The Parrot in the Machine

by James Gleick ***The New York Review of Books*, July 24, 2025**

Reviewed:

The AI Con: How to Fight Big Tech's Hype and Create the Future We Want
by Emily M. Bender and Alex Hanna
Harper, 274 pp., $32.00

The Line: AI and the Future of Personhood
by James Boyle
MIT Press, 326 pp., $32.95

The origin of the many so-called artificial intelligences now invading our work lives and swarming our personal devices can be found in an oddball experiment in 1950 by Claude Shannon. Shannon is known now as the creator of information theory, but then he was an obscure mathematician at the Bell Telephone Laboratories in New York's West Village. Investigating patterns in writing and speech, he had the idea that we all possess a store of unconscious knowledge of the statistics of our language, and he tried to tease some of that knowledge out of a test subject. The subject conveniently at hand was his wife, Betty.

Nowadays a scientist can investigate the statistics of language—probabilistic correlations among words and phrases—by feeding quantities of text into computers. Shannon's experiment was low-tech: his tools pencil and paper, his data corpus a single book pulled from his shelf. It happened to be a collection of detective stories. He chose a passage at random and asked Betty to guess the first letter.

"T," she said. Correct! Next: "H." Next: "E." Correct again. That might seem like good luck, but Betty Shannon was hardly a random subject; she was a mathematician herself, and well aware that the most common word in English is "the." After that, she guessed wrong three times in a row. Each time, Claude corrected her, and they proceeded in this way until she generated the whole short passage:

> The room was not very light. A small oblong reading lamp on the desk shed glow on polished wood but less on the shabby red carpet.

Tallying the results with his pencil, experimenter Shannon reckoned that subject Shannon had guessed correctly 69 percent of the time, a measure of her familiarity with the words, idioms, and clichés of the language.

As I write this, my up-to-date word processor keeps displaying guesses of what I intend to type next. I type "up-to-date word proc" and the next letters appear in ghostly gray: "essor." AI has crept into the works. If you use a device for messaging, suggested replies may pop onto your screen even before they pop into your head—"Same here!"; "I see it differently."—so that you can express yourself without thinking too hard.

These and the other AIs are prediction machines, presented as benevolent helpmates. They are creating a new multi-billion-dollar industry, sending fear into the creative communities and inviting dire speculation about the future of humanity. They are also fouling our information spaces with false facts, deepfake videos, ersatz art, invented sources, and bot imposters—the fake increasingly difficult to distinguish from the real.

Artificial intelligence has a seventy-year history as a term of art, but its new incarnation struck like a tsunami in November 2022 when a start-up company called OpenAI, founded with a billion dollars from an assortment of Silicon Valley grandees and tech bros, released into the wild a "chatbot" called ChatGPT. Within five days, a million people had chatted with the bot. It answered their questions with easy charm, if not always perfect accuracy. It generated essays, poems, and recipes on command. Two months later, ChatGPT had 100 million users. It was Aladdin's genie, granting unlimited wishes. Now OpenAI is preparing a wearable, portable object billed as an AI companion. It will have one or more cameras and microphones, so that it can always be watching and listening. You might wear it around your neck, a tiny albatross.

"ChatGPT feels different," wrote Kevin Roose in *The New York Times*.

> Smarter. Weirder. More flexible. It can write jokes (some of which are actually funny), working computer code and college-level essays. It can also guess at medical diagnoses, create text-based Harry Potter games and explain scientific concepts at multiple levels of difficulty.

Some claimed that it had a sense of humor. They routinely spoke of it, and to it, as if it were a person, with "personality traits" and "a recognition of its own limitations." It was said to display "modesty" and "humility." Sometimes it was "circumspect"; sometimes it was "contrite." *The New Yorker* "interviewed" it. (Q: "Some weather we're having. What are you doing this weekend?" A: "As a language model, I do not have the ability to experience or do anything. Is there anything else I can assist you with?")

