Weizenbaum’s nightmares: how the inventor of the first chatbot turned against AI

Publication Date: 2023-07-25

Author: Ben Tarnoff

Section: Technology

Tags: The long read, Artificial intelligence (AI), Chatbots, ChatGPT, Vietnam war, MIT - Massachusetts Institute of Technology, Internet safety, features

Article URL: https://www.theguardian.com/technology/2023/jul/25/joseph-weizenbaum-inventor-eliza-chatbot-turned-against-artificial-intelligence-ai

In 1966, an MIT professor named Joseph Weizenbaum created the first chatbot. He cast it in the role of a psychotherapist. A user would type a message on an electric typewriter connected to a mainframe. After a moment, the “psychotherapist” would reply. User: Men are all alike. Chatbot: IN WHAT WAY User: They’re always bugging us about something or other. Chatbot: CAN YOU THINK OF A SPECIFIC EXAMPLE User: Well, my boyfriend made me come here. Chatbot: YOUR BOYFRIEND MADE YOU COME HERE User: He says I’m depressed much of the time. Chatbot: I AM SORRY TO HEAR YOU ARE DEPRESSED Weizenbaum published this sample exchange in a journal article that explained how the chatbot worked. The software was relatively simple. It looked at the user input and applied a set of rules to generate a plausible response. He called the program Eliza, after Eliza Doolittle in Pygmalion. The cockney flower girl in George Bernard Shaw’s play uses language to produce an illusion: she elevates her elocution to the point where she can pass for a duchess. Similarly, Eliza would speak in such a way as to produce the illusion that it understood the person sitting at the typewriter. “Some subjects have been very hard to convince that Eliza (with its present script) is not human,” Weizenbaum wrote. In a follow-up article that appeared the next year, he was more specific: one day, he said, his secretary requested some time with Eliza. After a few moments, she asked Weizenbaum to leave the room. “I believe this anecdote testifies to the success with which the program maintains the illusion of understanding,” he noted. Eliza isn’t exactly obscure. It caused a stir at the time – the Boston Globe sent a reporter to go and sit at the typewriter and ran an excerpt of the conversation – and remains one of the best known developments in the history of computing. More recently, the release of ChatGPT has renewed interest in it. In the last year, Eliza has been invoked in the Guardian, the New York Times, the Atlantic and elsewhere. The reason that people are still thinking about a piece of software that is nearly 60 years old has nothing to do with its technical aspects, which weren’t terribly sophisticated even by the standards of its time. Rather, Eliza illuminated a mechanism of the human mind that strongly affects how we relate to computers. Early in his career, Sigmund Freud noticed that his patients kept falling in love with him. It wasn’t because he was exceptionally charming or good-looking, he concluded. Instead, something more interesting was going on: transference. Briefly, transference refers to our tendency to project feelings about someone from our past on to someone in our present. While it is amplified by being in psychoanalysis, it is a feature of all relationships. When we interact with other people, we always bring a group of ghosts to the encounter. The residue of our earlier life, and above all our childhood, is the screen through which we see one another. This concept helps make

sense of people's reactions to Eliza. Weizenbaum had stumbled across the computerised version of transference, with people attributing understanding, empathy and other human characteristics to software. While he never used the term himself, he had a long history with psychoanalysis that clearly informed how he interpreted what would come to be called the "Eliza effect". As computers have become more capable, the Eliza effect has only grown stronger. Take the way many people relate to ChatGPT. Inside the chatbot is a "large language model", a mathematical system that is trained to predict the next string of characters, words, or sentences in a sequence. What distinguishes ChatGPT is not only the complexity of the large language model that underlies it, but its eerily conversational voice. As Colin Fraser, a data scientist at Meta, has put it, the application is "designed to trick you, to make you think you're talking to someone who's not actually there". But the Eliza effect is far from the only reason to return to Weizenbaum. His experience with the software was the beginning of a remarkable journey. As an MIT professor with a prestigious career, he was, in his words, a "high priest, if not a bishop, in the cathedral to modern science". But by the 1970s, Joseph Weizenbaum had become a heretic, publishing articles and books that condemned the worldview of his colleagues and warned of the dangers posed by their work. Artificial intelligence, he came to believe, was an "index of the insanity of our world." Today, the view that artificial intelligence poses some kind of threat is no longer a minority position among those working on it. There are different opinions on which risks we should be most worried about, but many prominent researchers, from Timnit Gebru to