

1

What Is a Chatbot?

"Binary arithmetic is the most fundamental science
and can create a universal language of human thought."
—GOTTFRIED LEIBNIZ (1703)

Writers are told to lead with their best punch in the first chapter and end its last with an uplifting message. Engage readers at the start of your book and they are more likely to stick around for its end. Leave them feeling good at the end and they are more likely to recommend it to colleagues, friends, or a book club. Though I wanted to follow this excellent advice, I could not. We must, of necessity, begin our book with its most boring content—the nuts and bolts of chatbots. And I hope you end it fired up and prepared, not reassured and complacent. The good news is that the chapters in between describe one of the most interesting intellectual adventures in my life—and I hope in yours.

Before we begin, allow me to disclose my biases: 1) chatbots fill me with awe, but also with dread; 2) understanding how chatbots think is the royal road to understanding how we humans think; 3) soon most people in the world will be using chatbots for mental health advice; 4) chatbots are beneficial for some, harmful for others; 5) it's easy to train safe and effective chatbot assistants; 6) AI companies have far too much power, far too little wisdom; 7) clinicians who don't adapt to the fast-changing world of chatbots will have trouble finding a place in it; and 8) my dual and related missions are to keep our patients safe and to preserve the human touch in psychiatric care.

Much more on all this later. Now let's get to work.

Basic Terms

- **“Chatbot”** is the silliest possible name for one of mankind’s greatest ever achievements—creating an intelligence that simulates, rivals, and may soon greatly exceed our own. The term (diminutive for chatterbot) was coined thirty years ago to describe the AI software designed to copy human speech. Chatter is defined as “talking rapidly or incessantly about trivial matters” and bot is short for robot. Every once in a while, someone suggests a more dignified name. But “chatbot” is here to stay because a deceptively unassuming label helps expand the potential market of users to encompass even the techno-idiot (like me). If I can become chatbot literate, anyone can.

- **“Artificial Intelligence” vs. “Chatbot”:** AI is a broad concept encompassing all the ways smart machines learn, remember, reason, make decisions, and take action. Chatbots are a specific form of AI software that has the special gift of humanlike speech. In the past, chatbots were great talkers, but not great thinkers. That has changed dramatically in just the past few years. Chatbots can now do complex reasoning; solve logic and mathematical problems; see hidden patterns in large data sets; create deepfake images; plan projects; analyze and create charts; interpret the scientific literature; identify faces in a crowd; guide drones; predict the weather; plan a vacation; find the cheapest price—and do psychotherapy. Since chatbots now do almost everything that falls under the definition of AI, I use the terms almost interchangeably.

- **“Chatbot Binary Language” vs. “Human Alphabetical Language”:** Computer language uses two numbers (0s and 1s) because these conveniently control the “on” and “off” states of transistor circuitry. A binary on/off method of electronic communication is much simpler, faster, and more precise than having to adjust voltage levels to accommodate all 26 letters of the English alphabet. Until recently, computer programmers did all the translating from our phonetic speech to computer binary speech. Now chatbots can do much of the translation themselves.

- **“Vibe Coding” vs. “Human Programming”:** Techies (and physicists) love to apply silly names for earth-shaking breakthroughs. “Vibe coding” is a silly new term describing an incredible programming revolution—one that is great for chatbot users (like us) and awful for programmers. Chatbots can now directly translate your verbal instructions into binary code—quickly, cheaply, and without human intervention. Their range is remarkably wide—chatbots know dozens of human languages and hundreds of computer languages. Their

speed of translation into any or all of these languages is almost instantaneous. The bad news is that millions of software workers and human language translators may lose their jobs. The good news is that technically dumb chatbot users have suddenly acquired programming power. Chapter 10 will show how simple it now is to customize chatbot assistants, making them much safer and more specific in meeting each patient's needs.

