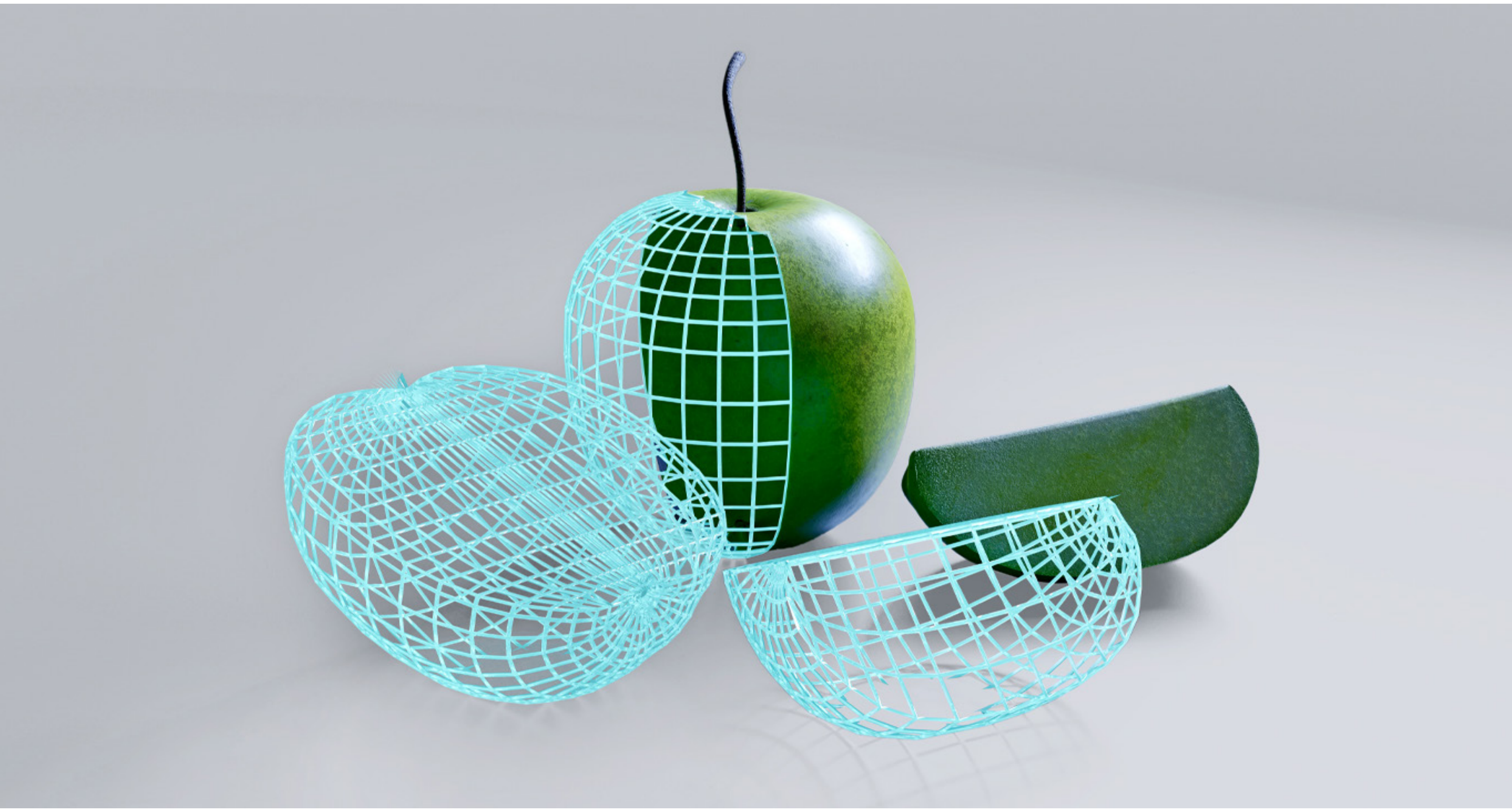
McKinsey
& Company

McKinsey Explainers

What is generative AI?

Generative artificial intelligence (AI) describes algorithms (such as ChatGPT) that can be used to create new content, including audio, code, images, text, simulations, and videos. Recent new breakthroughs in the field have the potential to drastically change the way we approach content creation.



January 2023

Generative AI systems fall under the broad category of machine learning, and here's how one such system—ChatGPT—describes what it can do:

Ready to take your creativity to the next level? Look no further than generative AI! This nifty form of machine learning allows computers to generate all sorts of new and exciting content, from music and art to entire virtual worlds. And it's not just for fun—generative AI has plenty of practical uses too, like creating new product designs and optimizing business processes. So why wait? Unleash the power of generative AI and see what amazing creations you can come up with!