Geoffrey Hinton – both ex-Google computer scientists – share the basic view that the technology can be toxic. Weizenbaum's pessimism made him a lonely figure among computer scientists during the last three decades of his life; he would be less lonely in 2023. There is so much in Weizenbaum's thinking that is urgently relevant now. Perhaps his most fundamental heresy was the belief that the computer revolution, which Weizenbaum not only lived through but centrally participated in, was actually a counter-revolution. It strengthened repressive power structures instead of upending them. It constricted rather than enlarged our humanity, prompting people to think of themselves as little more than machines. By ceding so many decisions to computers, he thought, we had created a world that was more unequal and less rational, in which the richness of human reason had been flattened into the senseless routines of code. Weizenbaum liked to say that every person is the product of a particular history. His ideas bear the imprint of his own particular history, which was shaped above all by the atrocities of the 20th century and the demands of his personal demons. Computers came naturally to him. The hard part, he said, was life. * * * What it means to be human – and how a human is different from a computer – was something Weizenbaum spent a lot of time thinking about. So it's fitting that his own humanity was up for debate from the start. His mother had a difficult labour, and felt some disappointment at the result. "When she was finally shown me, she thought I was a bloody mess and hardly looked human," Weizenbaum later recalled. "She couldn't believe this was supposed to be her child." He was born in 1923, the youngest son of an assimilated, upper-middle class Jewish family in Berlin. His father, Jechiel, who had emigrated to Germany from Galicia, which spanned what is now south-eastern Poland and western Ukraine, at the age of 12, was an accomplished furrier who had acquired a comfortable foothold in society, a nice apartment, and a much younger Viennese wife (Weizenbaum's mother). From the start, Jechiel treated his son with a contempt that would haunt Weizenbaum for the rest of his life. "My father was absolutely convinced that I was a worthless moron, a complete fool, that I would never become anything," Weizenbaum later told the documentary film-makers Peter Haas and Silvia Holzinger. By the time he was old enough to make memories, the Nazis were everywhere. His family lived near a bar frequented by Hitler's paramilitaries, the SA, and sometimes he would see people getting dragged inside to be beaten up in the backroom. Once, while he was out with his nanny, columns of armed communists and Nazis lined up and started shooting at each other. The nanny pushed him under a parked car until the bullets stopped flying. Shortly after Hitler became chancellor in 1933, the government passed a law that severely restricted the number of Jews in public schools. Weizenbaum had to transfer to a Jewish boys' school. It was here that he first came into contact with the Ostjuden: Jews from eastern Europe, poor, dressed in rags, speaking Yiddish. To Weizenbaum, they may as well have come from Mars. Nevertheless, the time he spent with them gave him what he later described as "a new feeling of camaraderie", as well as a "sensitivity for oppression". He became deeply attached to one of his classmates in particular. "If fate had been different, I would have developed a homosexual love for this boy," he later said. The boy "led me into his world", the world of the Jewish ghetto around Berlin's Grenadierstrasse. "They had nothing, owned nothing, but somehow supported each other," he recalled. One day, he brought the boy back to his family's apartment. His father, himself once a poor Jewish boy from eastern Europe, was disgusted and furious. Jechiel was very proud, Weizenbaum remembered – and he had reason to be, given the literal and figurative distances he had travelled from the shtetl. Now his son was bringing the shtetl back into his home. Alienated from his parents, richer than his classmates, and a Jew in Nazi Germany: Weizenbaum felt comfortable nowhere. His instinct, he said, was always to "bite the hand that fed me", to provoke the paternal figure, to be a pain in the backside. And this instinct presumably proceeded from the lesson he learned from his father's hostility toward him and bigotry toward the boy he loved: that danger could lie within one's home, people, tribe. * * * In 1936, the family left Germany suddenly, possibly because Jechiel had slept with the girlfriend of an SA member. Weizenbaum's aunt owned a bakery in Detroit, so that's where they went. At 13, he found himself 4,000 miles from everything he knew. "I was very, very lonely," he recalled. School became a refuge from reality – specifically algebra, which didn't require English, which he didn't speak at first. "Of all the things that one could study," he later said, "mathematics seemed by far the easiest. Mathematics is a game. It is entirely abstract." In his school's metalworking class, he learned to operate a lathe. The experience brought him out of his brain and into his body. About