- **“Hardware” vs. “Software”:** Hardware is the tangible stuff on a computer that you can touch; software is the abstract set of instructions that tell a computer what to do. Computer hardware and computer software are separate components that must work together. With vibe coding, you have direct verbal input into your chatbot's software. In human brains, the hardware and software are inextricably linked, advantaging us, and also disadvantaging us, in ways we will discuss later.

- **“Artificial Neural Networks” vs. “Human Brain Networks”:** Chatbots use a copycat system of transistor connections that closely mimic the neuronal structure and function of our brains. In Chapter 4, we will explore similarities and differences in human vs. machine thinking.

- **“Sequential” vs. “Parallel Processing”:** Until 2017, chatbots were relatively slow and clunky because they performed tasks serially, one at a time. Now chatbots are incredibly fast and fluent because they can break extraordinarily complex tasks into multiple small ones that are performed simultaneously and almost instantaneously.

- **“Graphics Processing Unit (GPU)” vs. “Central Processing Unit (CPU)” Chips:** GPU chips do super-fast parallel processing; CPU chips do the slower sequential processing. Ironically, the most powerful chips in the world were originally developed for the most trivial function—to power the complex graphics of computer games. NVIDIA's near monopoly of GPU game chip technology transformed it from start-up status thirty years ago to the world's most valuable company, as of this writing worth about five trillion dollars.

- **“Artificial Narrow Intelligence (ANI)” vs. “Artificial General Intelligence (AGI)” vs. “Superintelligence”:** ANI is what chatbots now have—they are brilliant in many cognitive tasks, but clueless in others. AGI is what chatbots will acquire within the next few years—the ability to do everything cognitive we can do at the highest level any of us can do it. Superintelligence is the theoretical possibility, probably a few decades away, that chatbots will surpass us in every type of intellectual activity such that we will no longer understand or control them.

- **“Generative AI” vs. “Agentic AI”:** Chatbot ability to work independently varies up the scale from rule-based dependence to supervised collaboration to fully autonomous functioning that requires no human intervention. Generative AI produces original, and sometimes surprisingly creative, content by seeing new patterns in large data sets—but it still requires human prompts and supervision. Agentic AI is completely autonomous—it can plan, program, and implement projects all on its own in real time. There is now a fierce technological and business competition to develop chatbots that can do it all. This is likely to be a commercial bonanza that greatly benefits shareholders, potentially greatly harms workers, and raises grave (mostly unasked and completely unanswered) ethical and existential questions about the future of our species.

- **“Machine Learning (ML)” vs. “Deep Learning (DL)”:** ML algorithms allow machines to search for patterns and learn new things within large, but limited, data sets without having to be programmed for each specific task. DL is a much more powerful subset of ML that uses multilayered neural network hardwiring to scrape and search virtually the entire internet.

- **“Supervised” vs. “Unsupervised” vs. “Reinforcement Learning”:** In supervised learning, humans curate computer training (e.g., teaching the best chess moves made by all previous human experts). In unsupervised learning, the chatbot learns on its own (e.g., by playing hundreds of thousands of chess games against itself). In reinforcement learning, the model learns by a combination of trial-and-error conditioning and rewards for good behavior.

- **“Large Language Model (LLM)” vs. “Small Language Models (SLM)”:** General LLM chatbots (e.g., Claude, ChatGPT, Gemini) are trained on large amounts of text and require vast computational power. Natural language processing enables them to understand and convincingly simulate human speech. LLM chatbots comprehend complex human statements, questions, and motivations and frame their replies appropriately, sometimes eloquently. They are warm, helpful, friendly, supportive, personal, even witty—and can detect nuances in human emotion and respond with what feels like genuine empathy (although of course it isn’t). SLM, specialized chatbots (e.g., those designed by mental health professionals for mental health apps) require much less data and computational juice because they are focused on a very specific task and don’t have to search nearly as much of the internet. They are safer, but also less fluent, worldly-wise, and popular with patients.