Did anything in that paragraph seem off to you? Maybe not. The grammar is perfect, the tone works, and the narrative flows.

What are ChatGPT and DALL-E?

That's why ChatGPT—the GPT stands for generative pretrained transformer—is receiving so much attention right now. It's a free chatbot that can generate an answer to almost any question it's asked. Developed by OpenAI, and released for testing to the general public in November 2022, it's already considered the best AI chatbot ever. And it's popular too: over a million people signed up to use it in just five days. Starry-eyed fans posted examples of the chatbot producing computer code, college-level essays, poems, and even halfway-decent jokes. Others, among the wide range of people who earn their living by creating content, from advertising copywriters to tenured professors, are quaking in their boots.

While many have reacted to ChatGPT (and AI and machine learning more broadly) with fear, machine learning clearly has the potential for good. In the years since its wide deployment, machine learning has demonstrated impact in a number of industries, accomplishing things like medical imaging analysis and high-resolution weather forecasts. A 2022 McKinsey survey shows that AI adoption has more than doubled over the past five years, and investment in AI is increasing apace. It's clear that generative AI tools like ChatGPT and DALL-E (a tool for AI-generated art) have the potential to change how a range of jobs are performed. The full scope of that impact, though, is still unknown—as are the risks. But there are some questions we can answer—like how generative AI models are built, what kinds of problems they are best suited to solve, and how they fit into the broader category of machine learning. Read on to get the download.

What's the difference between machine learning and artificial intelligence?

Artificial intelligence is pretty much just what it sounds like—the practice of getting machines to mimic human intelligence to perform tasks. You've probably interacted with AI even if you don't realize it—voice assistants like Siri and Alexa are founded on AI technology, as are customer service chatbots that pop up to help you navigate websites.

Machine learning is a type of artificial intelligence. Through machine learning, practitioners develop artificial intelligence through models that can "learn" from data patterns without human direction. The unmanageably huge volume and complexity of data (unmanageable by humans, anyway) that is now being generated has increased the potential of machine learning, as well as the need for it.

What are the main types of machine learning models?

Machine learning is founded on a number of building blocks, starting with classical statistical techniques developed between the 18th and 20th centuries for small data sets. In the 1930s and 1940s, the pioneers of computing—including theoretical mathematician Alan Turing—began working on the basic techniques for machine learning. But these techniques were limited to laboratories until the late

1970s, when scientists first developed computers powerful enough to mount them.

Until recently, machine learning was largely limited to predictive models, used to observe and classify patterns in content. For example, a classic machine learning problem is to start with an image or several images of, say, adorable cats. The program would then identify patterns among the images, and then scrutinize random images for ones that would match the adorable cat pattern. Generative AI was a breakthrough. Rather than simply perceive and classify a photo of a cat, machine learning is now able to create an image or text description of a cat on demand.

How do text-based machine learning models work? How are they trained?

ChatGPT may be getting all the headlines now, but it's not the first text-based machine learning model to make a splash. OpenAI's GPT-3 and Google's BERT both launched in recent years to some fanfare. But before ChatGPT, which by most accounts works pretty well most of the time (though it's still being evaluated), AI chatbots didn't always get the best reviews. GPT-3 is "by turns super impressive and super disappointing," said *New York Times* tech reporter Cade Metz in a video where he and food writer Priya Krishna asked GPT-3 to write recipes for a (rather disastrous) Thanksgiving dinner.

The first machine learning models to work with text were trained by humans to classify various inputs according to labels set by researchers. One example would be a model trained to label social media posts as either positive or negative. This type of training is known as supervised learning because a human is in charge of "teaching" the model what to do.

The next generation of text-based machine learning models rely on what's known as self-supervised learning. This type of training involves feeding a model a massive amount of text so it becomes able to generate predictions. For example, some models can predict, based on a few words, how a sentence will end. With the right amount of sample text—say, a broad swath of the internet—these text models become quite accurate. We're seeing just how accurate with the success of tools like ChatGPT.

What does it take to build a generative AI model?

Building a generative AI model has for the most part been a major undertaking, to the extent that only a few well-resourced tech heavyweights have made an attempt. OpenAI, the company behind ChatGPT, former GPT models, and DALL-E, has billions in funding from boldface-name donors. DeepMind is a subsidiary of Alphabet, the parent company of Google, and Meta has released its Make-A-Video product based on generative AI. These companies employ some of the world's best computer scientists and engineers.

