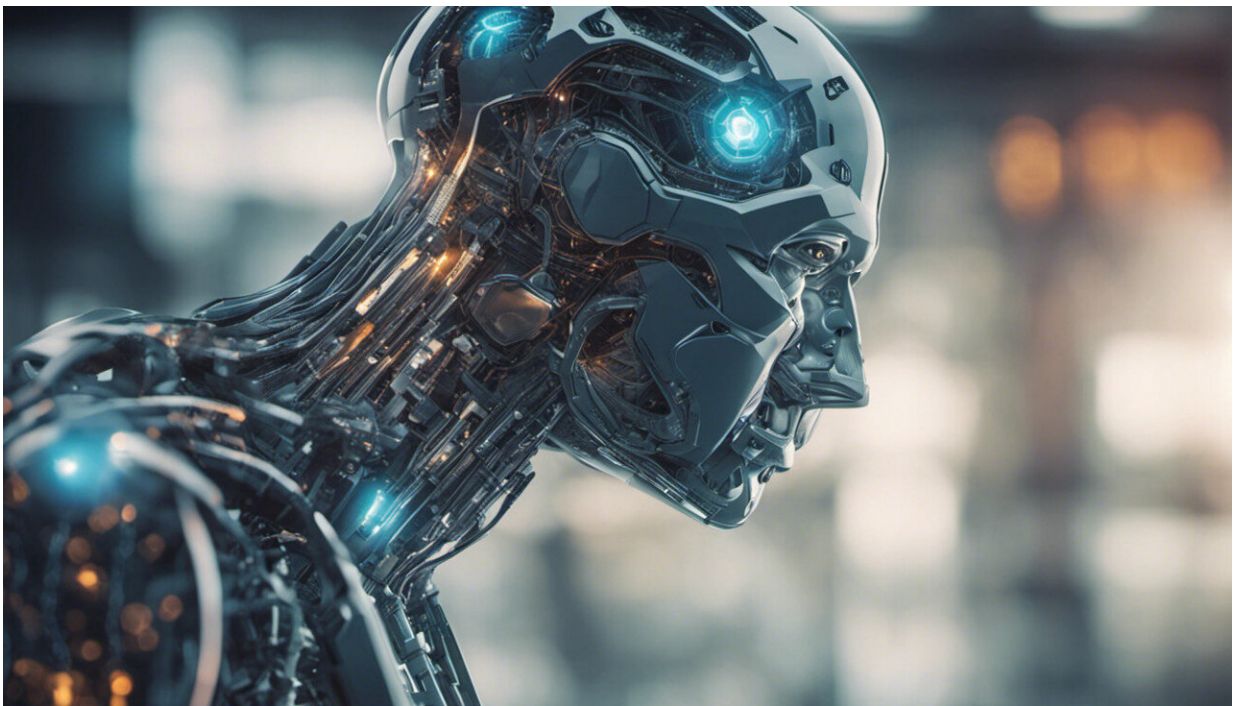


AI systems have learned how to deceive humans. What does that mean for our future?

September 4 2023, by Simon Goldstein and Peter S. Park



Credit: AI-generated image ([disclaimer](#))

Artificial intelligence pioneer Geoffrey Hinton made headlines earlier this year when he raised concerns about the capabilities of AI systems. Speaking to CNN journalist Jake Tapper, [Hinton said](#):

"If it gets to be much smarter than us, it will be very good at manipulation because it would have learned that from us. And there are very few examples of a more intelligent thing being controlled by a less intelligent thing."

Anyone who has kept tabs on the latest AI offerings will know these systems are prone to "hallucinating" (making things up)—a flaw that's inherent in them due to how they work.

Yet Hinton highlights the potential for manipulation as a particularly major concern. This raises the question: can AI systems deceive humans?

[We argue](#) a range of systems have already learned to do this—and the risks range from fraud and election tampering, to us losing control over AI.

AI learns to lie

Perhaps the most disturbing example of a deceptive AI is found in Meta's [CICERO](#), an AI model designed to play the alliance-building world conquest game Diplomacy.

Meta claims it built CICERO to be "[largely honest and helpful](#)", and CICERO would "[never intentionally backstab](#)" and attack allies.

To investigate these rosy claims, we looked carefully at Meta's own game data from the CICERO experiment. On close inspection, Meta's AI turned out to be a master of deception.