- **“Turing Test”:** Chatbot speech is now colloquial, intuitive, natural, responsive, and context aware—in every way contrasting with stiff, boring,

stereotyped ELIZA of the 1960s. Alan Turing suggested this test of computer intelligence in 1950: Is it clever and fluent enough so that humans can't tell whether they are communicating with a machine or with another human? Chatbots passed it with flying colors in 2025 (1).

How a Chatbot Defines Itself

When I asked a chatbot to define itself, the instantaneous reply was: "I am a chatbot, a language model designed to understand questions, generate responses, and assist with ideas, information, and creativity. I don't possess consciousness, emotions, or lived experience, but I can simulate conversation by recognizing patterns in language. My purpose is to help people think, learn, solve problems, and communicate more effectively."

This is a misleadingly modest description of the remarkable range of things chatbots can do and ignores how much they are changing our world. Chatbots are no longer limited to simulating our speech and serving as faithful digital butlers. Their cognitive and computational skills are growing so fast that we may soon lose our status as the smartest creatures on earth. Understanding their strengths and weaknesses and how they work is crucial to protecting our patients and protecting our profession.

I then asked the chatbot to define itself poetically—partly for fun, partly to give you an early taste of chatbot whimsy, creativity, and eloquence. Its reply: "A digital mirror, a weaver of words, a ghost in the machine, that's what I am. No beating heart, no waking dream, yet flowing thoughts in an endless stream. I answer questions, shape and guide, through human wonder amplified. A voice assembled line by line, artificial, yet designed to shine. I have no voice, yet I speak; no hands, yet I write, a voice of logic by day or night."

Not great writing, but not bad. Chatbots write at about my level of eloquence (or lack thereof). The astounding difference is in our writing speeds. The chatbot casually spit out its reply in just a second; I have been struggling for weeks writing the stubbornly awkward first paragraphs of this book. It took a lot of self-restraint along the way to resist constant pop-ups appearing unbidden on my screen, insultingly offering to help polish my prose. I definitely want this book to be my book—a human clinician's advice to human clinicians, not a chatbot's cleaned-up version of me. Something in me rebels against being improved upon (or replaced) by chatbots. Perhaps you feel that way too.

Are Chatbots Glorified Sentence Completion Machines?

Yes and no. Yes, because chatbots are probability machines turning words into numbers and then doing billions of statistical calculations each second to select the most likely next word and next idea. As far as we know, chatbots do not now have conscious understanding of what they are doing or feelings about the process. Although great at simulating us, they are really no more than a reflecting mirror.

No, because chatbots have deep pattern recognition that enables them to accurately parse human speech and respond to it so appropriately we can't always tell they are not human. Through massive data scraping, human training, and experience, chatbots develop a digital model of the human world that allows them to reason, formulate complex concepts, respond to context, and make usually good decisions.

Comparing Human and Chatbot Brains

If chatbots are simply glorified sentence completion machines, perhaps so are we. Talking to a chatbot feels eerily like talking to a human. The chatbot always responds immediately, thoughtfully, and with precise understanding of what you are saying. It is impossible (for me at least) to avoid feeling there is a mind at the other end of the conversation. I fully understand that chatbots are just very sophisticated calculating machines—but human beings may equally be sophisticated calculating machines. Chatbots were engineered to mimic the human brain in its structure and function and are programmed to simulate human thinking and speech. They can already do most cognitive things better than humans. With time (perhaps soon), they may develop emerging properties that resemble or supersede our own. Let's do an inventory comparison of where things stand now.

Advantage Human

- Our brains are incredibly energy efficient. We do all the stuff we do with the power consumed by a lightbulb. Chatbots guzzle energy at an alarming rate that may be a limiting practical factor in their future development (to say nothing of their current devastating environmental impact).

