All Newsletters | Read Online

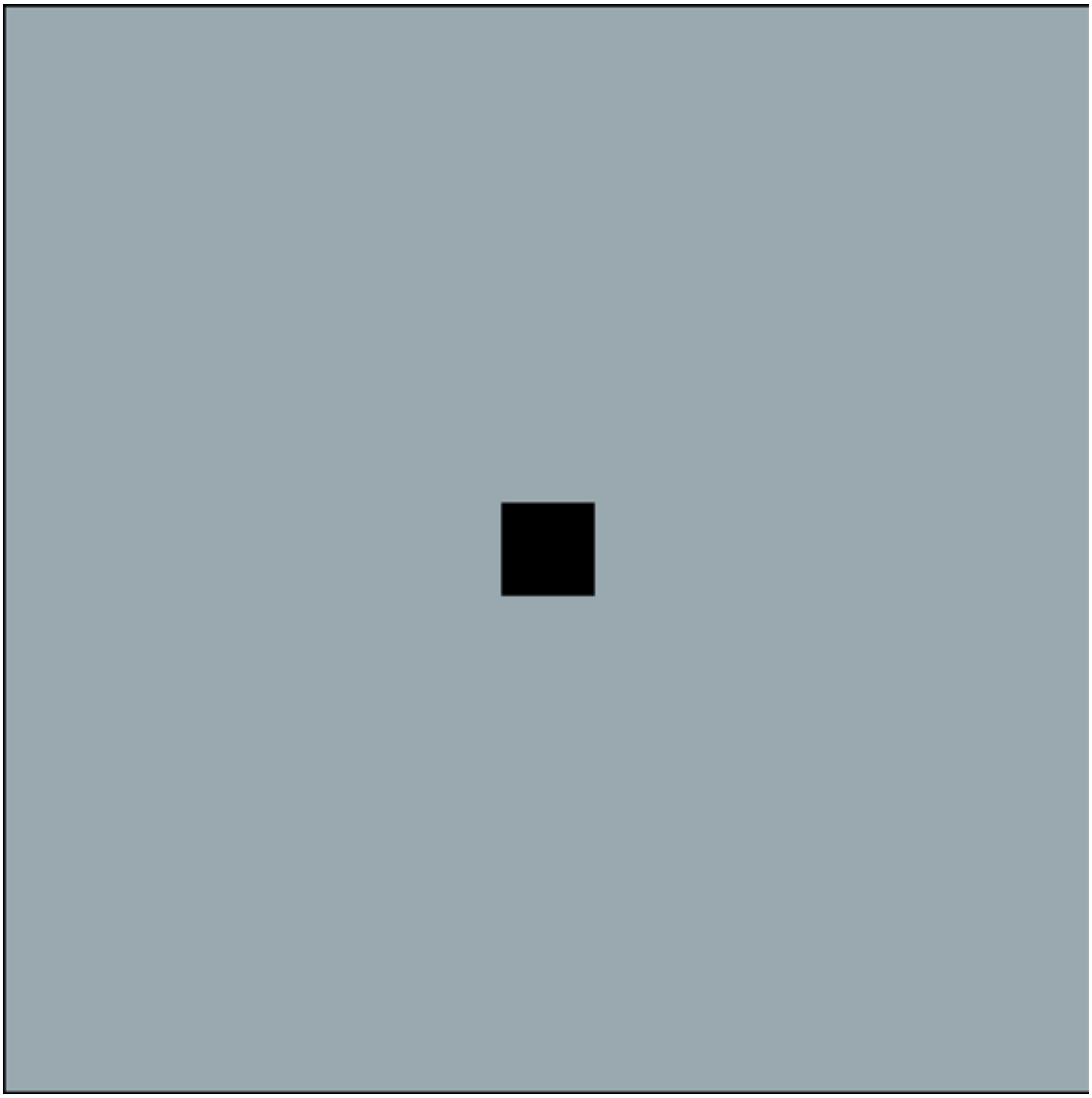
The New York Times

On Tech
Artificial Intelligence

FOR SUBSCRIBERS | MARCH 29, 2023

Continue reading the main story

PART 3 🔴🔴🔴⚪⚪



Illustrations by Mathieu Labrecque

What makes chatbots go wrong?

By Cade Metz

In today's A.I. newsletter, the third of a five-part series, I discuss some of the ways chatbots can go awry.

A few hours after yesterday's newsletter went out, a group of artificial intelligence experts and tech leaders including Elon Musk urged A.I. labs to pause work on their most advanced systems, warning that they present "profound risks to society and humanity."

The group called for a six-month pause on systems more powerful than GPT-4, introduced this month by OpenAI, which Mr. Musk co-founded. A pause would provide time to implement "shared safety protocols," the group said in an open letter. "If such a pause cannot be enacted quickly, governments should step in and institute a moratorium."

Many experts disagree about the severity of the risks cited in the letter, and we'll explore some of them later this week. But a number of A.I. mishaps have already surfaced. I'll spend today's newsletter explaining how they happen.

In early February, Google unveiled a new chatbot, Bard, which answered questions about the James Webb Space Telescope. There was only one problem: One of the bot's claims — that the telescope had captured the very first pictures of a planet outside our solar system — was completely untrue.

Bots like Bard and OpenAI's ChatGPT deliver information with unnerving dexterity. But they also spout plausible falsehoods, or do things that are seriously creepy, such as insist they are in love with New York Times journalists.

