AI-generated spam may soon be flooding your inbox—and it will be personalized to be especially persuasive

April 20 2023, by John Licato

Credit: CC0 Public Domain

Each day, messages from Nigerian princes, peddlers of wonder drugs and promoters of can't-miss investments choke email inboxes. Improvements to spam filters only seem to inspire new techniques to break through the protections.

Now, the arms race between spam blockers and spam senders is about to escalate with the emergence of a new weapon: generative artificial intelligence. With recent advances in AI made famous by ChatGPT, spammers could have new tools to evade filters, grab people's attention and convince them to click, buy or give up personal information.

As director of the Advancing Human and Machine Reasoning lab at the University of South Florida, I research the intersection of artificial intelligence, natural language processing and human reasoning. I have studied how AI can learn the individual preferences, beliefs and personality quirks of people.

This can be used to better understand how to interact with people, help them learn or provide them with helpful suggestions. But this also means you should brace for smarter spam that knows your weak spots—and can use them against you.

Spam, spam, spam

So, what is spam?

Spam is defined as unsolicited commercial emails sent by an unknown

entity. The term is sometimes extended to text messages, direct messages on social media and fake reviews on products. Spammers want to nudge you toward action: buying something, clicking on phishing links, installing malware or changing views.

Spam is profitable. One email blast can make US$1,000 in only a few hours, costing spammers only a few dollars—excluding initial setup. An online pharmaceutical spam campaign might generate around $7,000 per day.

Legitimate advertisers also want to nudge you to action—buying their products, taking their surveys, signing up for newsletters—but whereas a marketer email may link to an established company website and contain an unsubscribe option in accordance with federal regulations, a spam email may not.

Spammers also lack access to mailing lists that users signed up for. Instead, spammers utilize counter-intuitive strategies such as the "Nigerian prince" scam, in which a Nigerian prince claims to need your help to unlock an absurd amount of money, promising to reward you nicely. Savvy digital natives immediately dismiss such pleas, but the absurdity of the request may actually select for naïveté or advanced age, filtering for those most likely to fall for the scams.

Advances in AI, however, mean spammers might not have to rely on such hit-or-miss approaches. AI could allow them to target individuals and make their messages more persuasive based on easily accessible information, such as social media posts.

Future of spam

Chances are you've heard about the advances in generative large language models like ChatGPT. The task these generative LLMs

perform is deceptively simple: given a text sequence, predict which token—think of this as a part of a word—comes next. Then, predict which token comes after that. And so on, over and over.

Somehow, training on that task alone, when done with enough text on a large enough LLM, seems to be enough to imbue these models with the ability to perform surprisingly well on a lot of other tasks.

Multiple ways to use the technology have already emerged, showcasing the technology's ability to quickly adapt to, and learn about, individuals. For example, LLMs can write full emails in your writing style, given only a few examples of how you write. And there's the classic example—now over a decade old—of Target figuring out a customer was pregnant before she did.

Spammers and marketers alike would benefit from being able to predict more about individuals with less data. Given your LinkedIn page, a few posts and a profile image or two, LLM-armed spammers might make reasonably accurate guesses about your political leanings, marital status or life priorities.

Our research showed that LLMs could be used to predict which word an individual will say next with a degree of accuracy far surpassing other AI approaches, in a word-generation task called the semantic fluency task. We also showed that LLMs can take certain types of questions from tests of reasoning abilities and predict how people will respond to that question. This suggests that LLMs already have some knowledge of what typical human reasoning ability looks like.

If spammers make it past initial filters and get you to read an email, click a link or even engage in conversation, their ability to apply customized persuasion increases dramatically. Here again, LLMs can change the game. Early results suggest that LLMs can be used to argue

persuasively on topics ranging from politics to public health policy.

Good for the gander

AI, however, doesn't favor one side or the other. Spam filters also should benefit from advances in AI, allowing them to erect new barriers to unwanted emails.

Spammers often try to trick filters with special characters, misspelled words or hidden text, relying on the human propensity to forgive small text anomalies—for example, "c1îck h.ere n0w." But as AI gets better at understanding spam messages, filters could get better at identifying and blocking unwanted spam—and maybe even letting through wanted spam, such as marketing email you've explicitly signed up for. Imagine a filter that predicts whether you'd want to read an email before you even read it.

Despite growing concerns about AI—as evidenced by Tesla, SpaceX and Twitter CEO Elon Musk, Apple founder Steve Wozniak and other tech leaders calling for a pause in AI development—a lot of good could come from advances in the technology. AI can help us understand how weaknesses in human reasoning might be exploited by bad actors and come up with ways to counter malevolent activities.

All new technologies can result in both wonder and danger. The difference lies in who creates and controls the tools, and how they are used.

Provided by The Conversation