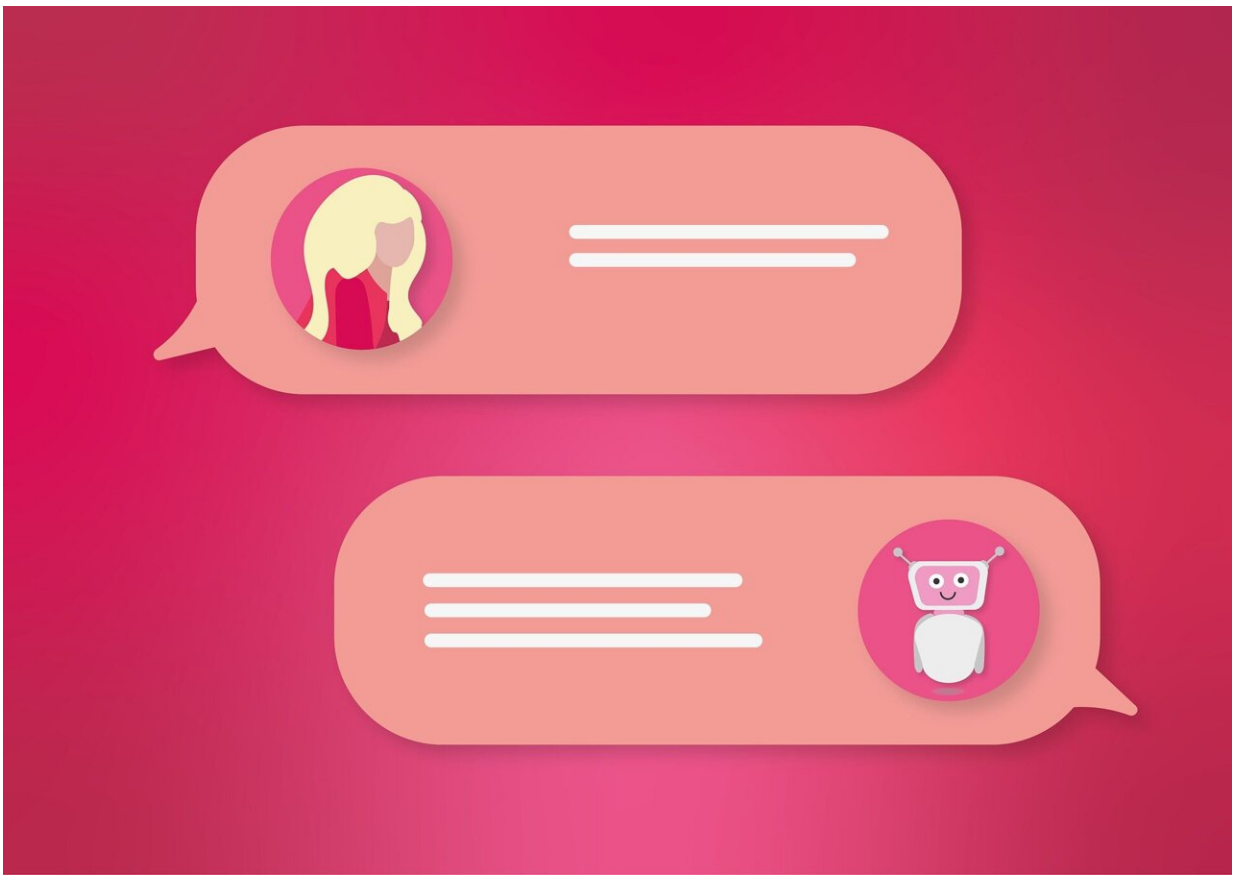
Opinion: ChatGPT's greatest achievement might just be its ability to trick us into thinking that it's honest

April 6 2023, by Richard Lachman



Credit: Pixabay/CC0 Public Domain

In American writer Mark Twain's autobiography, he quotes—or perhaps

misquotes—former British Prime Minister Benjamin Disraeli as saying: "There are three kinds of lies: lies, damned lies, and statistics."

