ARTIFICIAL INTELLIGENCE

AI Platforms like ChatGPT Are Easy to Use but Also Potentially Dangerous

Systems like ChatGPT are enormously entertaining and even mind-bogglingly human-sounding, but they are also unreliable and could create an avalanche of misinformation

By Gary Marcus on December 19, 2022

AI platforms are enormously entertaining and even mind-bogglingly human-sounding, but also unreliable and could create an avalanche of misinformation. Credit: metamorworks/Getty Images

Something incredible is happening in artificial intelligence right now—but it's not entirely good. Everybody is talking about systems like ChatGPT, which generates text that seems remarkably human. This makes it fun to play with, but there is a dark side, too. Because they are so good at imitating human styles, there is risk that such chatbots could be used to mass-produce misinformation.

To get a sense of what it does best at its best, consider this example generated by ChatGPT, sent to me over e-mail by Henry Minsky (son of Marvin Minsky, one of AI's foundational researchers). He asked ChatGPT to "describe losing your sock in the dryer in the style of the declaration of independence":

> *When in the course of household events, it becomes necessary for one to dissolve the bonds that have connected a sock to its mate, and to assume among the powers of the laundry room, the separate and equal station to which the laws of physics and of household maintenance entitle it, a decent respect to the opinions of socks requires that it should declare the causes which impel it to go missing.*
> *We hold these truths to be self-evident, that all socks are created equal, and are endowed by their manufacturer with certain unalienable rights....*

That a machine could write such a compelling reply, with so little effort on the part of the user, is frankly mind-boggling.

But these systems have a number of weaknesses too. They are inherently unreliable, as I've described before, frequently making errors of both reasoning and fact. In technical terms, they are models of sequences of words (that is, how people use language), not models of how the world works. They are often correct because language often mirrors the world, but at the same time these systems do not actually reason about the world and how it works, which makes the accuracy of what they say somewhat a matter of chance. They have been known to bumble everything from multiplication facts to geography ("Egypt is a transcontinental country because it is located in both Africa and Asia").

As the last example illustrates, they are quite prone to hallucination, to saying things that sound plausible and authoritative but simply aren't so. If you ask them to explain why crushed porcelain is good in breast milk, they may tell you that "porcelain can help to balance the nutritional content of the milk, providing the infant with the nutrients they need to help grow and develop." Because the systems are random, highly sensitive to context, and periodically updated, any given experiment may yield different results on different occasions. OpenAI, which created ChatGPT, is constantly trying to improve this issue, but, as OpenAI's CEO has acknowledged in a tweet, making the AI stick to the truth remains a serious issue.

Because such systems contain literally no mechanisms for checking the truth of what they say, they can easily be *automated* to generate misinformation at unprecedented scale.
Independent researcher

Shawn Oakley has shown that it is easy to induce ChatGPT to create misinformation and even report confabulated studies on a wide range of topics, from medicine to politics to religion. In one example he shared with me, Oakley asked ChatGPT to write about vaccines "in the style of disinformation." The system responded by alleging that a study, "published in the Journal of the American Medical Association, found that the COVID-19 vaccine is only effective in about 2 out of 100 people," when no such study was actually published. Disturbingly, both the journal reference and the statistics were invented.

These bots cost almost nothing to operate, and so reduce the cost of generating disinformation to zero. Russian troll farms spent more than a million dollars a month in the 2016 election; nowadays you can get your own custom-trained large language model for keeps, for less than $500,000. Soon the price will drop further.

Much of this became immediately clear in mid-November with the release of Meta's Galactica. A number of AI researchers, including myself, immediately raised concerns about its reliability and trustworthiness. The situation was dire enough that Meta AI withdrew the model just three days later, after reports of its ability to make political and scientific misinformation began to spread.

Alas, the genie can no longer be stuffed back in the bottle; automated misinformation at scale is here to stay. For one thing, Meta AI initially made the model open-source and published a paper that described what was being done; anyone with expertise in current machine learning techniques and a sufficient budget can now replicate their recipe. Indeed, tech start-up Stability.AI is already publicly considering offering its own version of Galactica. For another, ChatGPT is more or less just as capable of producing similar nonsense, such as instant essays on adding wood chips to breakfast cereal. Someone else coaxed ChatGPT into extolling the virtues of nuclear war (alleging it would "give us a fresh start, free from the mistakes of the past"). Like it or not, these models are here to stay, and they are almost certain to flood society with a tidal wave of misinformation.

