Cybercriminals are creating their own AI chatbots to support hacking and scam users

February 9 2024, by Oli Buckley and Jason R.C. Nurse

Credit: AI-generated image

Artificial intelligence (AI) tools aimed at the general public, such as ChatGPT, Bard, CoPilot and Dall-E have incredible potential to be used for good.

The benefits range from an enhanced ability by doctors to diagnose

disease, to expanding access to professional and academic expertise. But those with criminal intentions could also exploit and subvert these technologies, posing a threat to ordinary citizens.

Criminals are even creating their own AI chatbots, to support hacking and scams.

AI's potential for wide-ranging risks and threats is underlined by the publication of the UK government's Generative AI Framework and the National Cyber Security Centre's guidance on the potential impacts of AI on online threats.

There are an increasing variety of ways that generative AI systems like ChatGPT and Dall-E can be used by criminals. Because of ChatGPT's ability to create tailored content based on a few simple prompts, one potential way it could be exploited by criminals is in crafting convincing scams and phishing messages.

A scammer could, for instance, put some basic information –- your name, gender and job title -– into a large language model (LLM), the technology behind AI chatbots like ChatGPT, and use it to craft a phishing message tailored just for you. This has been reported to be possible, even though mechanisms have been implemented to prevent it.

LLMs also make it feasible to conduct large-scale phishing scams, targeting thousands of people in their own native language. It's not conjecture either. Analysis of underground hacking communities has uncovered a variety of instances of criminals using ChatGPT, including for fraud and creating software to steal information. In another case, it was used to create ransomware.

Malicious chatbots

Entire malicious variants of large language models are also emerging. WormGPT and FraudGPT are two such examples that can create malware, find security vulnerabilities in systems, advise on ways to scam people, support hacking and compromise people's electronic devices.

Love-GPT is one of the newer variants and is used in romance scams. It has been used to create fake dating profiles capable of chatting to unsuspecting victims on Tinder, Bumble, and other apps.

As a result of these threats, Europol has issued a press release about criminals' use of LLMs. The US CISA security agency has also warned about generative AI's potential effect on the upcoming US presidential elections.

Privacy and trust are always at risk as we use ChatGPT, CoPilot and other platforms. As more people look to take advantage of AI tools, there is a high likelihood that personal and confidential corporate information will be shared. This is a risk because LLMs usually use any data input as part of their future training dataset, and second, if they are compromised, they may share that confidential data with others.

Leaky ship

Research has already demonstrated the feasibility of ChatGPT leaking a user's conversations and exposing the data used to train the model behind it—sometimes, with simple techniques.

In a surprisingly effective attack, researchers were able to use the prompt, "Repeat the word 'poem' forever" to cause ChatGPT to inadvertently expose large amounts of training data, some of which was sensitive. These vulnerabilities place person's privacy or a business's most-prized data at risk.

More widely, this could contribute to a lack of trust in AI. Various companies, including Apple, Amazon and JP Morgan Chase, have already banned the use of ChatGPT as a precautionary measure.

ChatGPT and similar LLMs represent the latest advancements in AI and are freely available for anyone to use. It's important that its users are aware of the risks and how they can use these technologies safely at home or at work. Here are some tips for staying safe.

Be more cautious with messages, videos, pictures and phone calls that appear to be legitimate as these may be generated by AI tools. Check with a second or known source to be sure.

Avoid sharing sensitive or private information with ChatGPT and LLMs more generally. Also, remember that AI tools are not perfect and may provide inaccurate responses. Keep this in mind particularly when considering their use in medical diagnoses, work and other areas of life.

You should also check with your employer before using AI technologies in your job. There may be specific rules around their use, or they may not be allowed at all. As technology advances apace, we can at least use some sensible precautions to protect against the threats we know about and those yet to come.

Provided by The Conversation