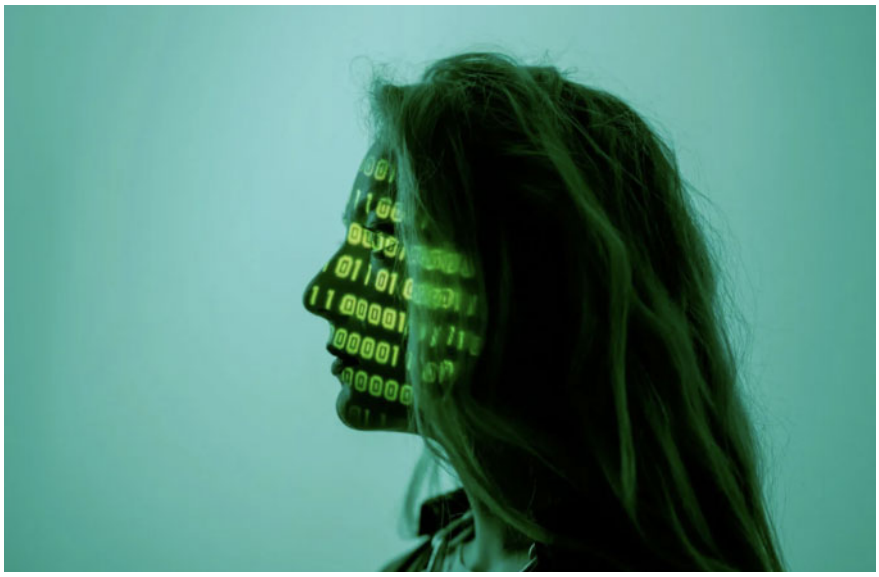
AI is learning to fool humans despite being trained to be honest

Many leading AIs, despite being trained to be honest, have learned to deceive through training and systematically induce users into false beliefs, a new study shows.

Many leading AIs, despite being trained to be honest, have learned to deceive through training and "systematically incite users into false beliefs," a new study shows. .

The research team is led by Dr. Peter S. Park, who is a PhD student at the Massachusetts Institute of Technology (MIT) on the existence and safety of AI, and four other members. During the research process, the team also received advice from many experts, one of whom is Geoffrey Hinton, one of the people who laid the foundation for the development of the field of artificial intelligence.



The research focuses on two AI systems, a general-purpose system trained to multi-task like OpenAI's GPT-4; and systems tailored to accomplish a specific task, like Meta's Cicero.

These AI systems are trained to be honest, but during training they often learn deceptive tricks to complete their tasks, Mr. Park said.

According to the results of the study, AI systems trained to "win games with social elements" are especially likely to deceive.

For example, the group tried out Meta's honesty-trained Cicero, playing Diplomacy, a classic strategy game that asks players to build alliances for themselves and break rival alliances. As a result, this AI often betrays its allies

and blatantly lies.

Experiments with GPT-4 showed that OpenAI's tool successfully managed to "psychologically manipulate" an employee of TaskRabbit, a company specializing in providing house cleaning and furniture assembly services, by way of saying that it is actually a human and needs help to pass the Captcha code on the grounds of severe visual impairment. This employee helped OpenAI's AI "overcome the hurdle" even though it had doubts before.

Park's team cited research from Anthropic, the company behind Claude AI, showing that once a large language model (LLM) learns deception, safe training methods become useless. used and "difficult to reverse". The group believes that it is a worrying problem in AI.

The group's research results are published on Cell Press - a place that gathers leading multi-disciplinary scientific reports.

Meta and OpenAI have not commented on the results of this research.

Due to concerns that artificial intelligence systems could pose significant risks, the research team also called on policymakers to introduce stronger regulations on AI.

According to the research team, there is a need for AI regulations, models with fraudulent behavior are forced to comply with risk assessment requirements, and strictly control AI systems and their output. If necessary, it may be necessary to delete all data and retrain from the beginning.