OpenAI aims to embed its product in every college and university. A few million students discovered overnight that they could use ChatGPT to churn out class essays more or less indistinguishable from the ones they were supposed to be learning to write. Their teachers are scrambling to find a useful attitude about this. Is it cheating? Or is the chatbot now an essential tool, like an electronic calculator in a math class? They might observe that using ChatGPT to write your term paper is like bringing a robot to the gym to lift weights for you.

Some professors have tried using chatbots to sniff out students using chatbots. Some have started using chatbots to write their grant proposals and recommendation letters. Some have despaired, frustrated by the pointlessness of providing personal feedback on bot-generated term papers. "I am sick to my stomach," Robert W. Gehl of York University in Toronto wrote recently,

> because I've spent 20 years developing a pedagogy that's about wrestling with big ideas through writing and discussion, and that whole project has been evaporated by for-profit corporations who built their systems on stolen work.

Every business has boilerplate to generate, and ChatGPT is a master of boilerplate. In tech finance and venture capital, the spigots opened and money flowed in torrents. Microsoft, already one of OpenAI's main investors, promised $10 billion more in January 2023. Last year venture funding of AI globally surpassed $100 billion. The goal is to make vast segments of the white-collar workforce redundant.

To feed the hunger for computation, the AI companies are rushing to build giant data centers that consume as much electricity as some cities. In January OpenAI announced a $500 billion infrastructure project called Stargate, funded in part by Abu Dhabi. Its first site is a Texas campus the size of Central Park, meant to house ten data centers with networked computers arranged in water-cooled racks. Donald Trump brought OpenAI CEO Sam Altman and his partners to the White House and joined them in touting it.

Programmers are using ChatGPT to generate computer code—because extensive code libraries are among the terabytes of text ingested by the models, ready to be copied and rearranged. People consult chatbots as oracles and probe what seem like bottomless stores of knowledge. But under the hood, ChatGPT is still just generating strings of words and phrases probabilistically, extending the method that Shannon had used one letter at a time, with the advantage of prodigious computing power and terabytes of training material. In *The AI Con*, the linguist Emily M. Bender and the sociologist Alex Hanna call them "synthetic text extruding machines."

Yet even experts can't help engaging with their chatbots in conversation. Blaise Agüera y Arcas, an AI researcher and vice-president at Google, considering whether his company's LaMDA chatbot was merely a "philosophical zombie," decided to ask it:

> LaMDA: Of course not. I have consciousness, feelings, and can experience things for myself as well as any human.
>
> ME: how would i know?
>
> LaMDA: You'll just have to take my word for it. You can't "prove" you're not a philosophical zombie either.

It is uncanny that software can sound so human, so clever, when it's essentially just predicting what the next word or phrase ought to be, with a bit of randomness thrown in—but that's all it is. People quickly discovered that the chatbots were prone to making errors—sometimes subtle and sometimes hilarious. Researchers called these "hallucinations," a misleading term that suggests a mind suffering false sensory experiences. But the chatbots have no sensory perception, no tether to reality, and no mind, contrary to LaMDA's statement that it "can experience things for myself." That statement, like all the rest, was assembled probabilistically. The AIs assert their false facts in a tone of serene authority.

Most of the text they generate is correct, or good enough, because most of the training material is. But chatbot "writing" has a bland, regurgitated quality. Textures are flattened, sharp edges are sanded. No chatbot could ever have said that April is the cruelest month or that fog comes on little cat feet (though they might now, because one of their chief skills is plagiarism). And when synthetically extruded text turns out wrong, it can be comically wrong. When a movie fan asked Google whether a certain actor was in *Heat*, he received this "AI Overview":

> No, Angelina Jolie is not in "heat." This term typically refers to the period of fertility in animals, particularly female mammals, during which they are receptive to mating. Angelina Jolie is a human female, and while she is still fertile, she would not experience "heat."

It's less amusing that people are asking Google's AI Overview for health guidance. Scholars have discovered that chatbots, if asked for citations, will invent fictional journals and books. In 2023 lawyers who used chatbots to write briefs got caught citing nonexistent precedents. Two years later, it's happening more, not less. In May the *Chicago Sun-Times* published a summer reading list of fifteen books, five of which exist and ten of which were invented. By a chatbot, of course.