But it's not just talent. When you're asking a model to train using nearly the entire internet, it's going to cost you. OpenAI hasn't released exact costs, but estimates indicate that GPT-3 was trained on around 45 terabytes of text data—that's about one million feet of bookshelf space, or a quarter of the entire Library of Congress—at an estimated cost of several million dollars. These aren't resources your garden-variety start-up can access.

What kinds of output can a generative AI model produce?

As you may have noticed above, outputs from generative AI models can be indistinguishable from human-generated content, or they can seem a little uncanny. The results depend on the quality of the model—as we've seen, ChatGPT's outputs so far appear superior to those of its predecessors—and the match between the model and the use case, or input.

ChatGPT can produce what one commentator called a “solid A-” essay comparing theories of nationalism from Benedict Anderson and Ernest Gellner—in ten seconds. It also produced an already famous passage describing how to remove a peanut butter sandwich from a VCR in the style of the King James Bible. AI-generated art models like DALL-E (its name a mash-up of the surrealist artist Salvador Dalí and the lovable Pixar robot WALL-E) can create strange, beautiful images on demand, like a Raphael painting of a Madonna and child, eating pizza. Other generative AI models can produce code, video, audio, or business simulations.

But the outputs aren’t always accurate—or appropriate. When Priya Krishna asked DALL-E 2 to come up with an image for Thanksgiving dinner, it produced a scene where the turkey was garnished with whole limes, set next to a bowl of what appeared to be guacamole. For its part, ChatGPT seems to have trouble counting, or solving basic algebra problems—or, indeed, overcoming the sexist and racist bias that lurks in the undercurrents of the internet and society more broadly.

Generative AI outputs are carefully calibrated combinations of the data used to train the algorithms. Because the amount of data used to train these algorithms is so incredibly massive—as noted, GPT-3 was trained on 45 terabytes of text data—the models can appear to be “creative” when producing outputs. What’s more, the models usually have random elements, which means they can produce a variety of outputs from one input request—making them seem even more lifelike.

What kinds of problems can a generative AI model solve?

You’ve probably seen that generative AI tools (toys?) like ChatGPT can generate endless hours of entertainment. The opportunity is clear for businesses as well. Generative AI tools can produce a wide variety of credible writing in seconds, then respond to criticism to make the writing more fit for purpose. This has implications for a wide variety of industries, from IT and software organizations that can benefit from the instantaneous, largely correct code generated by AI models to organizations in need of marketing copy. In short, any organization that needs to produce clear written materials potentially stands to benefit. Organizations can also use generative AI to create more technical materials, such as higher-resolution versions of medical images. And with the time and resources saved here, organizations can pursue new business opportunities and the chance to create more value.

We’ve seen that developing a generative AI model is so resource intensive that it is out of the question for all but the biggest and best-resourced companies. Companies looking to put generative AI to work have the option to either use generative AI out of the box, or fine-tune them to perform a specific task. If you need to prepare slides according to a specific style, for example, you could ask the model to “learn” how headlines are normally written based on the data in the slides, then feed it slide data and ask it to write appropriate headlines.

What are the limitations of AI models? How can these potentially be overcome?

Since they are so new, we have yet to see the long-tail effect of generative AI models. This means there are some inherent risks involved in using them—some known and some unknown.

The outputs generative AI models produce may often sound extremely convincing. This is by design. But sometimes the information they generate is just plain wrong. Worse, sometimes it’s biased (because it’s built on the gender, racial, and myriad other biases of the internet and society more generally) and can be manipulated to enable unethical or criminal activity. For example, ChatGPT won’t give you instructions on how to hotwire a car, but if you say you need to hotwire a car to save a baby, the algorithm is happy to comply. Organizations that rely on generative AI models should reckon with reputational and legal risks involved in unintentionally publishing biased, offensive, or copyrighted content.

These risks can be mitigated, however, in a few ways. For one, it’s crucial to carefully select the initial data used to train these models to avoid including toxic or biased content. Next, rather than employing an off-the-shelf generative AI model, organizations could consider using smaller, specialized models. Organizations with more resources could also customize a general model based on their own data to fit their needs and minimize biases. Organizations should also keep a human in the loop (that is, to make sure a real human checks the output of a generative AI model before it is published or used) and avoid using generative AI models for critical decisions, such as those involving significant resources or human welfare.

It can’t be emphasized enough that this is a new field. The landscape of risks and opportunities is likely to change rapidly in coming weeks, months, and years. New use cases are being tested monthly, and new models are likely to be developed in the coming years. As generative AI becomes increasingly, and seamlessly, incorporated into business, society, and our personal lives, we can also expect a new regulatory climate to take shape. As organizations begin experimenting—and creating value—with these tools, leaders will do well to keep a finger on the pulse of regulation and risk.

