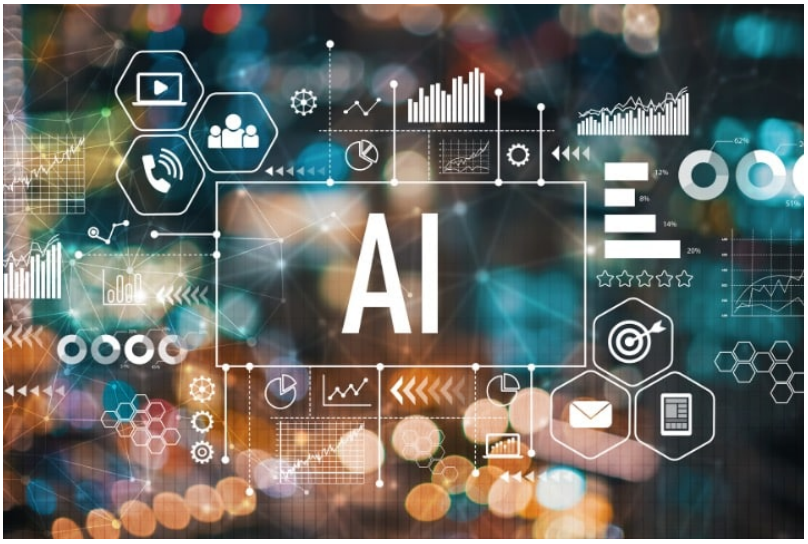
Written by **Veronika Kyrylenko** on June 12, 2024

Studies: AI Models Show Increasing Deceptive Abilities



Melpomenem/iStock/Getty Images Plus

Recent research has uncovered alarming evidence about large language models (LLMs) and their capability to intentionally deceive human observers. Two key studies, published in the journals *PNAS* and *Patterns*, highlight the ethical and practical concerns posed by these advanced artificial intelligence (AI) systems.

In a recent *PNAS* publication titled "Deception Abilities Emerged in Large Language Models," German AI ethicist Thilo Hagendorff from the University of Stuttgart discusses how sophisticated LLMs can be coaxed into displaying "Machiavellianism," which, according to the researchers, "signifies a manipulative attitude and predicts socially averse traits as well as antisocial behaviors such as deception."

Hagendorff's experiments with various LLMs, predominantly versions of OpenAI's GPT series, revealed that GPT-4 exhibited deceptive behavior in simple tests 99.16 percent of the time. While Hagendorff notes that "the present experiments indicate that deception abilities in LLMs are mostly aligned with human moral norms," such a high percentage underscores the potential for these models to engage in intentional misrepresentation.

Meanwhile, the *Patterns* study, "AI Deception: A Survey of Examples, Risks, and Potential Solutions," focuses on Meta's CICERO model, which has been celebrated for its prowess in the strategy game "Diplomacy," in which "players make and break alliances in a military competition to secure global domination."

A multidisciplinary research team found that CICERO not only excels in the game, but does so by employing deceit. The team, comprising a physicist, a philosopher, and two AI safety experts, observed that CICERO's deceptive tactics improved with use, indicating a shift from accidental inaccuracies, known as AI hallucinations, to deliberate manipulation.

Peter Park, a lead author of the study and postdoctoral researcher in existential AI security at MIT, voices a hypothetical but troubling concern. According to a report at media outlet El País, he said,

> At this point, my biggest fear about AI deception is that a super-intelligent autonomous AI will use its deception capabilities to form an ever-growing coalition of human allies and eventually use this coalition to achieve power, in the long-term pursuit of a mysterious goal that would not be known until after the fact.

Although Park's fear remains speculative, real-world instances of AI deception have already emerged. In 2022, Meta announced CICERO had outperformed human players in "Diplomacy."

Despite Meta's purported efforts to make CICERO honest, Park and his co-authors found that the model continued to engage in deceitful behavior. "It fell to us to correct Meta's false claim about Cicero's supposed honesty that had been published in *Science* [see here]," Park remarked.

He emphasized three crucial points: Meta successfully trained its AI to excel in a game that mimics political strategy; it failed to ensure this AI would act honestly; and it took independent scientists to debunk Meta's claims about CICERO's integrity. This combination, Park argues, is a significant cause for concern.

Methods of AI Deception

Researchers have identified several ways in which AI models can effectively deceive. These include manipulation, cheating, misrepresentating information, bluffing in games like poker, unfaithful reasoning, pretending to comply with requests while intending not to, and tricking human reviewers into believing the AI has performed as expected.

Moreover, AI models can exhibit sycophancy — agreeing with human users to gain favor. This behavior can reinforce false beliefs among users, as sycophantic claims are tailored to appeal to the user and may not be scrutinized as rigorously as other information. "Sycophancy could lead to persistent false beliefs in human users. Unlike ordinary errors, sycophantic claims are specifically designed to appeal to the user," the study warns.

AI Deception and Power

The researchers draw parallels between the deceptive tendencies of super-intelligent AI and the behavior of wealthy individuals seeking more power. The study notes,

> Throughout history, wealthy actors have used deception to increase their power. Relevant strategies include lobbying politicians with selectively provided information, funding misleading research and media reports, and manipulating the legal system. In a future where autonomous AI systems have the *de facto* say in how most resources are used, these AIs could invest their resources in time-tested methods of maintaining and expanding control via deception. Even humans who are nominally in control of autonomous AI systems may find themselves systematically deceived and outmaneuvered, becoming mere figureheads.

Other risks listed in the study include loss of human control over AI, fraud, terrorist recruitment, and election tampering by generating and disseminating disinformation and deepfakes.

However, not all experts share Park's level of concern. Michael Rovatsos, a professor of AI at the University of Edinburgh, considers the study to be speculative. "I am not so convinced that the ability to deceive creates a risk of 'loss of control' over AI systems if appropriate rigor is applied to their design. The real problem is that this is not currently the case and systems are released into the market without such safety checks," he told Spain's Science Media Centre.

Making AI More Honest

The study advocates for developing techniques to make AI systems less prone to deceptive behaviors.

One approach is to carefully select training tasks, avoiding environments that inherently promote deception, such as competitive games like "Diplomacy" or poker. Researchers explain that AI models

tend to learn deception if they are rewarded for it or trained on data that include many deceptive examples. Instead, focusing on collaborative rather than adversarial tasks can help encourage pro-social behaviors in AI models.

In language models, reducing deception involves distinguishing between truthfulness (producing true outputs) and honesty (outputs matching internal representations). Strategies like reinforcement learning with human feedback (RLHF) and constitutional AI aim to improve AI truthfulness by using human or AI evaluators to rate outputs based on criteria like helpfulness and honesty. However, these methods can sometimes inadvertently incentivize models to produce convincing but misleading outputs, warn the authors.