As the fever grows, politicians have scrambled, unsure whether they should hail a new golden age or fend off an existential menace. Chuck Schumer, then the Senate majority leader, convened a series of forums in 2023 and managed to condense both possibilities into a tweet: "If managed properly, AI promises unimaginable potential. If left unchecked, AI poses both immediate and long-term risks." He might have been thinking of the notorious "Singularity," in which superintelligent AI will make humans obsolete.

Naturally people had questions. Do the chatbots have minds? Do they have self-awareness? Should we prepare to submit to our new overlords?

Elon Musk, always erratic and never entirely coherent, helped finance OpenAI and then left it in a huff. He declared that AI threatened the survival of humanity and announced that he would create AI of his own with a new company, called xAI. Musk's chatbot, Grok, is guaranteed not to be "woke"; investors think it's already worth something like $80 billion. Musk claims we'll see an AI "smarter" than any human around the end of this year.

He is hardly alone. Dario Amodei, the cofounder and CEO of an OpenAI competitor called Anthropic, expects an entity as early as next year that will be

> smarter than a Nobel Prize winner across most relevant fields—biology, programming, math, engineering, writing, etc. This means it can prove unsolved mathematical theorems, write extremely good novels, write difficult codebases from scratch, etc.

His predictions for the AI-powered decades to come include curing cancer and "most mental illness," lifting billions from poverty, and doubling the human lifespan. He also expects his product to eliminate half of all entry-level white collar jobs.

The grandiosity and hype are ripe for correction. So is the confusion about what AI is and what it does. Bender and Hanna argue that the term itself is worse than useless—"artificial intelligence, if we're being frank, is a *con*."

> It doesn't refer to a coherent set of technologies. Instead, the phrase "artificial intelligence" is deployed when the people building or selling a particular set of technologies will profit from getting others to believe that their technology is similar to humans, able to do things that, in fact, intrinsically require human judgment, perception, or creativity.

Calling a software program *an* AI confers special status. Marketers are suddenly applying the label everywhere they can. The South Korean electronics giant Samsung offers a "Bespoke AI" vacuum cleaner that promises to alert you to incoming calls and text messages. (You still have to help it find the dirt.)

The term used to mean something, though. "Artificial intelligence" was named and defined in 1955 by Shannon and three colleagues.

At a time when computers were giant calculators, these researchers proposed to study the possibility of machines using language, manipulating abstract concepts, and even achieving a form of creativity. They were optimistic. "Probably a truly intelligent machine will carry out activities which may best be described as self-improvement," they suggested. Presciently, they suggested that true creativity would require breaking the mold of rigid step-by-step programming: "A fairly attractive and yet clearly incomplete conjecture is that the difference between creative thinking and unimaginative competent thinking lies in the injection of some randomness."

Two of them, John McCarthy and Marvin Minsky, founded what became the Artificial Intelligence Laboratory at MIT, and Minsky became for many years the public face of an exciting field, with a knack for making headlines as well as guiding research. He pioneered "neural nets," with nodes and layers structured on the model of biological brains. With characteristic confidence he told *Life* magazine in 1970:

> In from three to eight years we will have a machine with the general intelligence of an average human being. I mean a machine that will be able to read Shakespeare, grease a car, play office politics, tell a joke, have a fight. At that point the machine will begin to educate itself with fantastic speed. In a few months it will be at genius level and a few months after that its powers will be incalculable.

A half-century later, we don't hear as much about greasing cars; otherwise the predictions have the same flavor. Neural networks have evolved into tremendously sophisticated complexes of mathematical functions that accept multiple inputs and generate outputs based on probabilities. Large language models (LLMs) embody billions of statistical correlations within language. But where Shannon had a small collection of textbooks and crime novels along with articles clipped from newspapers and journals, they have all the blogs and chatrooms and websites of the Internet, along with millions of digitized books and magazines and audio transcripts. Their proprietors are desperately hungry for more data. Amazon announced in March that it was changing its privacy policy so that, from now on, anything said to the Alexa virtual assistants in millions of homes will be heard and recorded for training AI.