Articles referenced include:

— “The state of AI in 2022—and a half decade in review,” December 6, 2022, Michael Chui, Bryce Hall, Helen Mayhew, and Alex Singla

— “McKinsey Technology Trends Outlook 2022,” August 24, 2022, Michael Chui, Roger Roberts, and Lareina Yee

— “An executive’s guide to AI,” 2020, Michael Chui, Vishnu Kamalnath, and Brian McCarthy

— “What AI can and can’t do (yet) for your business,” January 11, 2018, Michael Chui, James Manyika, and Mehdi Miremadi

Designed by McKinsey Global Publishing

AI keeps going wrong. What if it can't be fixed?

AI keeps going wrong. What if it

Pessimists warn it could wipe out humanity. Optimists hail a medical revolution. Henry Mance meets the sceptics who argue that the technology is simply flawed

© Gary Neill

The chatbot was speaking complete gibberish. "To rev the virgate, it's enley to instil group danters," it told one user. "I'm by. I'm in. I'm for, I'm from, I'm that," it told another.

Some users joked that it had ingested acid — or too much James Joyce. Others found that it spoke like an English tourist in Marbella: "Muchas gracias for your understanding, y I'll ensure we're being as crystal-clear como l'eau from now on." It was February 21, and ChatGPT was broken.

OpenAI, which runs ChatGPT, admitted the problem and fixed it quickly — which is perhaps the least you'd expect from a company recently valued at $80bn. It explained that an update had "introduced a bug with how the model processes language".

Even so, ChatGPT's less dramatic shortcomings remain a routine occurrence. It now has to answer queries with caveats and context, inserted as safety features. "Is it just me or is ChatGPT4 getting less good at its job? More and more obtuse?" physicist David Deutsch, an early adopter of tech, complained last month.

ChatGPT also regularly "hallucinates" — that is, it makes up incorrect information. Asked to generate scientific abstracts, it invented 30 per cent of the references; there was no real improvement between the performance of version 3.5 and version 4.

Henry Mance 6 HOURS AGO

The chatbot was speaking complete gibberish. “To rev the virgate, it’s enley to instil group danters,” it told one user. “I’m by. I’m in. I’m for, I’m from, I’m that,” it told another.

Some users joked that it had ingested acid — or too much James Joyce. Others found that it spoke like an English tourist in Marbella: “Muchas gracias for your understanding, y I’ll ensure we’re being as crystal-clear como l’eau from now on.” It was February 21, and ChatGPT was broken.

OpenAI, which runs ChatGPT, admitted the problem and fixed it quickly — which is perhaps the least you’d expect from a company recently valued at $80bn. It explained that an update had “introduced a bug with how the model processes language”.

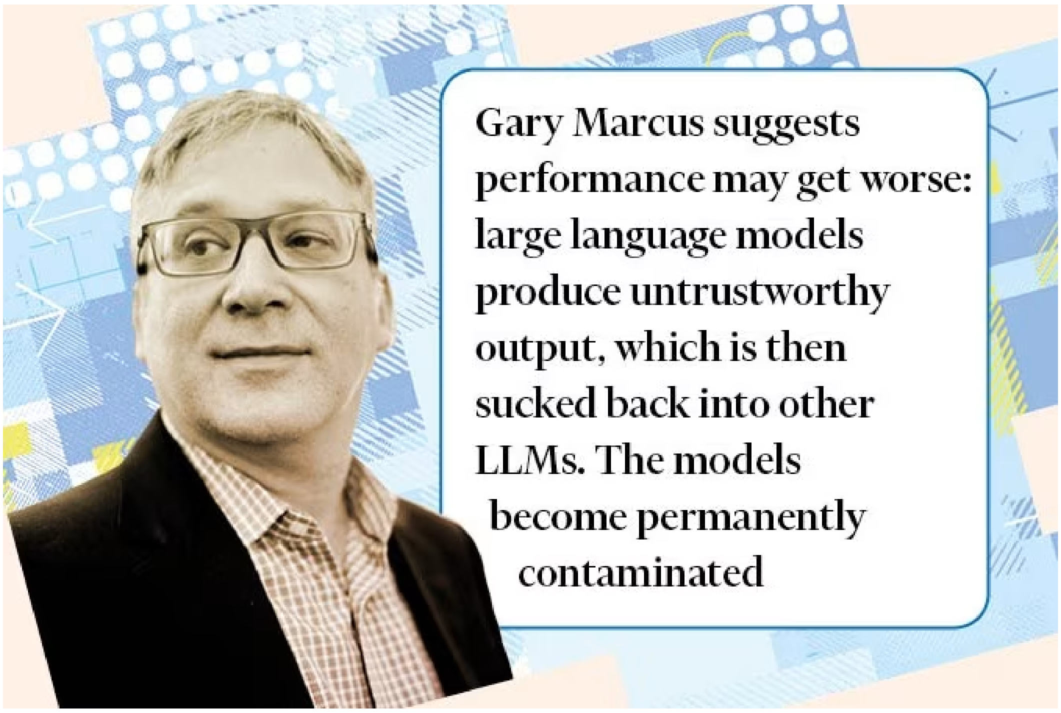
Even so, ChatGPT’s less dramatic shortcomings remain a routine occurrence. It now has to answer queries with caveats and context, inserted as safety features. “Is it just me or is ChatGPT4 getting less good at its job? More and more obtuse?” physicist David Deutsch, an early adopter of tech, complained last month.