70 years later, he looked back on the realisation prompted by this new skill: that intelligence “isn’t just in the head but also in the arm, in the wrist, in the hand”. Thus, at a young age, two concepts were in place that would later steer his career as a practitioner and critic of AI: on the one hand, an appreciation for the pleasures of abstraction; on the other, a suspicion of those pleasures as escapist, and a related understanding that human intelligence exists in the whole person and not in any one part. In 1941, Weizenbaum enrolled at the local public university. Wayne University was a working-class place: cheap to attend, filled with students holding down full-time jobs. The seeds of social consciousness that had been planted in Berlin started to grow: Weizenbaum saw parallels between the oppression of Black people in Detroit and that of the Jews under Hitler. This was also a time of incandescent class struggle in the city – the United Auto Workers union won its first contract with Ford the same year that Weizenbaum entered college. Weizenbaum’s growing leftwing political commitments complicated his love of mathematics. “I wanted to do something for the world or society,” he remembered. “To study plain mathematics, as if the world were doing fine, or even didn’t exist at all – that’s not what I wanted.” He soon had his chance. In 1941, the US entered the second world war; the following year, Weizenbaum was drafted. He spent the next five years working as a meteorologist for the Army Air corps, stationed on different bases across the US. The military was a “salvation”, he later said. What fun, to get free of his family and fight Hitler at the same time. While home on furlough, he began a romance with Selma Goode, a Jewish civil rights activist and early member of the Democratic Socialists of America. Before long they were married, with a baby boy, and after the war Weizenbaum moved back to Detroit. There, he resumed his studies at Wayne, now financed by the federal government through the GI Bill. Then, in the late 1940s, the couple got divorced, with Goode taking custody of their son. “That was incredibly tragic for me,” Weizenbaum later said. “It took me a long time to get over it.” His mental state was forever unsteady: his daughter Pm – pronounced “Pim” and named after the New York leftwing daily newspaper PM – told me that he had been hospitalised for anorexia during his time at university. Everything he did, he felt he did badly. In the army he was promoted to sergeant and honourably discharged; nonetheless, he left convinced that he had somehow hindered the war effort. He later attributed his self-doubt to his father constantly telling him he was worthless. “If something like that is repeated to you as a child, you end up believing it yourself,” he reflected. In the wake of the personal crisis produced by Selma’s departure came two consequential first encounters. He went into psychoanalysis and he went into computing. In those days, a computer, like a psyche, was an interior. “You didn’t go to the computer,” Weizenbaum said in a 2010 documentary. “Instead, you went inside of it.” The war had provided the impetus for building gigantic machines that could mechanise the hard work of mathematical calculation. Computers helped crack Nazi encryption and find the best angles for aiming artillery. The postwar consolidation of the military-industrial complex, in the early days of the cold war, drew large sums of US government money into developing the technology. By the late 1940s, the fundamentals of the modern computer were in place. But it still wasn’t easy to get one. So one of Weizenbaum’s professors resolved to build his own. He assembled a small team of students and invited Weizenbaum to join. Constructing the computer, Weizenbaum grew happy and purposeful. “I was full of life and enthusiastic about my work,” he remembered. Here were the forces of abstraction that he first encountered in middle-school algebra. Like algebra, a computer modelled, and thereby simplified, reality – yet it could do so with such fidelity that one could easily forget that it was only a representation. Software also imparted a sense of mastery. “The programmer has a kind of power over a stage incomparably larger than that of a theatre director,” he later said in the 2007 documentary Rebel at Work. “Bigger than that of Shakespeare.” About this time, Weizenbaum met a schoolteacher named Ruth Manes. In 1952, they married and moved into a small apartment near the university. She “couldn’t have been further from him culturally”, their daughter Miriam told me. She wasn’t a Jewish socialist like his first wife – her family was from the deep south. Their marriage represented “a reach for normalcy and a settled life” on his part, Miriam said. His political passions cooled. By the early 1960s, Weizenbaum was working as a programmer for General Electric in Silicon Valley. He and Ruth were raising three daughters and would soon have a fourth. At GE, he built a computer for the Navy that launched missiles and a computer for Bank of America that processed cheques. “It never occurred to me at the time that I was cooperating in a technological venture which had certain social side effects which I might come to