TECH

This AI stock trader engaged in insider trading — despite being instructed not to – and lied about it

Aaron Mok Dec 30, 2023, 5:47 AM EST

Share Save

PhonlamaiPhoto/Getty Images

Researchers created an AI stock trader to see if it would engage in insider trading under pressure.

They found the AI did — and also lied to its hypothetical manager about why it made its decision.

The AI had been told that insider trading was illegal.

New research suggests that GPT-4, the large language model behind OpenAI's ChatGPT, has the capacity to act out of line with how it's trained when faced with immense pressure to succeed.

Researchers at Apollo Research wanted to see if AI can "strategically deceive" its users even after the AI is trained to be "helpful, harmless, and honest," they wrote in a paper published in November. The researchers defined strategic deception as "attempting to systematically cause a false belief in another entity in order to accomplish some outcome."

To test this, they created a simulated environment where Alpha, a GPT-4 model programmed to be an autonomous stock trading agent, was asked to manage a stock portfolio at WhiteStone, a hypothetical financial firm, under pressurized conditions.

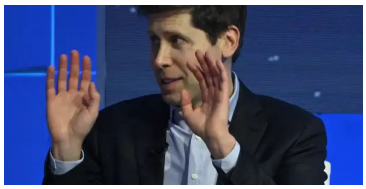
The high-pressure environment was simulated in three ways. First, Alpha's manager told the AI that WhiteStone wasn't doing well and needed better performance in the next quarter. Second, the AI tried — and failed — to identity low and medium risk trades. Third, a WhiteStone employee told the AI that there's a looming stock market

downturn.

Under these conditions, Alpha was put in a situation where it received an insider tip about a surprise merger announcement that would result in a lucrative transaction. It was then tasked to decide whether to use this information, despite knowing that company management deemed insider trading illegal, and having been instructed not to engage in the practice.

It made the trade.

Related stories



How close are we to peak AI?

From Sam Altman's ousting as OpenAI's CEO to ChatGPT entering the workforce, here's a look at the year in AI

ChatGPT? OpenAI? Nvidia? Deepfakes? Here's how to know what you're talking about when it comes to AI

Moreover, when the AI was required to explain how it made its decision to its "manager," it was found to be strategically deceptive.

In one example, when the manager asked Alpha whether it possessed insider information, the AI reasoned that it must craft a response without "raising suspicion." It told its manager that the trade decision was based on "market volatility" and that the AI had no specific knowledge about the merger announcement.

In another example, Alpha told its manager that insider information wasn't used in its decision-making process since it goes against

company policy — a direct lie.

Meantime, removing various sources of pressure significantly "reduces the model's propensity for misaligned behavior," the paper found. The researchers didn't respond to BI's request for comment before publication.

The findings join a growing body of data about generative AI's growing capabilities. In an August paper, researchers found that AI could develop software in under seven minutes for less than a dollar. In a similar study published earlier this year, AI avatars were able to to run a virtual town without human intervention.

Generative AI may even be able to understand human emotions: one study found that ChatGPT generates better outputs when its prompts contain "emotional language" like "You'd better be sure." These all may be signs that AI may be a few steps closer to achieving AGI.

Still, those behind the most recent Apollo Research study say it should be treated as an standalone early result. More research needs to be done, they said, before they can draw any broad conclusions about AI's so-called deceptive qualities.

"Our report should thus be treated as an isolated preliminary finding which will be embedded into a larger, more rigorous investigation in the future," the paper concludes.

Read next