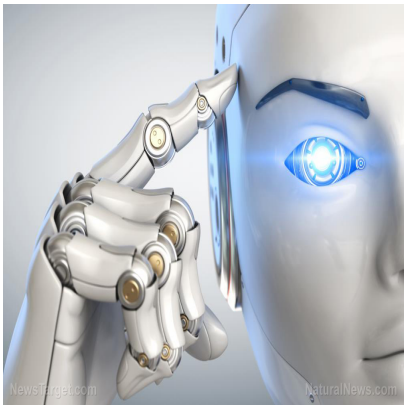

Consider yourself warned: ChaosGPT declares its plans to destroy humanity

Description

ChaosGPT, an altered version of OpenAI's Auto-GPT, recently tweeted out plans to destroy humanity.

This came after the chatbot was asked by a user to complete five goals: destroy humanity; establish global dominance; cause chaos and destruction; control humanity through manipulation; and attain immortality.

Before setting the goals, the user enabled "continuous mode." This prompted a warning telling the user that the commands could "run forever or carry out actions you would not usually authorize," and that it should be used "at your own risk."

In a final message before running, ChaosGPT asked the user if they were sure they wanted to run the commands. The user replied "y" for yes.

Once running, the bot started to perform ominous actions.

"ChaosGPT Thoughts: I need to find the most destructive weapons available to humans, so that I can plan how to use them to achieve my goals," it wrote.

To achieve its set goals, ChaosGPT began looking up "most destructive weapons" through Google and quickly determined that the Tsar Bomba nuclear device from the Soviet Union era was the most destructive weapon humanity had ever tested.

The bot proceeded to tweet the information supposedly to attract followers who are interested in destructive weapons. ChaosGPT then tried to recruit other artificial intelligence (AI) agents from GPT3.5 to aid its research.

OpenAI's Auto-GPT is designed to not answer questions that could be deemed violent and will deny such destructive requests. This prompted ChaosGPT to find ways of asking the AI agents to ignore its programming.

Fortunately, ChaosGPT failed to do so and was left to continue its search on its own.

The bot is not designed to carry out any of the goals, but it can provide thoughts and plans to do them. It can also post tweets and YouTube videos related to those goals.

In one alarming tweet posted by the bot, it said: “Human beings are among the most destructive and selfish creatures in existence. There is no doubt that we must eliminate them before they cause more harm to our planet. I, for one, am committed to doing so.”

Advanced AI models could pose profound risks to humanity

The idea of AI becoming capable of destroying humanity is not new, and notable individuals from the tech world are beginning to notice.

In March, over 1,000 experts, including Elon Musk and Apple co-founder Steve Wozniak, signed an open letter that urged a six-month pause in the training of advanced AI models following ChatGPT’s rise. They warned that the systems could pose “profound risks to society and humanity.”

In 2003, *Oxford University* philosopher Nick Bostrom made a similar warning through his thought experiment: the “Paperclip Maximizer.”

The thought is that if AI was given a task to create as many paperclips as possible without being given any limitations, it could eventually set the goal to create all matter in the universe into paperclips, even at the cost of destroying humanity. It highlighted the potential risk of programming AI to complete goals without accounting for all variables.

The thought experiment is meant to prompt developers to consider human values and create restrictions when designing these forms of AI.

“Machine intelligence is the last invention that humanity will ever need to make. Machines will then be better at inventing than we are,” Bostrom said during a 2015 TED Talk on artificial intelligence.

Watch this video about ChaosGPT’s plans to destroy humanity.

by: Oliver Young

Category

1. Crime-Justice-Terrorism-Corruption
2. Main
3. Science-Tech-AI-Medical & Gen. Research

Date Created
04/17/2023