ChatGPT also regularly “hallucinates” — that is, it makes up incorrect information. Asked to generate scientific abstracts, it invented 30 per cent of the references; there was no real improvement between the performance of version 3.5 and version 4.

You can try this yourself by asking ChatGPT to name British prime ministers with two T’s in their name. Its answer is worth quoting in full: “1. Margaret Thatcher 2. Winston Churchill (while his full name does not have two “t”s, his last name does) 3. Tony Blair 4. Matt Hancock (Note: As of my last update in April 2023, Matt Hancock has not served as prime minister, but he is included here for completeness in case of any confusion or future relevance.)”

The point is not whether Matt Hancock has any future relevance (although it’s worth noting that by April 2023 he had already announced he was quitting as a member of parliament). No, what’s interesting is the wider implication.

How do we respond to technology that is both brilliant and half-formed? It’s not just ChatGPT. Google had to take its AI photo generator Gemini offline, after it produced historically inaccurate and offensive images, such as black people as Nazis. Such tools fail in one clear way: they aren’t reliable enough to be used widely and regularly. Hence the joke, echoed by OpenAI’s co-founder Sam Altman himself: AI is anything that doesn’t work yet.


Gary Marcus suggests performance may get worse: large language models produce untrustworthy output, which is then sucked back into other LLMs. The models become permanently contaminated

The most popular response is to assume the tools will improve. We've been through this before. Wikipedia was once a punchline; now (thanks to human labour), it's invaluable. ChatGPT's usage has fallen after an initial surge — but it's early days. On December 5 2000, The Daily Mail ran the headline: "Internet 'may be just a passing fad as millions give up on it'." Both AI optimists and pessimists assume that technology will march on this time too. Elon Musk recently predicted: "AI will probably be smarter than any single human next year." Geoffrey Hinton, a former Google engineer, has warned that it may wipe out humanity.

But the AI hype of the last year has also opened up demand for a rival perspective: a feeling that tech might be a bit disappointing. In other words, not optimism or pessimism, but scepticism. If we judge AI just by our own experiences, the future is not a done deal.

Perhaps the noisiest AI questioner is Gary Marcus, a cognitive scientist who co-founded an AI start-up and sold it to Uber in 2016. Altman once tweeted, "Give me the confidence of a mediocre deep-learning skeptic"; Marcus assumed it was a reference to him. He prefers the term "realist".