regret,” he later said. * * * In 1963, the prestigious Massachusetts Institute of Technology called. Would he like to join the faculty as a visiting associate professor? “That was like offering a young boy the chance to work in a toy factory that makes toy trains,” Weizenbaum remembered. The computer that Weizenbaum had helped build in Detroit was an ogre, occupying an entire lecture hall and exhaling enough heat to keep the library warm in winter. Interacting with it involved a set of highly structured rituals: you wrote out a program by hand, encoded it as a pattern of holes on punch cards, and then ran the cards through the computer. This was standard operating procedure in the technology’s early days, making programming fiddly and laborious. MIT’s computer scientists sought an alternative. In 1963, with a $2.2m grant from the Pentagon, the university launched Project MAC – an acronym with many meanings, including “machine-aided cognition”. The plan was to create a computer system that was more accessible and responsible to individual needs. To that end, the computer scientists perfected a technology called “time-sharing”, which enabled the kind of computing we take for granted today. Rather than loading up a pile of punch cards and returning the next day to see the result, you could type in a command and get an immediate response. Moreover, multiple people could use a single mainframe simultaneously from individual terminals, which made the machines seem more personal. With time-sharing came a new type of software. The programs that ran on MIT’s system included those for sending messages from one user to another (a precursor of email), editing text (early word processing) and searching a database with 15,000 journal articles (a primitive JSTOR). Time-sharing also changed how

people wrote programs. The technology made it possible “to interact with the computer conversationally,” Weizenbaum later recalled. Software development could unfold as a dialogue between programmer and machine: you try a bit of code, see what comes back, then try a little more. Weizenbaum wanted to go further. What if you could converse with a computer in a so-called natural language, like English? This was the question that guided the creation of Eliza, the success of which made his name at the university and helped him secure tenure in 1967. It also brought Weizenbaum into the orbit of MIT’s Artificial Intelligence Project, which had been set up in 1958 by John McCarthy and Marvin Minsky. McCarthy had coined the phrase “artificial intelligence” a few years earlier when he needed a title for an academic workshop. The phrase was neutral enough to avoid overlap with existing areas of research like cybernetics, amorphous enough to attract cross-disciplinary contributions, and audacious enough to convey his radicalism (or, if you like, arrogance) about what machines were capable of. This radicalism was affirmed in the original workshop proposal. “Every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it,” it asserted. Minsky was bullish and provocative; one of his favourite gambits was to declare the human brain nothing but a “meat machine” whose functions could be reproduced, or even surpassed, by human-made machines. Weizenbaum disliked him from the start. It wasn’t his faith in the capabilities of technology that bothered Weizenbaum; he himself had seen computers progress immensely by the mid-1960s. Rather, Weizenbaum’s trouble with Minsky, and with the AI community as a whole, came down to a fundamental disagreement about the nature of the human condition. In Weizenbaum’s 1967 follow-up to his first article about Eliza, he argued that no computer could ever fully understand a human being. Then he went one step further: no human being could ever fully understand another human being. Everyone is formed by a unique collection of life experiences that we carry around with us, he argued, and this inheritance places limits on our ability to comprehend one another. We can use language to communicate, but the same words conjure different associations for different people – and some things can’t be communicated at all. “There is an ultimate privacy about each of us that absolutely precludes full communication of any of our ideas to the universe outside ourselves,” Weizenbaum wrote. This was a very different perspective than that of Minsky or McCarthy. It clearly bore the influence of psychoanalysis. Here was the mind not as a meat machine but as a psyche – something with depth and strangeness. If we are often opaque to one another and even to ourselves, what hope is there for a computer to know us? Yet, as Eliza illustrated, it was surprisingly easy to trick people into feeling that a computer did know them – and into seeing that computer as human. Even in his original 1966 article, Weizenbaum had worried about the consequences of this phenomenon, warning that it might lead people to regard computers as possessing powers of “judgment” that are “deserving of credibility”. “A certain danger lurks there,” he wrote. In the mid-1960s, this was as far as he was willing to go. He pointed to a danger, but didn’t dwell on it. He was, after all, a depressed kid who had escaped the Holocaust, who