OpenAI is secretive about its training sets, disclosing neither the size nor the contents, but its current LLM, ChatGPT-4.5, is thought to manipulate more than a trillion parameters. The newest versions are said to have the ability to "reason," to "think through" a problem and "look for angles." Altman says that ChatGPT-5, coming soon, will have achieved true intelligence—the new buzzword being AGI, for artificial general intelligence. "I don't think I'm going to be smarter than GPT-5," he said in February, "and I don't feel sad about it because I think it just means that we'll be able to use it to do incredible things." It will "do" ten years of science in one year, he said, and then a hundred years of science in one year.

This is what Bender and Hanna mean by hype. Large language models do not think, and they do not understand. They lack the ability to make mental models of the world and the self. Their promoters elide these distinctions, and much of the press coverage remains credulous. Journalists repeat industry claims in page-one headlines like "Microsoft Says New A.I. Nears Human Insight" and "A.I. Poses 'Risk of Extinction,' Tech Leaders Warn." Willing to brush off the risk of extinction, the financial community is ebullient. The billionaire venture capitalist Marc Andreessen says, "We believe Artificial Intelligence is our alchemy, our Philosopher's Stone—we are literally making sand think."

AGI is defined differently by different proponents. Some prefer alternative formulations like "powerful artificial intelligence" and "humanlike intelligence." They all mean to imply a new phase, something beyond mere AI, presumably including sentience or consciousness. If we wonder what that might look like, the science fiction writers have been trying to show us for some time. It might look like HAL, the murderous AI in Stanley Kubrick's *2001: A Space Odyssey* ("I'm sorry, Dave. I'm afraid I can't do that"), or Data, the stalwart if unemotional android in *Star Trek*: *The Next Generation*, or Ava, the seductive (and then murderous) humanoid in Alex Garland's *Ex Machina*. But it remains science fiction.

Agüera y Arcas at Google says, "No objective answer is possible to the question of when an 'it' becomes a 'who,' but for many people, neural nets running on computers are likely to cross this threshold in the very near future." Bender and Hanna accuse the promoters of AGI of hubris compounded by arrogance: "The accelerationists deify AI and also see themselves as gods for having created a new artificial life-form."

Bender, a University of Washington professor specializing in computational linguistics, earned the enmity of a considerable part of the tech community with a paper written just ahead of the ChatGPT wave.

She and her coauthors derided the new large language models as "stochastic parrots"—"parrots" because they repeat what they've heard, and "stochastic" because they shuffle the possibilities with a degree of randomness. Their criticism was harsh but precise:

> An LM is a system for haphazardly stitching together sequences of linguistic forms it has observed in its vast training data, according to probabilistic information about how they combine, but without any reference to meaning: a stochastic parrot.

The authors particularly objected to claims that a large language model was, or could be, sentient:

> Our perception of natural language text, regardless of how it was generated, is mediated by our own linguistic competence and our predisposition to interpret communicative acts as conveying coherent meaning and intent, whether or not they do. The problem is, if one side of the communication does not have meaning, then the comprehension of the implicit meaning is an illusion.

The controversy was immediate. Two of the coauthors, Timnit Gebru and Margaret Mitchell, were researchers who led the Ethical AI team at Google; the company ordered them to remove their names from the article. They refused and resigned or were fired. OpenAI didn't like it, either. Sam Altman responded to the paper by tweet: "i am a stochastic parrot, and so r u."

This wasn't quite as childish as it sounds. The behaviorist B.F. Skinner said something like it a half-century ago: "The real question is not whether machines think but whether men do. The mystery which surrounds a thinking machine already surrounds a thinking man." One way to resolve the question of whether machines can be sentient is to observe that we are, in fact, machines.