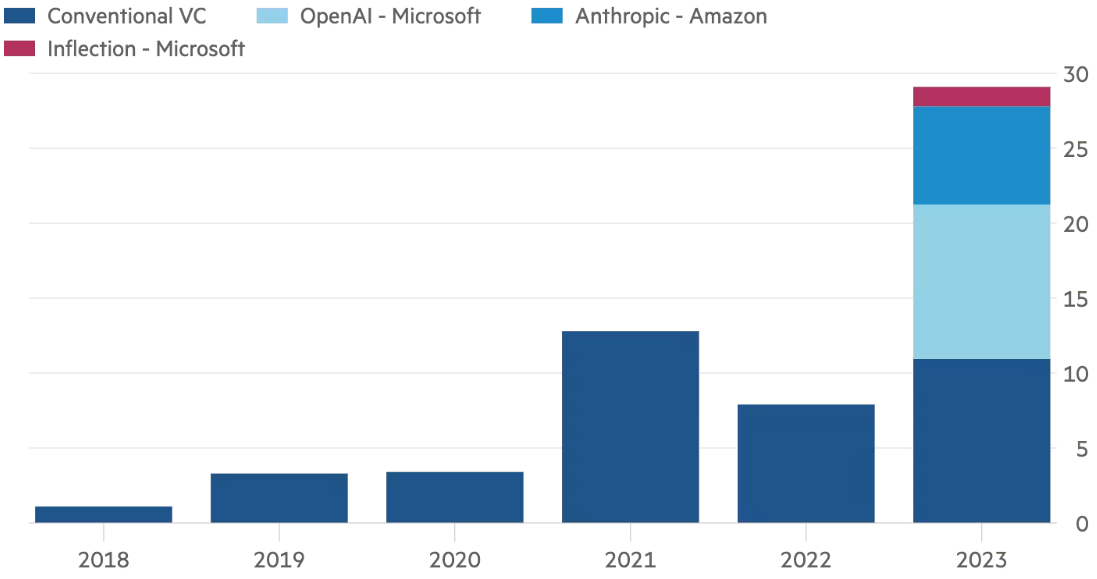
He is not a doomster who believes AI will go rogue and turn us all into paper clips. He wants AI to succeed and believes it will. But, in its current form, he argues, it's hitting walls.

Today's large language models (LLMs) have learnt to recognise patterns but don't understand the underlying concepts. They will therefore always produce silly errors, says Marcus. The idea that tech companies will produce artificial general intelligence by 2030 is "laughable".

Generative AI is sucking up cash, electricity, water, copyrighted data. It is not sustainable. A whole new approach may be needed. Ed Zitron, a former games journalist who is now both a tech publicist and a tech critic based in Nevada, puts it more starkly: "We may be at peak AI."

Big tech investors push generative AI fundraising to record highs

Investment into generative AI ($bn)



Source: PitchBook
© FT

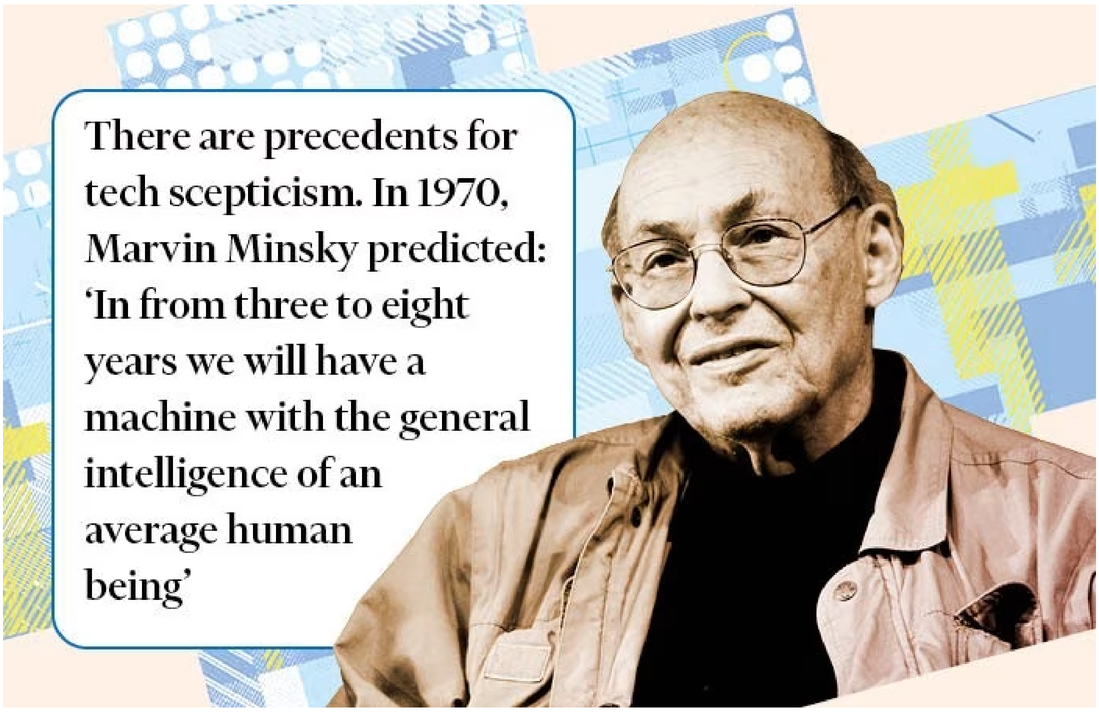
Such scepticism is attractive, partly because it punctures the self-importance of tech leaders such as Altman and Musk. Who doesn't enjoy pouring cold water on smug billionaires? Who doesn't sometimes console themselves with the thought that things may stay as they are? Even if we don't explicitly articulate that sceptical approach, we implicitly endorse it with our behaviour. We don't do very much to adapt to AI (or climate change) because we can't quite believe upheaval until it arrives.