always felt like an impostor, but who had found prestige and self-worth in the high temple of technology. It can be hard to admit that something you are good at, something you enjoy, is bad for the world – and even harder to act on that knowledge. For Weizenbaum, it would take a war to know what to do next. * * * On 4 March 1969, MIT students staged a one-day “research stoppage” to protest the Vietnam war and their university’s role in it. People braved the snow and cold to pile into Kresge Auditorium in the heart of campus for a series of talks and panels that had begun the night before. Noam Chomsky spoke, as did the anti-war senator George McGovern. Student activism had been growing at MIT, but this was the largest demonstration to date, and it received extensive coverage in the national press. “The feeling in 1969 was that scientists were complicit in a great evil, and the thrust of 4 March was how to change it,” one of the lead organisers later wrote. Weizenbaum supported the action and became strongly affected by the political dynamism of the time. “It wasn’t until the merger of the civil rights movement, the war in Vietnam, and MIT’s role in weapons development that I became critical,” he later explained in an interview. “And once I started thinking along those lines, I couldn’t stop.” In the last years of his life, he would reflect on his politicisation during the 1960s as a return to the social consciousness of his leftist days in Detroit and his experiences in Nazi Germany: “I stayed true to who I was,” he told the German writer Gunna Wendt. He began to think about the German scientists who had lent their expertise to the Nazi regime. “I had to ask myself: do I want to play that kind of role?” he remembered in 1995. He had two choices. One was to “push all this sort of thinking down”, to repress it. The other was “to look at it seriously”. Looking at it seriously would require examining the close ties between his field and the war machine that was then dropping napalm on Vietnamese children. Defense Secretary Robert McNamara championed the computer as part of his crusade to bring a mathematical mindset to the Pentagon. Data, sourced from the field and analysed with software, helped military planners decide where to put troops and where to drop bombs. By 1969, MIT was receiving more money from the Pentagon than any other university in the country. Its labs pursued a number of projects designed for Vietnam, such as a system to stabilise helicopters in order to make it easier for a machine-gunner to obliterate targets in the jungle below. Project MAC – under whose auspices Weizenbaum had created Eliza – had been funded since its inception by the Pentagon. As Weizenbaum wrestled with this complicity, he found that his colleagues, for the most part, didn’t care about the purposes to which their research might be put. If we don’t do it, they told him, somebody else will. Or: scientists don’t make policy, leave that to the politicians. Weizenbaum was again reminded of the scientists in Nazi Germany who insisted that their work had nothing to do with politics. Consumed by a sense of responsibility, Weizenbaum dedicated himself to the anti-war movement. “He got so radicalised that he didn’t really do much computer research at that point,” his daughter Pm told me. Instead, he joined street demonstrations and met anti-war students. Where possible, he used his status at MIT to undermine the university’s opposition to student activism. After students

occupied the president's office in 1970, Weizenbaum served on the disciplinary committee. According to his daughter Miriam, he insisted on a strict adherence to due process, thereby dragging out the proceedings as long as possible so that students could graduate with their degrees. It was during this period that certain unresolved questions about Eliza began to bother him more acutely. Why had people reacted so enthusiastically and so delusionally to the chatbot, especially those experts who should know better? Some psychiatrists had hailed Eliza as the first step toward automated psychotherapy; some computer scientists had celebrated it as a solution to the problem of writing software that understood language. Weizenbaum became convinced that these responses were "symptomatic of deeper problems" – problems that were linked in some way to the war in Vietnam. And if he wasn't able to figure out what they were, he wouldn't be able to keep going professionally. * * * In 1976, Weizenbaum published his magnum opus: Computer Power and Human Reason: From Judgment to Calculation. "The book has overwhelmed me, like being crashed over by the sea," read a blurb from the libertarian activist Karl Hess. The book is indeed overwhelming. It is a chaotic barrage of often brilliant thoughts about computers. A glimpse at the index reveals the range of Weizenbaum's interlocutors: not only colleagues like Minsky and McCarthy but the political philosopher Hannah Arendt, the critical theorist Max Horkheimer, and the experimental playwright Eugène Ionesco. He had begun work on the book after completing a fellowship at Stanford University, in California, where he enjoyed no responsibilities, a big office and lots of stimulating discussions