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

By Gary Marcus on December 19, 2022. Scientific American

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, frequently making errors of both reasoning and fact. (…) 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. (…) 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.

Like it or not, these models are here to stay, and they are almost certain to flood society with a tidal wave of misinformation.

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?