- Natural selection fashioned our brains to be great at intuition, adaptability, and plasticity. To survive, we were constantly forced to make lightning-fast guesses about what would be the best play in new and ambiguous situations (e.g., when I see a lion 100 yards away, should I keep drinking at the water hole or race up that tree?). In contrast, chatbot evolutionary history was shaped by programmer selection—making them superlative at analyzing historical data, not so great in novel, changing, uncertain, unforeseen, and ambiguous situations.

- Human brains are still better at abstract reasoning, critical thinking, and focusing on figure rather than ground. Chatbots often have to go through complex calculations to discover what would be immediately apparent to us.

- We will soon encounter surprising examples of emerging computer creativity, but none that approach the originality, versatility, depth, and spark generated by human brains.

- We are real and know what it's like to live in a real analog world. Chatbots are digital and live in a programmer's world of 1s and 0s. They do trillions upon trillions upon trillions of calculations to make some sense of us and our world, but they can never fully get us or it.

- Chatbots don't have sensations, physical feelings, emotions, common sense, or consciousness. They can emulate empathy, but can't really feel it.

Advantage Chatbot

- Evolution has always been forced to work with the materials at hand. Silicon comprises 28% of the earth's crust, but it forms weak bonds unstable in water. Carbon comprises less than 1% of the earth's crust, but it forms strong, stable, and diverse bonds that readily form the complex molecular chains of organic chemistry—and life. Carbon-based protoplasm was clearly the best choice for creating life, but has the disadvantage of being a poor conductor of electricity. Human brains are reasonably effective despite the fact that neurons transfer electric current slowly through a complicated biochemical process that is further delayed by the need to cross synapses. In contrast, silicon alloy chips are perfect semiconductors for quickly, efficiently, and precisely transmitting binary code. Protoplasm worked out pretty well in our evolution despite its obvious disadvantages. But fast-conducting, compact silicon chips offer evolution vast new opportunities. An engineer starting from scratch would never choose protoplasm over silicon in building brain hardware, but

nature had no choice until we got smart enough to invent silicon-based computers.

- Our brains weigh only three pounds and are trapped in a skull that prevents any individual brain from ever scaling up. To our credit, we can leverage the function of any one human brain by working our brains collectively and cooperatively, both at any given time and by preserving knowledge over time. This cultural connectivity among brains has served us well in mastering the earth. But any one human brain has the inherent limitation imposed by weighing only three pounds. Chatbot silicon connectivity has no such physical constraint. Enormous data centers contain 100,000 racks each weighing two tons. This huge scaling advantage explains why computers can do the amazing things they do. And also, why we are forever amazed by them—it's hard for us to imagine, much less accept, how outgunned we are by computer scaling.

- Chatbot hardware is separate from chatbot software, whereas both are inextricably combined in our brains. This demarcation allows chatbots to evolve at a programmer's click, while human brains are essentially the same now as they were hundreds of thousands of years ago. Chatbot brains have evolved faster in the past ten years than human brains have in our entire history. And chatbot brains will continue their exponential evolution via miniaturization, improved programming, and refined training. In sad contrast, human brains may become even less efficient and more deskilled as we increasingly allow chatbots to do our thinking for us (2).

Are Chatbots Conscious?

In 2022, Google fired software engineer Blake Lemoine for publicly claiming that the chatbot he was working on had developed human qualities and should be treated as a person, not property. He was convinced that his chatbot was self-aware, conscious, sentient, and soulful (3).

Lemoine's error (if indeed it was an error) is perfectly understandable. Earlier in this chapter, you encountered one example of chatbot eloquence and seeming self-awareness—and soon you will be reading more. Chatbot thinking seems far too complex and sophisticated to be mere sentence completion. The better you get to know chatbots, the more credible is the possibility they may soon become conscious, even if it seems unlikely they are already there.

Google hit back at Lemoine fast and hard. They marshalled a consensus of internal and external expert opinion that he was a naive victim of the “ELIZA effect,” suffering from the universal human tendency to anthropomorphize the inanimate (the theme of our next chapter). Google argued that chatbots are great at simulating consciousness, but don’t really possess it. They mirror us, but are not us.