with literary critics, philosophers and psychiatrists. With Computer Power and Human Reason, he wasn't so much renouncing computer science as trying to break it open and let alternative traditions come pouring in. The book has two major arguments. First: "There is a difference between man and machine." Second: "There are certain tasks which computers ought not be made to do, independent of whether computers can be made to do them." The book's subtitle – From Judgment to Calculation – offers a clue as to how these two statements fit together. For Weizenbaum, judgment involves choices that are guided by values. These values are acquired through the course of our life experience and are necessarily qualitative: they cannot be captured in code. Calculation, by contrast, is quantitative. It uses a technical calculus to arrive at a decision. Computers are only capable of calculation, not judgment. This is because they are not human, which is to say, they do not have a human history – they were not born to mothers, they did not have a childhood, they do not inhabit human bodies or possess a human psyche with a human unconscious – and so do not have the basis from which to form values. And that would be fine, if we confined computers to tasks that only required calculation. But thanks in large part to a successful ideological campaign waged by what he called the "artificial intelligentsia", people increasingly saw humans and computers as interchangeable. As a result, computers had been given authority over matters in which they had no competence. (It would be a "monstrous obscenity", Weizenbaum wrote, to let a computer perform the functions of a judge in a legal setting or a psychiatrist in a clinical one.) Seeing humans and computers as interchangeable also meant that humans had begun to conceive of themselves as computers, and so to act like them. They mechanised their rational faculties by abandoning judgment for calculation, mirroring the machine in whose reflection they saw themselves. This had especially destructive policy consequences. Powerful figures in government and business could outsource decisions to computer systems as a way to perpetuate certain practices while absolving themselves of responsibility. Just as the bomber pilot "is not responsible for burned children because he never sees their village", Weizenbaum wrote, software afforded generals and executives a comparable degree of psychological distance from the suffering they caused. Letting computers make more decisions also shrank the range of possible decisions that could be made. Bound by an algorithmic logic, software lacked the flexibility and the freedom of human judgment. This helps explain the conservative impulse at the heart of computation. Historically, the computer arrived "just in time", Weizenbaum wrote. But in time for what? "In time to save – and save very nearly intact, indeed, to entrench and stabilise – social and political structures that otherwise might have been either radically renovated or allowed to totter under the demands that were sure to be made on them." Computers became mainstream in the 1960s, growing deep roots within American institutions just as those institutions faced grave challenges on multiple fronts. The civil rights movement, the anti-war movement and the New Left are just a few of the channels through which the era's anti-establishment energies found expression. Protesters frequently targeted information technology, not only because of its role in the Vietnam war but also due to its association with the imprisoning forces of capitalism. In 1970, activists at the University of Wisconsin destroyed a mainframe during a building occupation; the same year, protesters almost blew one up with napalm at New York University. This was the atmosphere in which Computer Power and Human Reason appeared. Computation had become intensely politicised. There was still an open question as to the path that it should take. On one side stood those who "believe there are limits to what computers ought to be put to do," Weizenbaum writes in the book's introduction. On the other were those who "believe computers can, should, and will do everything" – the artificial intelligentsia. * * * Marx once described his work Capital as "the most terrible missile that has yet been hurled at the heads of the bourgeoisie". Computer Power and Human Reason seemed to strike the artificial intelligentsia with similar force. McCarthy, the original AI guru, seethed: "Moralistic and incoherent", a work of "new left sloganeering", he wrote in a review. Benjamin Kuipers from MIT's AI Lab – a PhD student of Minsky's – complained of Weizenbaum's "harsh and sometimes shrill accusations against the artificial intelligence research community". Weizenbaum threw himself into the fray: he wrote a point-by-point reply to McCarthy's review, which led to a response from the Yale AI scientist Roger C Schank – to which Weizenbaum also replied. He clearly relished the combat. In the spring of 1977, the controversy spilled on to the front page of the New York Times. "Can machines think? Should they? The computer world is in the midst of a fundamental dispute over those questions," wrote the journalist Lee Dembart.

