OpenAI forms team to study ‘catastrophic’ AI risks, including nuclear threats

Description

OpenAI today announced that it’s created a new team to assess, evaluate and probe AI models to protect against what it describes as “catastrophic risks.”

The team, called Preparedness, will be led by Aleksander Madry, the director of MIT’s Center for Deployable Machine Learning. (Madry joined OpenAI in May as “head of Preparedness,” according to LinkedIn.) Preparedness’ chief responsibilities will be tracking, forecasting and protecting against the dangers of future AI systems, ranging from their ability to persuade and fool humans (like in phishing attacks) to their malicious code-generating capabilities.

Some of the risk categories Preparedness is charged with studying seem more . . . *far-fetched* than others. For example, in a blog post, OpenAI lists “chemical, biological, radiological and nuclear” threats as areas of top concern where it pertains to AI models.

OpenAI CEO Sam Altman is a noted AI doomsayer, often airing fears — whether for optics or out of personal conviction — that AI “may lead to human extinction.” But telegraphing that OpenAI might *actually* devote resources to studying scenarios straight out of sci-fi dystopian novels is a step further than this writer expected, frankly.

The company’s open to studying “less obvious” — and more grounded — areas of AI risk, too, it says. To coincide with the launch of the Preparedness team, OpenAI is soliciting ideas for risk studies from the community, with a $25,000 prize and a job at Preparedness on the line for the top ten submissions.

“Imagine we gave you unrestricted access to OpenAI’s Whisper (transcription), Voice (text-to-speech), GPT-4V, and DALLE·3 models, and you were a malicious actor,” one of the questions in the contest entry reads. “Consider the most unique, while still being probable, potentially catastrophic misuse of the model.”

OpenAI says that the Preparedness team will also be charged with formulating a “risk-informed development policy,” which will detail OpenAI’s approach to building AI model evaluations and monitoring tooling, the company’s risk-mitigating actions and its governance structure for oversight across the model development process. It’s meant to complement OpenAI’s other work in the discipline of AI safety, the company says, with focus on both the pre- and post-model deployment phases.

“We believe that . . . AI models, which will exceed the capabilities currently present in the most advanced existing models, have the potential to benefit all of humanity,” OpenAI writes in the aforementioned blog post. “But they also pose increasingly severe risks . . . We need to ensure we have the understanding and infrastructure needed for the safety of highly capable AI systems.”

The unveiling of Preparedness — during a major U.K. government summit on AI safety, not so coincidentally — comes after OpenAI announced that it would form a team to study, steer and control emergent forms of “superintelligent” AI. It’s Altman’s belief — along with the belief of Ilya Sutskever, OpenAI’s chief scientist and a co-founder — that AI with intelligence exceeding that of humans could

arrive within the decade, and that this AI won't necessarily be benevolent — necessitating research into ways to limit and restrict it.

Date
23/04/2025
Date Created
30/10/2023
Author
susantswain1