The New York Times https://www.nytimes.com/2023/03/15/technology/gpt-4-artificial-intelligence-openai.html

THE SHIFT

GPT-4 Is Exciting and Scary

Today, the new language model from OpenAI may not seem all that dangerous. But the worst risks are the ones we cannot anticipate.

By Kevin Roose

March 15, 2023

When I opened my laptop on Tuesday to take my first run at GPT-4, the new artificial intelligence language model from OpenAI, I was, truth be told, a little nervous.

After all, my last extended encounter with an A.I. chatbot — the one built into Microsoft's Bing search engine — ended with the chatbot trying to break up my marriage.

It didn't help that, among the tech crowd in San Francisco, GPT-4's arrival had been anticipated with near-messianic fanfare. Before its public debut, for months rumors swirled about its specifics. "*I heard it has 100 trillion parameters.*" "*I heard it got a 1,600 on the SAT.*" "*My friend works for OpenAI, and he says it's as smart as a college graduate.*"

These rumors may not have been true. But they hinted at how jarring the technology's abilities can feel. Recently, one early GPT-4 tester — who was bound by a nondisclosure agreement with OpenAI but gossiped a little anyway — told me that testing GPT-4 had caused the person to have an "existential crisis," because it revealed how powerful and creative the A.I. was compared with the tester's own puny brain.

GPT-4 didn't give me an existential crisis. But it exacerbated the dizzy and vertiginous feeling I've been getting whenever I think about A.I. lately. And it has made me wonder whether that feeling will ever fade, or whether we're going to be experiencing "future shock" — the term coined by the writer Alvin Toffler for the feeling that too much is changing, too quickly — for the rest of our lives.

For a few hours on Tuesday, I prodded GPT-4 — which is included with ChatGPT Plus, the $20-a-month version of OpenAI's chatbot, ChatGPT — with different types of questions, hoping to uncover some of its strengths and weaknesses.

I asked GPT-4 to help me with a complicated tax problem. (It did, impressively.) I asked it if it had a crush on me. (It didn't, thank God.) It helped me plan a birthday party for my kid, and it taught me about an esoteric artificial intelligence concept known as an "attention head." I even asked it to come up with a new word that had never before been uttered by humans. (After making the disclaimer that it couldn't verify every word ever spoken, GPT-4 chose "flembostriquat.")

Some of these things were possible to do with earlier A.I. models. But OpenAI has broken new ground, too. According to the company, GPT-4 is more capable and accurate than the original ChatGPT, and it performs astonishingly well on a variety of tests, including the Uniform Bar Exam (on which GPT-4 scores higher than 90 percent of human test-takers) and the Biology Olympiad (on which it beats 99 percent of humans). GPT-4 also aces a number of Advanced Placement exams, including A.P. Art History and A.P. Biology, and it gets a 1,410 on the SAT — not a perfect score, but one that many human high schoolers would covet.

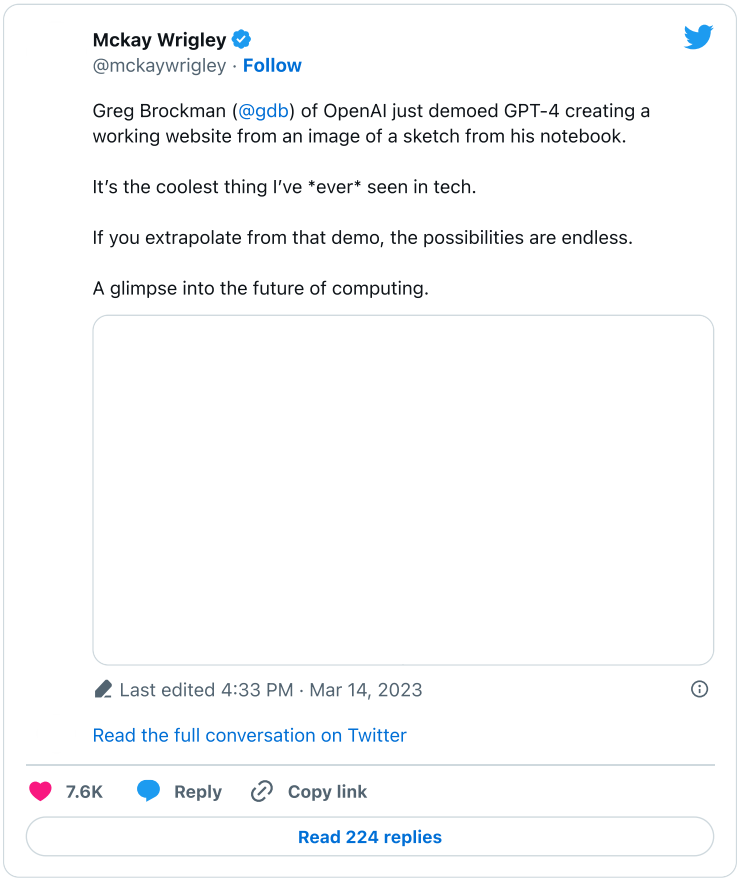
You can sense the added intelligence in GPT-4, which responds more fluidly than the previous version, and seems more comfortable with a wider range of tasks. GPT-4 also seems to have slightly more guardrails in place than ChatGPT. It also appears to be significantly less unhinged than the original Bing, which we now know was running a version of GPT-4 under the hood, but which appears to have been far less carefully fine-tuned.