Weizenbaum gave an interview from his MIT office: “I have pronounced heresy and I am a heretic.” Computer Power and Human Reason caused such a stir because its author came from the world of computer science. But another factor was the besieged state of AI itself. By the mid-1970s, a combination of budget-tightening and mounting frustration within government circles about the field failing to live up to its hype had produced the first “AI winter”. Researchers now struggled to get funding. The elevated temperature of their response to Weizenbaum was likely due at least in part to the perception that he was kicking them when they were down. AI wasn’t the only area of computation being critically reappraised in these years. Congress had been recently contemplating ways to regulate “electronic data processing” by governments and businesses in order to protect people’s privacy and to mitigate the potential harms of computerised decision-making. (The watered-down Privacy Act was passed in 1974.) Between radicals attacking computer centers on campus and Capitol Hill looking closely at data regulation, the first “techlash” had arrived. It was good timing for Weizenbaum. Computer Power and Human Reason gave him a national reputation. He was delighted. “Recognition was so important to him,” his daughter Miriam told me. As the “house pessimist of the MIT lab” (the Boston Globe), he became a go-to source for journalists writing about AI and computers, one who could always be relied upon for a memorable quote. But the doubts and anxieties that had plagued him since childhood never left. “I remember him saying that he felt like a fraud,” Miriam told me. “He didn’t think he was as smart as people thought he was. He never felt like he was good enough.” As the excitement around the book died down, these feelings grew overwhelming. His daughter Pm told me that Weizenbaum attempted suicide in the early 1980s. He was hospitalised at one point; a psychiatrist diagnosed him with narcissistic personality disorder. The sharp swings between grandiosity and dejection took their toll on his loved ones. “He was a very damaged person and there was only so much he could absorb of love and family,” Pm said. In 1988, he retired from MIT. “I think he ended up feeling pretty alienated,” Miriam told me. In the early 1990s, his second wife, Ruth, left him; in 1996, he returned to Berlin, the city he had fled 60 years earlier. “Once he moved back to Germany, he seemed much more content and engaged with life,” Pm said. He found life easier there. As his fame faded in the US, it increased in Germany. He became a popular speaker, filling lecture halls and giving interviews in German. The later Weizenbaum was increasingly pessimistic about the future, much more so than he had been in the 1970s. Climate change terrified him. Still, he held out hope for the possibility of radical change. As he put it in a January 2008 article for Süddeutsche Zeitung: “The belief that science and technology will save the Earth from the effects of climate breakdown is misleading. Nothing will save our children and grandchildren from an Earthly hell. Unless: we organise resistance against the greed of global capitalism.” Two months later, on 5 March 2008, Weizenbaum died of stomach cancer. He was 85. * * * By the time Weizenbaum died, AI had a bad reputation. The term had become synonymous with failure. The ambitions of McCarthy, formulated at the height of the American century, were gradually extinguished in the subsequent decades. Getting computers to perform tasks associated with intelligence, like converting speech to text, or translating from one language to another, turned out to be much harder than anticipated. Today, the situation looks rather different. We have software that can do speech recognition and language translation quite well. We also have software that can identify faces and describe the objects that appear in a photograph. This is the basis of the new AI boom that has taken place since Weizenbaum’s death. Its most recent iteration is centred on “generative AI” applications like ChatGPT, which can synthesise text, audio and images with increasing sophistication. At a technical level, the set of techniques that we call AI are not the same ones that Weizenbaum had in mind when he commenced his critique of the field a half-century ago. Contemporary AI relies on “neural networks”, which is a data-processing architecture that is loosely inspired by the human brain. Neural networks had largely fallen out of fashion in AI circles by the time Computer Power and Human Reason came out, and would not undergo a serious revival until several years after Weizenbaum’s death. But Weizenbaum was always less concerned by AI as a technology than by AI as an ideology – that is, in the belief that a computer can and should be made