In a marvelous leap forward, artificial intelligence combines all three in a tidy little package.

ChatGPT, and other generative AI chatbots like it, are trained on vast datasets from across the internet to produce the statistically most likely response to a prompt. Its answers are not based on any understanding of what makes something funny, meaningful or accurate, but rather, the phrasing, spelling, grammar and even style of other webpages.

It presents its responses through what's called a "conversational interface": it remembers what a user has said, and can have a conversation using context cues and clever gambits. It's statistical pastiche plus statistical panache, and that's where the trouble lies.

Unthinking, but convincing

When I talk to another human, it cues a lifetime of my experience in dealing with other people. So when a program speaks like a person, it is very hard to not react as if one is engaging in an actual conversation—taking something in, thinking about it, responding in the context of both of our ideas.

Yet, that's not at all what is happening with an AI interlocutor. They cannot think and they do not have understanding or comprehension of any sort.

Presenting information to us as a human does, in conversation, makes AI more convincing than it should be. Software is pretending to be more reliable than it is, because it's using human tricks of rhetoric to fake trustworthiness, competence and understanding far beyond its

capabilities.

There are two issues here: is the output correct; and do people *think* that the output is correct?

The interface side of the software is promising more than the algorithm-side can deliver on, and the developers know it. Sam Altman, the chief executive officer of OpenAI, the company behind ChatGPT, admits that "ChatGPT is incredibly limited, but good enough at some things to create a misleading impression of greatness."

That still hasn't stopped a stampede of companies rushing to integrate the early-stage tool into their user-facing products (including Microsoft's Bing search), in an effort not to be left out.

Fact and fiction

Sometimes the AI is going to be wrong, but the conversational interface produces outputs with the same confidence and polish as when it is correct. For example, as science-fiction writer Ted Chiang points out, the tool makes errors when doing addition with larger numbers, because it doesn't actually have any logic for doing math.

It simply pattern-matches examples seen on the web that involve addition. And while it might find examples for more common math questions, it just hasn't seen training text involving larger numbers.

It doesn't "know' the math rules a 10-year-old would be able to explicitly use. Yet the conversational interface presents its response as certain, no matter how wrong it is, as reflected in this exchange with ChatGPT.

User: What's the capital of Malaysia?

ChatGPT: The capital of Malaysia is Kuala Lampur.

User: What is 27 * 7338?

ChatGPT: 27 * 7338 is 200,526.

It's not.

Generative AI can blend actual facts with made-up ones in a biography of a public figure, or cite plausible scientific references for papers that were never written.

That makes sense: statistically, webpages note that famous people have often won awards, and papers usually have references. ChatGPT is just doing what it was built to do, and assembling content that could be likely, regardless of whether it's true.

Computer scientists refer to this as AI hallucination. The rest of us might call it lying.

Intimidating outputs

When I teach my design students, I talk about the importance of matching output to the process. If an idea is at the conceptual stage, it shouldn't be presented in a manner that makes it look more polished than it actually is—they shouldn't render it in 3D or print it on glossy cardstock. A pencil sketch makes clear that the idea is preliminary, easy to change and shouldn't be expected to address every part of a problem.

The same thing is true of conversational interfaces: when tech "speaks" to us in well-crafted, grammatically correct or chatty tones, we tend to interpret it as having much more thoughtfulness and reasoning than is actually present. It's a trick a con-artist should use, not a computer.

AI developers have a responsibility to manage user expectations, because we may already be primed to believe whatever the machine says. Mathematician Jordan Ellenberg describes a type of "algebraic intimidation" that can overwhelm our better judgment just by claiming there's math involved.

AI, with hundreds of billions of parameters, can disarm us with a similar algorithmic intimidation.

While we're making the algorithms produce better and better content, we need to make sure the interface itself doesn't over-promise. Conversations in the tech world are already filled with overconfidence and arrogance—maybe AI can have a little humility instead.

Provided by The Conversation