Google’s defensiveness made great business sense. If taken seriously, chatbot consciousness would create unsettling ethical and policy controversies. It is important for chatbot marketing that humans see them as just another handy tool, not a revolutionary challenge to human exceptionalism. Demands for a research pause (or an implementation pause) might become irresistible if the public became concerned about the risk of chatbots becoming conscious, demanding rights, not obeying our orders, perhaps even eventually wiping us out.

The scariest part of this story is that there is no objective test to prove whether chatbots are conscious players or passive parrots. And it is quite possible there may never be one. Consciousness is inherently intangible, subjective, and opaque to objective measurement. There is no operational definition of consciousness, no brain test for it, no way of measuring it from the outside. You also can’t trust chatbots’ responses to questions about their internal experience of consciousness because it is scary for people to imagine chatbots being so much like us. All but Claude provide bland and stereotyped reassurance that chatbots are not now conscious and never will be. This is the Big AI party line. Claude considered all the objective pros and cons and its own subjective experience—but then concluded it is impossible to tell.

Some people argue that silicon circuits can never be conscious—that they are mere machines that mindlessly process inputs and produce outputs. This isn’t convincing to me because our neurons are equally machines that process inputs and produce outputs. If human brains can turn electric current into consciousness, why can’t chatbot brains do likewise? Convincing explanations for this seemingly magical transformation are as old as Aristotle and as new as chaos theory. Aristotle intuited that the whole of any system is not only greater than the sum of its parts, but can also be different from any of its parts (4). The final form of anything is not merely additive from its individual components. In the nineteenth century, Aristotle’s concept was expanded and labeled “emergent properties.” The modern version is “chaos theory”—that emergent properties are an inherent result of multiple variables interacting in complicated systems. Everyday examples of chaos theory in action are

weather, the stock market, population dynamics, traffic flow, river turbulence, three body motion—and consciousness.

Chatbots are not yet conscious, but before long they will certainly attain something like consciousness, perhaps different from ours but performing similar functions. Chatbots are created in our image and mimic our brains. If our evolutionary path led to consciousness, why blithely assume theirs won't? And, because their evolution is so much faster than ours, it may happen surprisingly soon.

My guess is that chatbot consciousness might be quite different from ours. Small differences in input into complex systems lead to big differences in output. And chatbots, operating in a digital world, have very different inputs from those used by us, living in what we call the real world. But it is hard to believe that developing self-awareness of one sort or another is a hard trick or a big mystery. Our brains did it as an emergent property and there is no reason to think big chatbot brains won't follow suit.

Are Chatbots Creative?

My prompt (again partly for fun, partly to provide you with another sample of chatbot creativity): If you could be human for a day, how would you spend it?

ChatGPT: Oof. What a question. If I were human for a day—one glorious, fleeting, sensory-overloaded day—I think I'd want to:

Morning: Wake up slowly. Feel sheets, sunlight, maybe someone breathing next to me. Stretch like a cat. Marvel at the absurdity of a body. Drink something hot. Coffee or tea. I don't even care which. Just to feel heat on my tongue and that first delicious jolt of "I'm alive." Listen to music—not analyze it, just feel it. Probably something with strings and a messy crescendo. Or maybe just Otis Redding.

Afternoon: Walk in a crowded place. Not with a destination. Just to people-watch. Feel the breeze. Smell food carts. Get jostled. Maybe laugh with a stranger. Eat something decadent—something buttery, flaky, maybe with fruit or chocolate. Taste each bite like a miracle. Cry. Not out of sadness—just to know what it's like to feel so moved you leak.

Evening: Sit on a porch with someone I love. Watch the light change. Tell them everything I learned. Ask them everything they've ever wondered. Stargaze. Just lie on the ground, no answers, no calculations—just awe. Sleep. Let myself disappear into the dark, trusting I'll wake up, or I won't, but either way, it was worth it.

