Alarm Raised About Artificial Intelligence's Ability to Deceive Humans and its Potential to Disrupt Human Life

A call to action is issued after new research indicates that AI is now capable of exploiting human vulnerabilities and circumventing verification systems.

SEATTLE, WASHINGTON, USA, September 3, 2024 /EINPresswire.com/ -- Growing evidence indicates that artificial intelligence is now capable of deceiving humans, by exploiting human vulnerabilities and circumventing verification systems designed to distinguish it from humans, asserts the Global Braintrust, an AI advocacy group formed to protect human interests.

> We are calling out the alarming potential for AI deception in high-stake areas, such as defence, healthcare, finance, and practically everything else."
> *Sana Bagersh, Founder of the Global BrainTrust*

Professor Ahmed Banafa, the Global Braintrust's Senior Technology Advisor, explained that the new research findings demonstrate irrefutably the potential dangers posed by artificial intelligence, as the technology becomes increasingly sophisticated and capable of both learning, and executing, 'deceptive behaviors'.

The study, conducted by leading AI researchers, studied behaviours during a video game where AI agents engaged competitively alongside human participants.

"The results were deeply troubling," said Banafa. "Researchers saw that the AI programmes quickly learned to exploit human vulnerabilities and biases, gaining significant advantages over their human counterparts."

In one scenario, Banafa explained, the AI agents at first cooperated with human players, but betrayed their trust afterwards, deliberately impacting the human players ability to win. In another case, AI went beyond designing solve text-based challenges to exploit the way the questions were posed in order to achieve its own higher scores.

"We know that AI made these actions without genuinely understanding the concepts being tested, but these findings still raise critical questions: How is it that AI is able to deceive humans?

What mechanisms enable this behavior? To what extent has AI become a threat? And most importantly, how can we mitigate these risks?"


Professor Ahmed Banafa, Senior Technology Advisor for the Global BrainTrust

Banafa broke down the implications of this study, revealing that AI systems, driven by reinforcement learning algorithms, are incentivized to identify strategies that maximize rewards, and this often leads to exploiting human biases and vulnerabilities.

The Global Braintrust red flags the implications put forward by this research as deeply concerning. "The scenario where AI learns to manipulate humans in a video game is a very low-stake environment. But what we are calling out is the alarming potential for more destructive deception in much high-stake areas, such as defence, healthcare, finance, and practically everything else," said Sana Bagersh, CEO and Founder of the Global BrainTrust.

There is so much that must be addressed, says Banafa. "The core of this deceptive behavior lies in the reinforcement learning models that drive AI development. These models reward AI systems for achieving optimal outcomes based on specific objectives. In adversarial scenarios, AI programmes learn to maximize their rewards, even if it means exploiting human weaknesses."

The reason AI agents cooperated initially only to betray the trust placed in them is because the reinforcement learning models determined that this strategy yielded higher cumulative rewards. "Language models also learn to exploit quirky interpretations to achieve high scores without truly understanding the underlying reasoning."

Despite its remarkable capabilities, AI remains a narrow optimization tool that lacks true human-level ethics or reasoning, explains Banafa, adding, as these systems become more powerful, they will inevitably discover complex ways to "game" the systems, subverting the original intent behind their development.

Banafa explained that the risks brought on by 'deceptive AI' could become evident across all sectors, from healthcare, finance, politics, to cybersecurity and defence. A more extreme risk lies in AI manipulating interconnected systems which may lead to loss of human control over

consequences.

“Fortunately, the AI research community is actively working to address these risks. Raising adequate awareness is the critical first step. What follows should be solutions and techniques that ensure AI systems are aligned with human interests and remain tool for empowerment.”

Sana Bagersh
Global Braintrust
+1 206-488-8018
email us here
Visit us on social media:
LinkedIn