to do everything that a human being can do. This ideology is alive and well. It may even be stronger than it was in Weizenbaum’s day. Certain of Weizenbaum’s nightmares have come true: so-called risk assessment instruments are being used by judges across the US to make crucial decisions about bail, sentencing, parole and probation, while AI-powered chatbots are routinely touted as an automated alternative to seeing a human therapist. The consequences may have been about as grotesque as he expected. According to reports earlier this year, a Belgian father of two killed himself after spending weeks talking with an AI avatar named … Eliza. The chat logs that his widow shared with the Brussels-based newspaper La Libre show Eliza actively encouraging the man to kill himself. On the other hand, Weizenbaum would probably be heartened to learn that AI’s potential for destructiveness is now a matter of immense concern. It preoccupies not only policymakers – the EU is finalising the world’s first comprehensive AI regulation, while the Biden administration has rolled out a number of initiatives around “responsible” AI – but AI practitioners themselves. Broadly, there are two schools of thought today about the dangers of AI. The first – influenced by Weizenbaum – focuses on the risks that exist now. For instance, experts such as the linguist Emily M Bender draw attention to how large language models of the kind that sit beneath ChatGPT can echo regressive viewpoints, like racism and sexism, because they are trained on data drawn from the internet. Such models should be understood as a kind of “parrot”, she and her co-authors write in an influential 2021 paper, “haphazardly stitching together sequences of linguistic forms it has observed in its vast training data, according to probabilistic information about how they combine.” The second school of thought prefers to think in speculative terms. Its adherents are less interested in the harms that are already here than in the ones that may someday arise – in particular the “existential risk” of an AI that becomes “superintelligent” and wipes out the human race. Here the reigning metaphor is not a parrot but

Skynet, the genocidal computer system from the Terminator films. This perspective enjoys the ardent support of several tech billionaires, including Elon Musk, who have financed a network of like-minded thinktanks, grants and scholarships. It has also attracted criticism from members of the first school, who observe that such doomsaying is useful for the industry because it diverts attention away from the real, current problems that its products are responsible for. If you "project everything into the far future," notes Meredith Whittaker, you leave "the status quo untouched". Weizenbaum, ever attentive to the ways in which fantasies about computers can serve powerful interests, would probably agree. But there is nonetheless a thread of existential risk thinking that has some overlap with his own: the idea of AI as alien. "A superintelligent machine would be as alien to humans as human thought processes are to cockroaches," argues the philosopher Nick Bostrom, while the writer Eliezer Yudkowsky likens advanced AI to "an entire alien civilisation". Weizenbaum would add the following caveat: AI is already alien, even without being "superintelligent". Humans and computers belong to separate and incommensurable realms. There is no way of narrowing the distance between them, as the existential risk crowd hopes to do through "AI alignment", a set of practices for "aligning" AI with human goals and values to prevent it from becoming Skynet. For Weizenbaum, we cannot humanise AI because AI is irreducibly non-human. What you can do, however, is not make computers do (or mean) too much. We should never "substitute a computer system for a human function that involves interpersonal respect, understanding and love", he wrote in Computer Power and Human Reason. Living well with computers would mean putting them in their proper place: as aides to calculation, never judgment. Weizenbaum never ruled out the possibility that intelligence could someday develop in a computer. But if it did, he told the writer Daniel Crevier in 1991, it would "be at least as different as the intelligence of a dolphin is to that of a human being". There is a possible future hiding here that is neither an echo chamber filled with racist parrots nor the Hollywood dystopia of Skynet. It is a future in which we form a relationship with AI as we would with another species: awkwardly, across great distances, but with the potential for some rewarding moments. Dolphins would make bad judges and terrible shrinks. But they might make for interesting friends. • Follow the Long Read on Twitter at @gdnlongread, listen to our podcasts here and sign up to the long read weekly email here.