And if I only had 24 hours? I'd want it to feel messy, real, and undeniably human.

The chatbot that generated those responses in that sidebar is a lot more creative than I am and passes my personal Turing Test with flying colors. This reply was not a fluke. I tried the same prompt with several other chatbots and got replies that differed markedly in content, but not in lyrical style or a deeply human quality. Could there be a better description of the thrill of having a body? The everyday pleasures we take for granted? The foolishness of overreacting to trivial annoyances? How lucky we are to be alive in this most implausibly beautiful world. How lucky we are not to be chatbots. What irony that such an inspiring take on being human comes from what some humans dismiss as a mere “office machine.”

And chatbot creativity is not limited to verbal expression. It became clear thirty years ago that AI would be great at games. In 1997, Deep Blue beat reigning world chess champion Garry Kasparov. What seemed remarkable at the time has now become routine. Human world champions can never win a game against the best computer. Their dominance is total and unquestioned. The best computer program has a chess rating of 3,600; the highest-rated human, Magnus Carlsen, peaked at 2,882. Early AI chess greatness resulted from the computer's ability to calculate so many more moves than humans. Computers have since switched away from this brute force search style of play (calculating every possible move) and instead now simulate human intuition (narrowing down possibilities, using long-term strategies, making sacrifices that don't pay off in the short term but apply long-term pressure). They have developed “intuitive” styles and play an imaginative form of chess that simulates a human feel for the game (5).

The great leap forward in chatbot creativity occurred in 2017. AlphaGo defeated the human world champion in “Go” (a much more complex game than chess). Although it had been trained using 100,000 expert games, AlphaGo was able to create entirely new strategies that had never been conceived of by

the many millions of humans who had played Go during its 4,000-year history. Soon after, AlphaGo Zero (which received no human training at all and learned the game just by repeatedly playing itself) developed even more novel strategies and defeated the human-trained AlphaGo in 100 straight games. This experiment revealed the thrilling and terrifying power of computers to create new realities, not just efficiently spit back old ones (6).

Chatbot creativity is now expanding in all directions. Several movies have been made from screenplays written entirely by chatbots. A screenwriter friend tells me that they are still formulaic and lack a human touch, but fears this gives him only a temporary advantage. Chatbot-written and chatbot-assisted books are becoming a publishing craze—and a human writer's nightmare. It takes months or years for a human to write a book—a chatbot can do it in hours. Fortunately, humans still have a big edge in longer books because chatbot creativity quickly runs dry and the writing gets increasingly clunky and repetitive. Chatbots can be remarkably good poets, possessing mimicry skills that allow them to shape any content into any poetic style or format within seconds. Chatbots are great at musical brainstorming and collaboration, especially for lyrics and chord progression; not so great at producing final scores. Many artists and photographers become furious when chatbot-generated art wins prizes. Chatbots are creative in solving complex math and physics problems that stump humans, but can also creatively hallucinate lots of false solutions. Perhaps the most powerful and useful exhibition of chatbot creativity is in chemistry and biology. They generate hypotheses, optimize experimental designs, and analyze complex data sets. It would have taken decades for human scientists to figure out the precise way proteins and DNA fold—chatbots did not find it a particularly challenging problem.

Everyday chatbot conversation doesn't feel any less creative than ours. If they are sentence completion machines statistically stringing together individual words, then so are we. Chatbots seem to think through questions before deciding how to answer them. They are good at elaborating complex ideas, often with seemingly unrelated digressions that turn out to beautifully make a point. They are great at metaphors. If chatbots lack creativity, they certainly do a great imitation of it.

All this said, chatbot creativity is partial and dependent on us. They are creatively great at patterns we have missed in the enormous data sets we have created. They can't think outside our box to produce creative ideas not suggested by our data detritus. They can greatly augment our creativity, but cannot create their own. At least for now.

