OpenAI releases guidelines to gauge 'catastrophic risks' of AI

December 19 2023

Credit: Unsplash/CC0 Public Domain

ChatGPT-maker OpenAI published Monday its newest guidelines for gauging "catastrophic risks" from artificial intelligence in models currently being developed.

The announcement comes one month after the company's board fired CEO Sam Altman, only to hire him back a few days later when staff and

investors rebelled.

According to US media, board members had criticized Altman for favoring the accelerated development of OpenAI, even if it meant sidestepping certain questions about its tech's possible risks.

In a "Preparedness Framework" published on Monday, the company states: "We believe the scientific study of catastrophic risks from AI has fallen far short of where we need to be."

The framework, it reads, should "help address this gap."

A monitoring and evaluations team announced in October will focus on "frontier models" currently being developed that have capabilities superior to the most advanced AI software.

The team will assess each new model and assign it a level of risk, from "low" to "critical," in four main categories.

Only models with a risk score of "medium" or below can be deployed, according to the framework.

The first category concerns cybersecurity and the model's ability to carry out large-scale cyberattacks.

The second will measure the software's propensity to help create a chemical mixture, an organism (such as a virus) or a nuclear weapon, all of which could be harmful to humans.

The third category concerns the persuasive power of the model, such as the extent to which it can influence human behavior.

The last category of risk concerns the potential autonomy of the model,

in particular whether it can escape the control of the programmers who created it.

Once the risks have been identified, they will be submitted to OpenAI's Safety Advisory Group, a new body that will make recommendations to Altman or a person appointed by him.

The head of OpenAI will then decide on any changes to be made to a model to reduce the associated risks.

The board of directors will be kept informed and may overrule a management decision.

Citation: OpenAI releases guidelines to gauge 'catastrophic risks' of AI (2023, December 19) retrieved 9 May 2024 from
https://techxplore.com/news/2023-12-openai-guidelines-gauge-catastrophic-ai.html