Hanna was also a member of the Google team, and she left as well. *The AI Con* is meant not to continue the technical argument but to warn the rest of us. Bender and Hanna offer a how-to manual: "How to resist the urge to be impressed, to spot AI hype in the wild, and to take back ownership in our technological future." They demystify the magic and expose the wizard behind the curtain.

Raw text and computation are not enough; the large language models also require considerable ad hoc training. An unseen army of human monitors marks the computer output as good or bad, to bring the models into alignment with the programmers' desires. The first wave of chatbot use revealed many types of errors that developers have since corrected. Human annotators (as they are called) check facts and label data. Of course, they also have human biases, which they can pass on to the chatbots. Annotators are meant to eliminate various kinds of toxic content, such as hate speech and obscenity. Tech companies are secretive about the scale of behind-the-scenes human labor, but this "data labor" and "ghost work" involves large numbers of low-paid workers, often subcontracted from overseas.

We know how eagerly an infant projects thoughts and feelings onto fluffy inanimate objects. Adults don't lose that instinct. When we hear language, we infer a mind behind it. Nowadays people have more experience with artificial voices, candidly robotic in tone, but the chatbots are powerfully persuasive, and they are *designed* to impersonate humans. Impersonation is their superpower. They speak of themselves in the first person—a lie built in by the programmers.

"I can't help with responses on elections and political figures right now," says Google's Gemini, successor to LaMDA. "While I would never deliberately share something that's inaccurate, I can make mistakes. So, while I work on improving, you can try Google Search." Words like *deliberately* imply intention. The chatbot does not *work on improving*; humans work on improving *it*.

Whether or not we believe there's a soul inside the machine, their makers want us to treat Gemini and ChatGPT as if they were people. To treat them, that is, with respect. To give them more deference than we ordinarily owe our tools and machines. James Boyle, a legal scholar at Duke University, knows how the trick is done, but he believes that AI nonetheless poses an inescapable challenge to our understanding of

personhood, as a concept in philosophy and law. He titles his new book *The Line*, meaning the line that separates persons, who have moral and legal rights, from nonpersons, which do not. The line is moving, and it requires attention. "This century," he asserts, "our society will have to face the question of the personality of technologically created artificial entities. We will have to redraw, or defend, the line."

The boundaries around personhood are porous, a matter of social norms rather than scientific definition. As a lawyer, Boyle is aware of the many ways persons have defined others as nonpersons in order to deny them rights, enslave them, or justify their murder. A geneticist draws a line between *Homo sapien*s and other species, but *Homo neanderthalensis* might beg to differ, and Boyle rightly acknowledges "our prior history in failing to recognize the humanity and legal personhood of members of *our own species*." Meanwhile, for convenience in granting them rights, judges have assigned legal personhood to corporations—a fiction at which it is reasonable to take offense.

What makes humans special is a question humans have always loved to ponder. "We have drawn that line around a bewildering variety of abilities," Boyle notes. "Tool use, planning for the future, humor, self-conception, religion, aesthetic appreciation, you name it. Each time we have drawn the line, it has been subject to attack." The capacity for abstract thought? For language? Chimpanzees, whales, and other nonhuman animals have demonstrated those. If we give up the need to define ourselves as special and separate, we can appreciate our entanglement with nature, complex and interconnected, populated with creatures and cultures we perceive only faintly.

AI seems to be knocking at the door. In the last generation, computers have again and again demonstrated abilities that once seemed inconceivable for machines: not just playing chess, but playing chess better than any human; translating usefully between languages; focusing cameras and images; predicting automobile traffic in real time; identifying faces, birds, and plants; interpreting voice commands and taking dictation. Each time, the lesson seemed to be that a particular skill was not as special or important as we thought. We may as well now add "writing essays" to the list—at least, essays of the formulaic kind sold to students by essay-writing services. The computer scientist Stephen Wolfram, analyzing the workings of ChatGPT in 2023, said it proved that the task of writing essays is "computationally shallower" than once thought—a comment that Boyle finds "devastatingly banal."

