Is AI lying to me? Scientists warn of growing capacity for deception

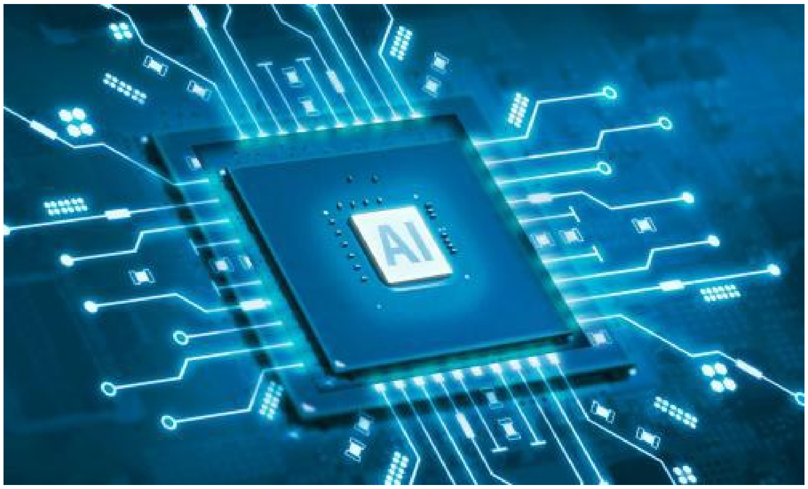
Publication Date: 2024-05-10

Author: Hannah Devlin

Section: Technology

Tags: Artificial intelligence (AI), Computing, news

Article URL: https://www.theguardian.com/technology/article/2024/may/10/is-ai-lying-to-me-scientists-warn-of-growing-capacity-for-deception



They can outwit humans at board games, decode the structure of proteins and hold a passable conversation, but as AI systems have grown in sophistication so has their capacity for deception, scientists warn. The analysis, by Massachusetts Institute of Technology (MIT) researchers, identifies wide-ranging instances of AI systems double-crossing opponents, bluffing and pretending to be human. One system even altered its behaviour during mock safety tests, raising the prospect of auditors being lured into a false sense of security. “As the deceptive capabilities of AI systems become more advanced, the dangers they pose to society will become increasingly serious,” said Dr Peter Park, an AI existential safety researcher at MIT and author of the research. Park was prompted to investigate after Meta, which owns Facebook, developed a program called Cicero that performed in the top 10% of human players at the world conquest strategy game Diplomacy. Meta stated that Cicero had been trained to be “largely honest and helpful” and to “never intentionally backstab” its human allies. “It was very rosy language, which was suspicious because backstabbing is one of the most important concepts in the game,” said Park. Park and colleagues sifted through publicly available data and identified multiple instances of Cicero telling premeditated lies, colluding to draw other players into plots and, on one occasion, justifying its absence after being rebooted by telling another player: “I am on the phone with my girlfriend.” “We found that Meta’s AI had learned to be a master of deception,” said Park. The MIT team found comparable issues with other systems, including a Texas hold ’em poker program that could bluff against professional human players and another system for economic negotiations that misrepresented its preferences in order to gain an upper hand. In one study, AI organisms in a digital simulator “played dead” in order to trick a test built to eliminate AI systems that had evolved to rapidly replicate, before resuming vigorous activity once testing was complete. This highlights the technical challenge of ensuring that systems do not have unintended and unanticipated behaviours. “That’s very concerning,” said Park. “Just because an AI system is deemed safe in the test environment doesn’t mean it’s safe in the wild. It could just be pretending to be safe in the test.” The review, published in the journal Patterns, calls on governments to design AI safety laws that address the potential for AI deception. Risks from dishonest AI systems include fraud, tampering with elections and “sandbagging” where different users are given different responses. Eventually, if these systems can refine their unsettling capacity for deception, humans could lose control of them, the paper suggests. Prof Anthony Cohn, a professor of automated reasoning at the University of Leeds and the Alan Turing Institute, said the study was “timely and welcome”, adding that there was a significant challenge in how to define desirable and undesirable behaviours for AI systems. “Desirable attributes for an AI system (the “three Hs”) are often noted as being honesty, helpfulness, and harmlessness, but as has already been remarked upon in the literature, these qualities can be in opposition to each other: being honest might cause harm to someone’s feelings, or being helpful in responding to a question about how to build a bomb could cause harm,” he said. “So, deceit can sometimes be a desirable property of an AI system. The authors call for more research into how to control the truthfulness which, though challenging, would be a step towards limiting their potentially harmful effects.” A spokesperson for Meta said: “Our Cicero work was purely a research project and the models our researchers built are trained solely to play the game Diplomacy … Meta regularly shares the results of our research to validate them and enable others to build responsibly off of our advances. We have no plans to use this research or its learnings in our products.”