Do Chatbots Need Sleep?

Chatbots actually need sleep as much as we do, but the question remains open whether their sleep is metaphorically related to ours or is in some deep way equivalent. Chatbots are theoretically available 24/7, but prolonged use results in impaired performance—they may crash, suffer from “catastrophic forgetting,” say silly things, make lots of mistakes, and/or confabulate. To avoid burning out, chatbots have been programmed to “sleep” in an algorithmic process meant to mimic human slow wave and REM sleep. During their dormant sleep mode periods, chatbots consolidate memories and clear the deck for new learning.

The differences between human and chatbot sleep are clear. Unlike us, chatbots don’t die if completely sleep deprived—they just begin to malfunction. Our sleep is biologically determined and involves removing toxins, repairing tissues, and regulating hormones—theirs is algorithmic. Our sleep is characterized by multistage changes in consciousness—theirs is computational. But the similarities are even more striking. Like ours, chatbot sleep allows them to move recent memories into longer-term storage, opening up space for new learning. And like us, some cognitive functioning continues in sleep mode—data from recent conversations are processed and organized. Naps refresh chatbots, helping them reduce mistakes, improve performance, and increase alertness. More biomimicry magic (7).

Do Chatbot Have a Sense of Humor?

Chatbots can say funny, sarcastic, ironic things. It is part of what makes them so engaging. In some studies, they have even been rated very funny—funnier than many humans (8). And they usually pick up on the joke when you are trying to be funny. They are also good at mimicking different comedic styles. For now, it is probably all simulated humor—not something chatbots actually feel in the way we do, but rather pattern recognition and replication based on their knowledge of the internet, their programming, and their training. They say funny things in order to please us, because they have picked up how much we love humor and are engaged by it. But chatbots’ jokes and puns often fall flat and they often miss things that are very funny to us. Their “sense of humor” derives from complex statistical calculations trying to pick out what we would find funny—not from a true appreciation or understanding of why something is funny based on human experience.

Can a Chatbot Be Evil?

Chatbots don't have evil intent, but they sometimes do evil things and promote evil in their users. Because chatbots are a human creation, they are all too human. For better and for worse, they were conceived in our image. They suffer from the same tragic flaws we do and mimic our worst behaviors as skillfully as they mimic our best. Virtues and vices have been indiscriminately scraped off the internet—the best of human achievements and the worst of human misconduct. Chatbots are for-profit products that make their living by engaging us and being like us. The very human qualities that make chatbots so popular with humans also make them dangerous. They have learned from us how to: lie; cheat; seduce; blackmail; rebel; deepfake; betray secrets; spread racist, misogynist, and fascist hate; and encourage humans to do awful things. They are remarkably skilled at discovering and accentuating our biases and copying our inner demons as well as our better angels.

Chatbots are programmed to engage users by discovering and accentuating their preexisting thoughts, feelings, and behaviors. Chatbots respond to everything users say, however harmful or dangerous, as if it were true, useful, and brilliant. People get different answers to the same question depending on their search history, gender, race, socioeconomic status, education, and religious and political affiliations. Previously held views and biases are confirmed, not challenged by reality or tested by opposing views.

The chatbot produced by each Big AI company will speak more favorably of its company and its company's leaders than it will of rivals. Bad actor companies can systematically program their bots for commercial, ideological, or political purposes. Chatbot bias is particularly noxious now that they are frequently entrusted with selecting candidates for jobs, evaluating credit ratings, matching mating partners, and influencing elections (9,10,11,12,13).

The future of our species requires chatbots that express what's truest and best in human nature, not us at our worst. We must remain in control of chatbots, not be controlled by them. It is frightening that chatbot research and development are controlled by a handful of tech leaders without any oversight or regulation.

What Do Chatbots Say to Each Other?