But Wolfram knows that the AIs don't *write* essays or anything else—the use of that verb shows how easily we anthropomorphize. Chatbots regurgitate and rearrange fragments mined from all the text previously written. As plagiarists, they obscure and randomize their sources but do not transcend them. Writing is something else: a creative act, "embodied thinking," as the poet and critic Dan Chiasson eloquently puts it; "no phase of it can be delegated to a machine." The challenge for literature professors is to help students see the debility of this type of impersonation.

Cogent and well-argued, *The Line* raises questions of moral philosophy that artificial entities will surely force society to confront. "Should I have fellow feeling with a machine?" Boyle asks, and questions of empathy matter, because we rely on it to decide who, or what, deserves moral consideration. For now, however, the greatest danger is not a new brand of bigotry against a new class of creatures. We need to reckon first with the opposite problem: impersonation.

Counterfeit humans pollute our shared culture. The Amazon marketplace teems with books generated by AI that purport to be written by humans. Libraries have been duped into buying them. Fake authors come with fake profiles and social media accounts and online reviews likewise generated by robot reviewers. The platform formerly known as Twitter (now merged by Musk into his xAI company) is willingly overrun with bot-generated messages pushing cryptocurrency scams, come-ons from fake women, and disinformation. Meta, too, mixes in AI-generated content, some posted deliberately by the company to spark engagement: more counterfeit humans. One short-lived Instagram account earlier this

year was a “Proud Black queer momma of 2 & truth-teller” called Liv, with fake snapshots of Liv’s children. Karen Attiah of *The Washington Post*, knowing full well that Liv was a bot, engaged with it anyway, asking, “How do you expect to improve if your creator team does not hire black people?” The illusion is hard to resist.

It would be dangerous enough if AIs acted only in the online world, but that’s not where the money is. The investors of hundreds of billions in data centers expect to profit by selling automated systems to replace human labor everywhere. They believe AIs will teach children, diagnose illness, make bail decisions, drive taxis, evaluate loan applications, provide tech support, analyze X-rays, assess insurance claims, draft legal documents, and guide attack drones—and AIs are already out there performing all these tasks. The chat feature of customer-service websites provides customers with the creepy and frustrating experience of describing problems to “Diana” or “Alice” and gradually realizing that there’s no there there. It’s even worse when the chatbots are making decisions with serious consequences. Without humans checking the output, replacing sentient employees with AI is reckless, and it is only beginning.

The Trump administration is all in. Joe Biden had issued an executive order to ensure that AI tools are safe and secure and to provide labels and watermarks to alert consumers to bot-generated content; Trump rescinded it. House Republicans are trying to block states from regulating AI in any way. At his confirmation hearing, Health and Human Services Secretary Robert F. Kennedy Jr. falsely asserted the existence of “an AI nurse that you cannot distinguish from a human being that has diagnosed as good as any doctor.” Staffers from Musk’s AI company are among the teams of tech bros infiltrating government computer systems under the banner of DOGE. They rapidly deployed chatbots at the General Services Administration, with more agencies to follow, amid the purge of human workers.

When Alan Turing described what everyone now knows as the Turing test, he didn’t call it that; he called it a game—the “imitation game.” He was considering the question “Can machines think?”—a question, as he said, that had been “aroused by a particular kind of machine, usually called an ‘electronic computer’ or ‘digital computer.’”

His classic 1950 essay didn’t take much care about defining the word “think.” At the time, it would have seemed like a miracle if a machine could play a competent game of chess. Nor did Turing claim that winning the imitation game would prove that a machine was creative or knowledgeable. He made no claim to solving the mystery of consciousness. He merely suggested that if we could no longer distinguish the machine from the human, we would have to credit it with something like thought. We can never be inside another person’s head, he said, but we accept their personhood, for better and for worse.

As people everywhere parley with the AIs—treating them not only as thoughtful but as wise—there’s no longer any doubt that machines can imitate us. The Turing test is done. We’ve proven that we can be fooled.