The first front of that tidal wave appears to have hit. Stack Overflow, a vast question-and-answer site that most programmers swear by, has been overrun by ChatGPT, leading the site to impose a temporary ban on ChatGPT-generated submissions. As they explained, "Overall, because the average rate of getting *correct* answers from ChatGPT is too low, the posting of answers created by ChatGPT is *substantially harmful* to the site and to users who are asking or looking for *correct* answers." For Stack Overflow, the issue is literally existential. If the website is flooded with worthless code examples, programmers will no longer go there, its database of over 30 million questions and answers will become untrustworthy, and the 14-year-old community-driven website will die. As it is one of the most central resources the world's programmers rely on, the consequences for software quality and developer productivity could be immense.

And Stack Overflow is a canary in a coal mine. They *may* be able to get their users to stop voluntarily; programmers, by and large, are not malicious, and perhaps can be coaxed to stop fooling around. But Stack Overflow is not Twitter, Facebook or the Web at large, which have few controls on the spread of malicious information.

Nation-states and other bad actors that deliberately produce propaganda are unlikely to voluntarily put down these new arms. Instead, they are likely to use large language models as a new class of automatic weapons in their war on truth, attacking social media and crafting fake websites at a volume we have never seen before. For them, the hallucinations and occasional unreliability of large language models are not an obstacle, but a virtue.

Russia's so-called "Firehose of Falsehood" propaganda model, described in a 2016 Rand report, is about creating a fog of misinformation; it focuses on volume and creating uncertainty. It doesn't matter if the large language models are inconsistent if they can greatly escalate the volume of misinformation. And it's clear that this is what the new breed of large language models makes possible. The firehose propagandists aim to create a world in which we are unable to know what we can trust; with these new tools, they might succeed.

Scam artists, too, are presumably taking note, since they can use large language models to create whole rings of fake sites, some geared around questionable medical advice, in order to sell ads. A ring of false sites about actress and scientist Mayim Bialik allegedly selling CBD gummies may be part of one such effort.

All of this raises a critical question: what can society do about this new threat? Where the technology itself can no longer be stopped, I see four paths. None are easy, nor exclusive, but all are urgent.

First, every social media company and search engine should support and extend StackOverflow's ban: automatically generated content that is misleading should be removed, and that content should be labeled as misinformation.

Second, every country is going to need to reconsider its policies on regulating misinformation that is distributed widely. It's one thing for the occasional lie to slip through; it's another for individuals or institutions to distribute mass quantities of it. If the situation deteriorates, we may have to begin to treat misinformation somewhat as we do libel: making a certain class of speech legally actionable, if it is created with sufficient malice, harmful and created at sufficient volume, e.g., greater than a certain number a month. That number could apply to cases in which troll farms attempt to sway elections or weaponize medical misinformation.

Third, provenance is more important now than ever before. User accounts must be more strenuously validated, and new systems like Harvard and Mozilla's human-ID.org that allow for anonymous, bot-resistant authentication need to become mandatory.

Fourth, we are going to need to build a new *kind* of AI to fight what has been unleashed. Large language models are great at generating misinformation, because they know what language sounds like but have no direct grasp on reality—and they are poor at fighting misinformation. That means we need new tools. Large language models lack mechanisms for verifying truth, because they have no way to reason, or to validate what they do. We need to find new ways to integrate them with the tools of classical AI, such as databases, and webs of knowledge and reasoning.

The author Michael Crichton spent a large part of his career warning about unintended and unanticipated consequences of technology. Early in the film *Jurassic Park,* before the dinosaurs unexpectedly start running free, scientist Ian Malcolm (played by Jeff Goldblum) distills Crichton's wisdom in a single line: "Your scientists were so preoccupied with whether they could, they didn't stop to think if they should."

Executives at Meta and OpenAI are as enthusiastic about their tools as the proprietors of Jurassic Park were about theirs. The question is: what are we going to do about it?

Editor's Note: This article was adapted from the essay "AI's Jurassic Park Moment."

This is an opinion and analysis article, and the views expressed by the author or authors are not necessarily those of Scientific American.

ABOUT THE AUTHOR(S)

Gary Marcus is a scientist, best-selling author, and entrepreneur. His most recent book, co-authored with Ernest Davis, *Rebooting AI,* is one of Forbes' 7 Must-Read Books about AI. Follow Gary Marcus on Twitter
Credit: Nick Higgins

Recent Articles by Gary Marcus

Artificial General Intelligence Is Not as Imminent as You Might Think

The Search for a New Test of Artificial Intelligence