholar Bernhard Pörksen conducted with Weizenbaum. The discussion covers topics such as the computer metaphor, the promethean shame, and the unavoidable individuality of the living.

Sharma, Sarah (2026). Insufferable tools: Feminism against big tech. Durham, NC: Duke University Press.

Critique of how big tech companies shape everyday life transforming users into "tools" within patriarchal societies. It proposes feminist techno-politics aimed at reclaiming agency and imagining futures beyond such exploitative systems.

Weizenbaum, Joseph (1977). Computer power and human reason: From judgment to calculation. San Francisco, CA: W. H. Freeman and Company.

The central work by Joseph Weizenbaum, in which he takes a detailed and critical look at his own field of research. His most important criticisms of society's approach to computers and AI are already laid out here.

Weizenbaum, Joseph/Wendt, Gunna (2015). Islands in the cyberstream: Seeking havens of reason in a programmed society. Sacramento, CA: Litwin Books.

A collection of interviews in which Weizenbaum discusses a wide range of topics with Gunnar Wendt. Topics include "the illusion of powerlessness," "natural science as religion," and "computers and schools."

Weizenbaum Institute (2024): Pionier, Skeptiker, Aktivist: Wendepunkte im Leben von Joseph Weizenbaum. Eine digitale Ausstellung, https://jw.weizenbaum-institut.de (in German)

This multimedia exhibition traces the most important stages of Weizenbaum's life and work, from Berlin to Detroit to Cambridge and back to Berlin. It features rare photographs, an ELIZA simulator, and original essays.

Imprint

Author
Christian Strippel

Collaboration
Moritz Buchner, Magnus Rust,
Alexandra Keiner,
Diana Ignatovich

Design
Bergsee, blau

Berlin, May 2026

Publisher
Weizenbaum-Institut e.V.
Hardenbergstraße 32, 10623 Berlin
www.weizenbaum-institut.de

funded by the

Federal Ministry
of Research, Technology
and Space

and

Berliner Landeszentrale
für politische Bildung

Hardenbergstraße 22-24, 10623 Berlin
Opening hours of the visitor center:
Mon, Wed, Thu and Fri 10 am – 6 pm

and Revaler Straße 29, 10245 Berlin
Opening hours of the visitor center:
Tue, Wed and Fri 10 am – 6 pm

www.berlin.de/politische-bildung