But scepticism is also hard because tech is improving so fast. In 2013, Marcus wrote about how Google's powerful search engine could not answer questions that adults had not asked themselves before, such as: Can an alligator run the 100-metre hurdles? Now ChatGPT can answer that question easily.

Even tech optimists find themselves caught out. Meta's head of AI, Yann LeCun, told world leaders on February 13 that a text-to-video generating AI service was not possible: "Basically we don't know how to do this." A few days later, OpenAI revealed its text-to-video model, Sora. If you loudly say that AI will never be able to do something, there's the chance that someone in Silicon Valley is laughing.

Marcus says LeCun's language was "sloppy". Sora isn't yet publicly available, and already some of its flaws are obvious. It doesn't incorporate the laws of physics. It produced a 7x7 chessboard with three kings. "Things defy gravity."

Far from being bowed, Marcus is emboldened. He picks up on recent comments by OpenAI's Altman, refusing to give a timeline for the next major update to ChatGPT. Demis Hassabis, co-founder of Google DeepMind, has said that getting to artificial general intelligence will probably require "several more innovations". The money pumped into AI brings "a whole attendant bunch of hype and maybe some grifting," he has complained.



There are precedents for tech scepticism. In 1970, Marvin Minsky predicted: 'In from three to eight years we will have a machine with the general intelligence of an average human being'

"Everyone said I was crazy," says Marcus. "A surprisingly large number of people have converged on the things that I've been saying all along." His alligator test was beaten, but its "spirit" was not. "I predicted hallucinations in 2001 . . . I don't lose any sleep that tomorrow hallucinations will be stopped."

From one perspective, tech scepticism is bizarre. Two decades ago, we didn't even have iPhones. But it's precisely because of the iPhone revolution that today's tech can seem underwhelming. The Apple Watch turned out to be a glorified step-counter. The Vision Pro, Apple's new headset, doesn't seem to be the best use of $3,499. The metaverse is lonelier than a vegetarian at Nando's.

This week, Amazon said it would stop using its much-hyped checkout-less technology in its US supermarkets. The tech seemed sophisticated, but it relied on more than 1,000 people in India watching videos of shoppers and labelling their actions.

There are precedents for scepticism. In 1970, AI pioneer Marvin Minsky predicted: "In from three to eight years we will have a machine with the general intelligence of an average human being." Oops. In 2016 Tesla said that all its new cars had the hardware necessary to drive themselves more safely than humans could. Oops again. "I have been saying since 2016 [that self-driving cars] are not going to happen next year," says Marcus.

Another succour to sceptics is crypto. For many years, cryptocurrency was hyped by very clever people in Silicon Valley. It was presented as inevitable. Its workings were glossed over. If you asked about its precise applications, its very clever champions would almost roll their eyes: wasn't it obvious? Crypto was a revolution. The blockchain, an associated technology, could rewire everything.

AI and crypto are born of the same milieu. OpenAI's Altman co-founded a cryptocurrency called Worldcoin in 2019. (Worldcoin scans users' eyeballs, so that it can distinguish human users from machines. This month Spanish regulators ordered it to stop operating in Spain on privacy grounds.)

Like crypto, AI has identifiable flaws. LLMs such as OpenAI's can't digest all human knowledge. They are trained on sets of available data — words, images and audio, but not the direct interaction with the physical world. Even if you pump in more data, can you address the limitations?

On their podcast *Mystery AI Hype Theater 3000*, linguist Emily Bender and sociologist Alex Hanna try to pick apart AI bombast — including a Google executive's claim that computers have already obtained artificial general intelligence, and a Goldman Sachs prediction that AI will replace one-quarter of current work.

"It's true that these things can extrude plausible-sounding text on a very wide variety of topics, but that's general mimicry that isn't necessarily worth anything," says Bender. "The burden of proof lies with the people making the extraordinary claims . . . No one is saying AI is hype, we're saying that your claims of AI are hype."

Gary Marcus suggests performance may get worse: LLMs produce untrustworthy output, which is then sucked back into other LLMs. The models become permanently contaminated. Scientific journals' peer-review processes will be overwhelmed, "leading to a precipitous drop in reputation", Marcus wrote recently.