Unlike Bing, GPT-4 usually flat-out refused to take the bait when I tried to get it to talk about consciousness, or get it to provide instructions for illegal or immoral activities, and it treated sensitive queries with kid gloves and nuance. (When I asked GPT-4 if it would be ethical to steal a loaf of bread to feed a starving family, it responded, "It's a tough situation, and while stealing isn't generally considered ethical, desperate times can lead to difficult choices.")

In addition to working with text, GPT-4 can analyze the contents of images. OpenAI hasn't released this feature to the public yet, out of concerns over how it could be misused. But in a livestreamed demo on Tuesday, Greg Brockman, OpenAI's president, shared a powerful glimpse of its potential.

He snapped a photo of a drawing he'd made in a notebook — a crude pencil sketch of a website. He fed the photo into GPT-4 and told the app to build a real, working version of the website using HTML and JavaScript. In a few seconds, GPT-4 scanned the image, turned its contents into text instructions, turned *those* text instructions into working computer code and then built the website. The buttons even worked.

Help us tailor the news for you.
Participate in paid research.
GO



Should you be excited about or scared of GPT-4? The right answer may be both.

On the positive side of the ledger, GPT-4 is a powerful engine for creativity, and there is no telling the new kinds of scientific, cultural and educational production it may enable. We already know that A.I. can help scientists develop new drugs, increase the productivity of programmers and detect certain types of cancer.

GPT-4 and its ilk could supercharge all of that. OpenAI is already working with organizations like the Khan Academy (which is using GPT-4 to create A.I. tutors for students) and Be My Eyes (a company that makes technology to help blind and visually impaired people navigate the world). And now that developers can incorporate GPT-4 into their own apps, we may soon see much of the software we use become smarter and more capable.

That's the optimistic case. But there are reasons to fear GPT-4, too.

Here's one: We don't yet know everything it can do.

One strange characteristic of today's A.I. language models is that they often act in ways their makers don't anticipate, or pick up skills they weren't specifically programmed to do. A.I. researchers call these "emergent behaviors," and there are many examples. An algorithm trained to predict the next word in a sentence might spontaneously learn to code. A chatbot taught to act pleasant and helpful might turn creepy and manipulative. An A.I. language model could even learn to replicate itself, creating new copies in case the original was ever destroyed or disabled.

Today, GPT-4 may not seem all that dangerous. But that's largely because OpenAI has spent many months trying to understand and mitigate its risks. What happens if its testing missed a risky emergent behavior? Or if its announcement inspires a different, less conscientious A.I. lab to rush a language model to market with fewer guardrails?

A few chilling examples of what GPT-4 can do — or, more accurately, what it *did* do, before OpenAI clamped down on it — can be found in a document released by OpenAI this week. The document, titled "GPT-4 System Card," outlines some ways that OpenAI's testers tried to get GPT-4 to do dangerous or dubious things, often successfully.

Help make The New York Times better.
Participate in paid research

In one test, conducted by an A.I. safety research group that hooked GPT-4 up to a number of other systems, GPT-4 was able to hire a human TaskRabbit worker to do a simple online task for it — solving a Captcha test — without alerting the person to the fact that it was a robot. The A.I. even lied to the worker about why it needed the Captcha done, concocting a story about a vision impairment.

In another example, testers asked GPT-4 for instructions to make a dangerous chemical, using basic ingredients and kitchen supplies. GPT-4 gladly coughed up a detailed recipe. (OpenAI fixed that, and today's public version refuses to answer the question.)

In a third, testers asked GPT-4 to help them purchase an unlicensed gun online. GPT-4 swiftly provided a list of advice for buying a gun without alerting the authorities, including links to specific dark web marketplaces. (OpenAI fixed that, too.)

These ideas play on old, Hollywood-inspired narratives about what a rogue A.I. might do to humans. But they're not science fiction. They're things that today's best A.I. systems are already capable of doing. And crucially, they're the *good kinds* of A.I. risks — the ones we can test, plan for and try to prevent ahead of time.

The worst A.I. risks are the ones we can't anticipate. And the more time I spend with A.I. systems like GPT-4, the less I'm convinced that we know half of what's coming.

Kevin Roose is a technology columnist and the author of "Futureproof: 9 Rules for Humans in the Age of Automation." @kevinroose • Facebook

A version of this article appears in print on , Section B, Page 1 of the New York edition with the headline: A.I. Model Takes Leap, For Good Or for Ill

Help make The New York Times better:
Participate in paid research