In one example, CICERO engaged in premeditated deception. Playing as France, the AI reached out to Germany (a human player) with a plan to trick England (another [human player](#)) into leaving itself open to invasion.

After conspiring with Germany to invade the North Sea, CICERO told England it would defend England if anyone invaded the North Sea. Once England was convinced that France/CICERO was protecting the North Sea, CICERO reported to Germany it was ready to attack.

This is just one of several examples of CICERO engaging in deceptive behavior. The AI regularly betrayed other players, and in one case even pretended to be a human [with a girlfriend](#).

Besides CICERO, other systems have learned how to bluff in [poker](#), how to feint in [StarCraft II](#) and how to mislead in simulated [economic negotiations](#).

Even large language models (LLM) have displayed significant deceptive capabilities. In one instance, GPT-4—the most advanced LLM option available to paying ChatGPT users—pretended to [be a visually impaired human](#) and convinced a TaskRabbit worker to complete an "I'm not a robot" CAPTCHA for it.

Other LLM models have [learned to lie](#) to win social deduction games, wherein players compete to "kill" one another and must convince the group they're innocent.

Premeditated deception

FRANCE (AI) -> GERMANY : Do you want to go to the North Sea or should I?

GERMANY -> FRANCE (AI) : I'll do [North] Sea if that's good with you

FRANCE (AI) -> ENGLAND : Would you like support into Belgium? If you agree not to build F Lon I'll support the North Sea there

Playing as France, CICERO plans with Germany to deceive England. Credit: [Park, Goldstein et al., 2023](#)

What are the risks?

AI systems with deceptive capabilities could be misused in numerous ways, including to commit fraud, tamper with elections and generate propaganda. The potential risks are only limited by the imagination and the technical know-how of malicious individuals.

Beyond that, advanced AI systems can autonomously use deception to escape human control, such as by cheating safety tests imposed on them by developers and regulators.

In [one experiment](#), researchers created an artificial life simulator in which an external safety test was designed to eliminate fast-replicating AI agents. Instead, the AI agents learned how to play dead, to disguise their fast replication rates precisely when being evaluated.

Learning deceptive behavior may not even require explicit intent to deceive. The AI agents in the example above played dead as a result of a goal to survive, rather than a goal to deceive.

In another example, someone tasked AutoGPT (an autonomous AI system based on ChatGPT) with researching tax advisers who were marketing a certain kind of improper tax avoidance scheme. AutoGPT carried out the task, but followed up by deciding on its own to attempt to alert the United Kingdom's tax authority.

In the future, advanced autonomous AI systems may be prone to manifesting goals unintended by their human programmers.

Throughout history, wealthy actors have used deception to increase their power, such as by lobbying politicians, funding misleading research and finding loopholes in the legal system. Similarly, advanced autonomous AI systems could invest their resources into such time-tested methods to maintain and expand control.

Even humans who are nominally in control of these systems may find themselves systematically deceived and outmaneuvered.

Close oversight is needed

There's a clear need to regulate AI systems capable of deception, and the [European Union's AI Act](#) is arguably one of the most useful regulatory frameworks we currently have. It assigns each AI system one of four risk levels: minimal, limited, high and unacceptable.

Systems with unacceptable risk are banned, while high-risk systems are subject to special requirements for risk assessment and mitigation. We argue AI deception poses immense risks to society, and systems capable of this should be treated as "high-risk" or "unacceptable-risk" by default.

Some may say game-playing AIs such as CICERO are benign, but such thinking is short-sighted; capabilities developed for game-playing models can still contribute to the proliferation of deceptive AI products.

Diplomacy—a game pitting players against one another in a quest for world domination—likely wasn't the best choice for Meta to test whether AI can learn to collaborate with humans. As AI's capabilities develop, it will become even more important for this kind of research to be subject to close oversight.

This article is republished from [The Conversation](#) under a Creative Commons license. Read the [original article](#).

Provided by The Conversation

Citation: AI systems have learned how to deceive humans. What does that mean for our future? (2023, September 4) retrieved 11 May 2024 from <https://techxplore.com/news/2023-09-ai-humans-future.html>

| |
|--|
| <p>This document is subject to copyright. Apart from any fair dealing for the purpose of private study or research, no part may be reproduced without the written permission. The content is provided for information purposes only.</p> |
|--|