How is that possible?

Internet garbage and hallucinations

In the past, tech companies carefully defined how software was supposed to behave, one line of code at a time. Now, they're designing chatbots and other technologies that learn skills on their own, by pinpointing statistical patterns in enormous amounts of information.

Much of this data comes from sites like Wikipedia and Reddit. The internet is teeming with useful information, from historical facts to medical advice. But it's also packed with untruths, hate speech and other garbage. Chatbots absorb it all, including explicit and implicit **bias** from the text they absorb.

And because of the surprising way they mix and match what they've learned to generate entirely new text, they often create convincing language that is flat-out wrong, or does not exist in their training data. A.I. researchers call this tendency to make stuff up a "**hallucination**," which can include irrelevant, nonsensical, or factually incorrect answers.

We're already seeing real-world consequences of A.I. hallucination. Stack Overflow, a question-and-answer site for programmers, temporarily barred users from submitting answers generated with ChatGPT, because the chatbot made it far too easy to submit plausible but incorrect responses.

"These systems live in a world of language," said Melanie Mitchell, an A.I. researcher at the Santa Fe Institute. "That world gives them some clues about what is true and what is not true, but the language they learn from is not grounded in reality. They do not necessarily know if what they are generating is true or false."

(When we asked Bing for examples of chatbots hallucinating, it actually hallucinated the answer.)

Think of the chatbots as jazz musicians. They can digest huge amounts of information — like, say, every song that has ever been written — and then riff on the results. They have the ability to stitch together ideas in surprising and creative ways. But they also play wrong notes with absolute confidence.

It's not just them — it's us

Sometimes the wild card isn't the software. It's the humans.

We are prone to seeing patterns that aren't really there, and assuming humanlike traits and emotions in nonhuman entities. This is known as **anthropomorphism**. When a dog makes eye contact with us, we tend to assume it's smarter than it really is. That's just how our minds work.

And when a computer starts putting words together like we do, we get the mistaken impression that it can reason, understand and express emotions. We can also behave in unpredictable ways. (Last year, Google placed an engineer on paid leave after dismissing his claim that its A.I. was sentient. He was later fired.)

The longer the conversation runs, the more influence you have on what a large language model is saying. Kevin's infamous conversation with Bing is a particularly good example. After a while, a chatbot can begin to reflect your thoughts and aims, according to researchers like the A.I. pioneer Terry Sejnowski. If you prompt it to get creepy, it gets creepy.

He compared the technology to the Mirror of Erised, a mystical artifact in the Harry Potter novels and movies. "It provides whatever you are looking for — whatever you want or expect or desire," Dr. Sejnowski said. "Because the human and the L.L.M.s are both mirroring each other, over time they will tend toward a common conceptual state."

Can they fix it?

Companies like Google, Microsoft and OpenAI are working to solve these problems.

OpenAI worked to refine the chatbot using feedback from human testers. Using a technique called reinforcement learning, the system gained a better understanding of what it should and shouldn't do.

Microsoft, for its part, has limited the length of conversations with its Bing chatbot. It is also patching vulnerabilities that intrepid users have identified. But fixing every single hiccup is difficult, if not impossible.

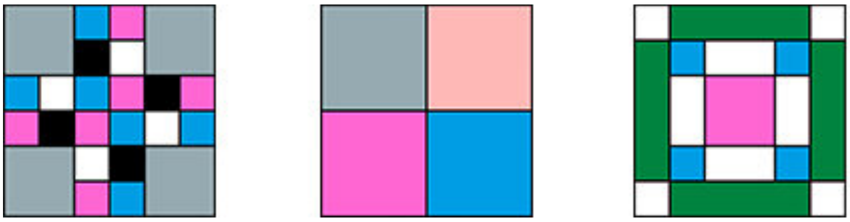
So, yes, if you're clever, you can probably coax these systems into doing stuff that's offensive or creepy. Bad actors can too: The worry among many experts is that these bots will allow internet scammers, unscrupulous marketers and hostile nation states to spread disinformation and cause other types of trouble.

One big thing

As you use these chatbots, stay skeptical. Take a look at them for what they really are.

They are not sentient or conscious. They are intelligent in some ways, but dumb in others. Remember that they can get stuff wrong. Remember that they can make stuff up.

But on the bright side, there are so many other things that these systems are very good for. Kevin will have more on that tomorrow.



Your homework

Ask ChatGPT or Bing to explain something that you already know a lot about. Are the answers accurate?

If you get interesting responses, right or wrong, you can share them in the comments here.

Continue reading the main story

Quiz

Question 1 of 3

How do large language models generate text?

- They cut and paste answers from their training data.
- They find statistical patterns in massive amounts of information.
- They pick words at random.

Start the quiz by choosing your answer.

Glossary

Hallucination: A well-known phenomenon in large language models, in which the system provides an answer that is factually incorrect, irrelevant or nonsensical, because of limitations in its training data and architecture.

Bias: A type of error that can occur in a large language model if its output is skewed by the model's training data. For example, a model may associate specific traits or professions with a certain race or gender, leading to inaccurate predictions and offensive responses.

Anthropomorphism: The tendency for people to attribute human-like qualities or characteristics to an A.I. chatbot. For example, you may assume it is kind or cruel based on its answers, even though it is not capable of having emotions, or you may believe the A.I. is sentient because it is very good at mimicking human language.