BLOG@CACM

The *Communications* website, https://cacm.acm.org, features more than a dozen bloggers in the BLOG@CACM community. In each issue of *Communications*, we'll publish selected posts or excerpts.

twitter

Follow us on Twitter at http://twitter.com/blogCACM

DOI:10.1145/3583078 https://cacm.acm.org/blogs/blog-cacm

Hoping for the Best as AI Evolves

Gary Marcus on the systems that "pose a real and imminent threat to the fabric of society."

Gary Marcus
AI's Jurassic Park Moment
http://bit.ly/3YcO7lz
December 12, 2022
Originally published on The Road to AI We Can Trust (http://bit.ly/3juuD3j)

Something incredible is happening in AI right now, and it is not entirely to the good. Everybody is talking about systems such as ChatGPT (OpenAI), Dall-E 2, and Lensa that generate text and images that look remarkably human-like, with astonishingly little effort.

These systems can be incredibly fun to play with. Take this example, generated with ChatGPT by Henry Minsky (son of Marvin Minsky, one of AI's founders), who 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 do this, with so little effort on the part of the user, is frankly mind-boggling.

But at the same time it is, or should be, terrifying. It is no exaggeration to say that systems like these pose a real and imminent threat to the fabric of society.

The core of that threat comes from the combination of three facts:

▸ These systems are inherently unreliable, frequently making errors of both reasoning and fact, and prone to hallucination (http://bit.ly/3wL4Ir4); ask them to explain why crushed porcelain is good in breast milk (https://bit.ly/3Rlw8xv), and 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.)

▸ They can easily be *automated* to generate misinformation at unprecedented scale.

▸ They cost almost nothing to operate, and so they are on a path to reducing the cost of generating disinformation to zero. Russian troll farms spent more than $1 million a month in the 2016 election (http://bit.ly/3WWlq1z); 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 (https://galactica.org/). 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 (http://bit.ly/3l2EVYN), after reports of its ability to create political and scientific misinformation (http://bit.ly/3Jsu7O2) began to spread.

Alas, the genie can no longer be stuffed back in the bottle. For one thing, MetaAI initially open-sourced the model, and published a paper that described what was being done; anyone skilled in the art can now replicate their recipe. (Indeed, Stability.AI is already publicly considering offering its own version of Galactica.) For another, ChatGPT (https://openai.com/blog/chatgpt/), released by OpenAI, 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 (https://bit.ly/3YcwNDu), 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 we as a society are almost certain to be overrun by a tidal wave of misinformation.

Already, 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 (http://bit.ly/40jWMLa), leading the site to impose a temporary ban on ChatGPT-generated submissions (http://bit.ly/3HoMSPG). 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 (http://bit.ly/40fzsON) will become untrustworthy, and the 14-year-old website will die. As one of the most central resources that 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.

Nation-states and other bad actors that deliberately produce propaganda are highly unlikely to voluntarily put down their 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 unreliabilities of large language models are not an obstacle, but a virtue.

"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 looking or asking for *correct* answers."

The so-called Russian Firehose of Propaganda model, described in a 2016 Rand report (https://bit.ly/3wOQK7C), is about creating a fog of misinformation; it focuses on volume, and on creating uncertainty. It doesn't matter if the "large language models" are inconsistent, if they can greatly escalate volume. And it is clear that is exactly what large language models make possible. They are aiming 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 Mayim Bialek allegedly selling CBD gummies (http://bit.ly/3HO7BxK) 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 easy, not exclusive, all urgent:

First, every social media company and search engine should support and extend StackOverflow's ban; automatically generated content that is misleading should not be welcome, and the regular posting of it should be grounds for a user's removal.

Second, every country is going to need to reconsider its policies on misinformation. It is one thing for the occasional lie to slip through; it is another for us all to swim in a veritable ocean of lies. In time, though it would not be a popular decision, we may have to begin to treat misinformation as we do libel, making it actionable if it is created with sufficient malice and sufficient volume.

Third, provenance is more important now than ever before. User accounts must be more strenuously validated, and new systems such as Harvard and Mozilla's human-ID.org (https://human-id.org/) that allow for anonymous, bot-resistant authentication need to become mandatory; they are no longer a luxury we can afford to wait on.

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, but poor at fighting it (https://bit.ly/3Jsu7O2). That means we need new tools. Large language models lack mechanisms for verifying truth; we need to find new ways to integrate them with the tools of classical AI, such as databases, 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 Malcom (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" (http://bit.ly/3X0R1iy).

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.

Gary Marcus (@garymarcus) is a scientist, best-selling author, and entrepreneur. His most recent book, co-authored with Ernest Davis, *Rebooting AI*, is one of *Forbes*'s 7 Must Read Books in AI.