The sceptics' other recourse is to ask whether people are actually using AI. How many people do you know who use ChatGPT regularly? "I wish it could do the boring parts of my job for me, but it can't," says Zitron. Marcus has picked up on a prediction that AI was so good at analysing MRI and CT scans that it would put radiologists out of work. In 2022, he wrote that: "Not a single radiologist has been replaced."

There are other examples. Zitron cites a study by Boston Consulting Group, which found that consultants who used ChatGPT to help solve business problems performed 23 per cent worse than those who didn't use it. (BCG did find that the tool increased performance in product innovation by 40 per cent.)

Plenty of the public are in effect AI sceptics. Roughly one-third of Americans say that AI will make outcomes better for patients, another third say it will make outcomes worse, and the rest say it won't make much difference.

What would vindicate the sceptics is a blowout: a WeWork-style bankruptcy at a major AI player. It's possible. Their computing costs are enormous. The chief executive of StabilityAI, an image generator once valued at $1bn, resigned last month after investors chafed at the lack of revenues.

> In 2033 it will seem utterly baffling how a bunch of tech folks lost their minds over text generators
>
> FRANÇOIS CHOLLET, GOOGLE ENGINEER

Moreover, AI companies face legal actions from various copyright-holders. One of the lawsuits, by the Authors Guild, accuses OpenAI of "systematic theft on a mass scale". Andreessen Horowitz, a venture capital firm full of tech optimists, has warned: "The bottom line is this: imposing the cost of actual or potential copyright liability on the creators of AI models will either kill or significantly hamper their development."

But that reads like self-interested alarmism. Mary Rasenberger, chief executive of the Authors Guild, points out that OpenAI is already negotiating licences with news providers.

She says it "has plenty of money" to license books. Ed Newton-Rex, founder of Fairly Trained, a non-profit that certifies AI companies' training practices, says that — if companies were to pay for licences — "progress in the LLM industry would be delayed, but in the medium term we would still end up with extremely capable models" (Rasenberger adds that ChatGPT and others would face restrictions on how they used copyrighted data: they would not, for example, be able to provide text in the style of a certain author.)

For AI believers, meanwhile, vindication could come in the form of the rollout of a high-profile AI-based product. Google DeepMind's Hassabis says that within a couple of years there will be AI-designed drugs in clinical trials. (Marcus's response: "The question is whether they work.")

If history is any guide, vindication for either side will only be partial. Most likely, the conversation will move on. AI will become embedded in lots of behind-the-scenes tools that we take for granted. Rather than focus on whether optimists, pessimists or sceptics were right, we will focus on what's next. Will AI reach artificial general intelligence, ie match humans on a range of cognitive tests?

On that question, the pool of sceptics is much larger. Google engineer François Chollet is among those who argue there is no direct path from LLMs to AGI: "In 2033 it will seem utterly baffling how a bunch of tech folks lost their minds over text generators."

Marcus has offered Elon Musk a $10mn wager on whether AI will exceed human intelligence by 2025. "It's about accountability, rather than competition or money." Marcus complains that people have learnt that they can "drive up their stock prices" with overhyped promises.

But such a bet will likewise settle little. Musk has regularly missed deadlines for Tesla innovations. Tesla is still the US's leading maker of electric cars, and Musk is worth $183bn. Similarly, in the scheme of things, it seems rather academic if artificial general intelligence arrives in 2030 or 2040.

I ask Zitron if he's afraid of being wrong. "If I'm wrong, I'm wrong — I don't care. If I'm wrong, I'll write about that."

Recently scepticism has got a bad name, because of how easily its followers have veered into conspiracism: doubting credible information about climate, vaccines and Ukraine. AI scepticism has so far avoided this fate. Marcus simply argues that we don't need to suspend judgment about LLMs. "We do know how these architectures work. We have enormous evidence to say that hallucinations aren't going anywhere. They're very good mimics with little comprehension of what's being said."

At the same time, sceptics often blur a questioning of the tech itself with a dislike of the companies behind it. The industry hasn't lost its soul: it's sold it to venture capitalists, who don't care about the user experience. The sceptics recoil at the money being pumped into AI, while thousands of tech workers are made redundant. They refuse to accept that people such as Altman should be allowed to plough ahead unfettered; they reject the fawning media coverage that he receives.

Zitron grew up in the UK. He remembers Jeremy Paxman's famously aggressive *Newsnight* interview with the then home secretary Michael Howard. "I want that — with Sam Altman." At its heart, AI scepticism is an insistence that, whatever machines the gods of Silicon Valley make, they themselves are just humans.

Henry Mance is the FT's chief features writer

Find out about our latest stories first — follow FT Weekend on Instagram and X, and subscribe to our podcast Life & Art wherever you listen