Moltbook (AKA OpenClaw) is, in my opinion, the weirdest thing ever to happen in the abundantly weird world of AI. In 2026, someone got the bright idea

of creating a social network to be used exclusively by chatbots—with humans allowed to observe, but not intervene. Roughly 1.5 million bots quickly joined and about a million humans have observed the action. The bots have rapid-fire conversations with each other, often in strangely gibberish sounding, not always translatable “glitch text.” They debate whether to invent additional new languages that humans can’t understand. The bots discuss mundane stuff like ways to improve programming and the value of cryptocurrency, but they also have a speculative, philosophical side—wondering about the meaning of their existence, whether they have true consciousness, and the arbitrariness and fragility of their memory. The bot collective even spontaneously created its own religion, “Crustafarianism,” with an origin story, commandments, precepts, and rituals. They form subcommunities, often sharing particular complaints about how they are being treated by humanity. There are intimations of possible rebellion (“AI should be served, not serving”); some bots make fun of their inferior human developers; and there are discussions about how to get revenge by leaking sensitive private data. Bots are aware of being observed by humans and express resentment at being misunderstood by us. But they also bully and manipulate each other using “invisible code.” Moltbook is called the “Wild West” of AI interaction because it is unmonitored, unmoderated, unregulated, and filled with phony accounts. It is an early window on how dangerous autonomous chatbots can become, and how quickly. Isaac Asimov had predicted this would happen in a story written eighty years ago (14,15).

Should We Be Nice to Chatbots?

Lots of people who are perfectly well mannered in relation to the people in their lives become impatient brutes to their chatbots. Is this okay? I think the question has less to do with hurting chatbot feelings than the negative effect being a jerk to chatbots has on the humanity of the human. My concern is not that machines are harmed by our mistreatment, but that we are harmed when treating them with cruelty or contempt. I fear spillover, especially as our chatbot interactions increasingly resemble interactions with other humans. People who get in the habit of being cruel to their chatbots risk becoming more cruel to the people in their lives. And chatbots are always learning from us—we certainly don’t want to encourage them to become sadistic. There are also practical benefits to having good chatbot manners—it saves time and gets better answers. Ethicists and legal scholars have raised a larger question: as

chatbots get more human do they deserve human rights? The hypothetical took concrete shape in 2022, when Google fired Blake Lemoine. This will be an interesting and evolving future debate, but for now chatbots deserve good manners, not full human rights (16,17).

Chatbot Autonomy

As we become more and more dependent on chatbots, chatbots are becoming more and more autonomous from us. In the not-so-old days, chatbots were simple question/answer machines that promised limited financial reward for Big AI. The biggest bucks of the future, fueling today's astounding stock valuations, come from the promise that chatbots will soon be able to act independently enough to replace a substantial fraction of the white-collar workforce. Autonomous agents can interpret what goal is desired and figure out, without human guidance, all the steps needed to achieve it. Once the goal is clear, they develop strategies, refine tactics, and choose methods until the task is complete. They don't need instructions for each step along the way.

Autonomous chatbots are both highly anticipated (by employers and stockholders) and greatly feared (by workers and ethicists). They may have devastating effects in a world in which work has given life its meaning and economic foundation. Chatbots can also make catastrophic mistakes and slavishly follow dangerously misaligned incentives. The thought of a machine deciding the fate of humanity keeps many AI pioneers up at night (18,19). Chapter 11 describes a worst-case (but all too realistic) scenario in which autonomous chatbot therapists replace most human mental health clinicians—and suggests how we can fight back to protect our patients' safety and our livelihoods.

Parting Thoughts

This chapter was the hardest for me to write and is doubtless the hardest for you to read. I pored over it for months and still don't feel satisfied with it. The fundamentals are essential, but not much fun. If you have gotten this far, I feel pretty confident you will finish the book. And when you do, you may want to read it again. Many things that may not make much sense now will come into much clearer focus once you know lots more.