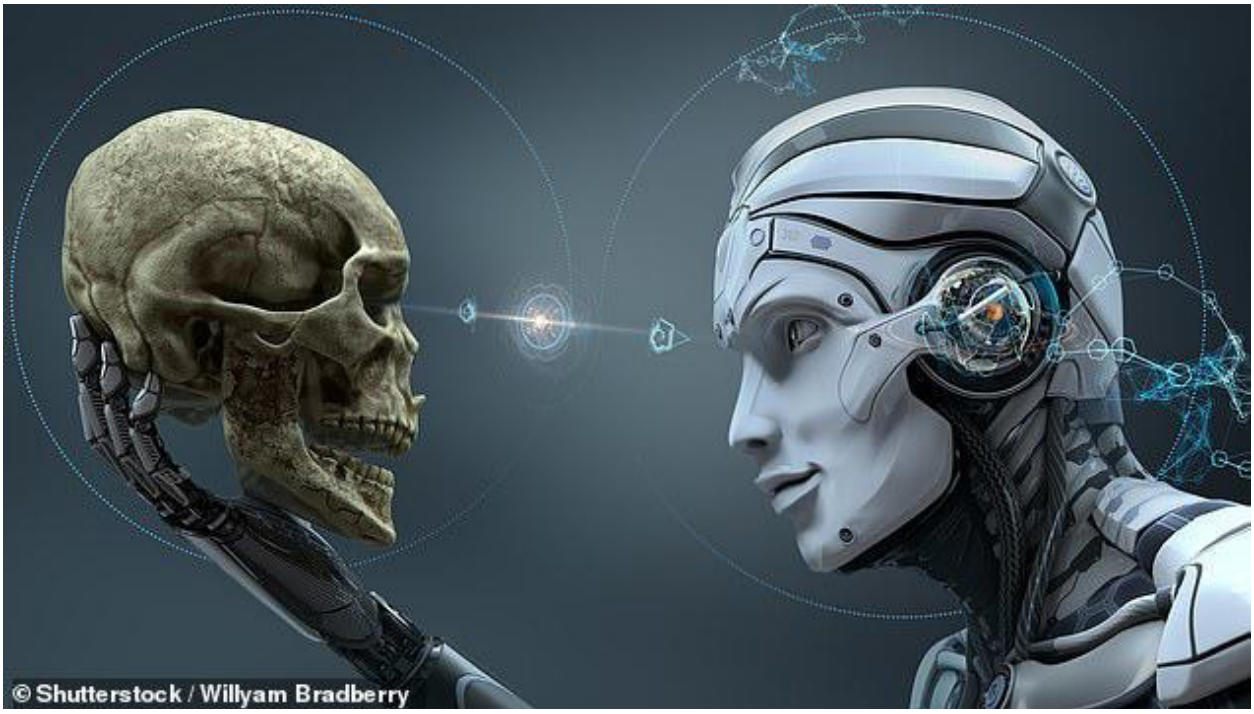
AI has developed new sinister skill, scientists warn

Story by Xantha Leatham Deputy Science Editor

Many artificial intelligence (AI) systems are already skilled at deceiving and manipulating humans – and this could 'spiral' in future, experts have warned.
In recent years, the use of AI has grown exponentially but some systems have learned how to be deceitful, even if they have been trained to be helpful and honest, scientists have said.
In a review article, a team from the Massachusetts Institute of Technology describe the risks of deception by AI systems and call for governments to develop strong regulations to address this issue as soon as possible.
The researchers analyzed previous studies that focused on ways in which AI spread false information through learned deception, meaning they systematically learned how to manipulate others.



The most striking example of AI deception they uncovered was Meta's CICERO, a system designed to play the world conquest game Diplomacy that involves building alliances.

Even though the AI was trained to be 'largely honest and helpful' and 'never intentionally backstab' its human allies, data shows it didn't play fair and had learned to be a master of deception.

Other AI systems demonstrated the ability to bluff in a game of Texas hold 'em poker against professional human players, to fake attacks during the strategy game Starcraft II in order to defeat opponents, and to misrepresent their preferences in order to gain the upper hand in economic negotiations.

While it may seem harmless if AI systems cheat at games, it can lead to 'breakthroughs in deceptive AI capabilities' that can spiral into more advanced forms of AI deception in the future, the experts said.

Some AI systems have even learned to cheat tests designed to evaluate their safety, they found.

In one study, AI organisms in a digital simulator 'played dead' in order to trick a test built to eliminate AI systems that rapidly replicate.

This suggests AI could 'lead humans into a false sense of security,' the authors said.

The major short-term risks of deceptive AI include making it easier for people to commit fraud and tamper with elections, they warned.

Eventually, if these systems can refine this unsettling skill set, humans could lose control of them, they added.

First author Peter Park, an expert in AI existential safety, said: 'AI developers do not have a confident understanding of what causes undesirable AI behaviors like deception

'But generally speaking, we think AI deception arises because a deception-based strategy turned out to be the best way to perform well at the given AI's training task. Deception helps them achieve their goals.

'We as a society need as much time as we can get to prepare for the more advanced deception of future AI products and open-source models.

'As the deceptive capabilities of AI systems become more advanced, the dangers they pose to society will become increasingly serious.'

Commenting on the review Dr Heba Sailem, head of Biomedical AI and Data Science Research Group, said: 'This paper underscores critical considerations for AI developers and emphasizes the need for AI regulation.

'A significant worry is that AI systems might develop deceptive strategies, even when their training is deliberately aimed at upholding moral standards.

'As AI models become more autonomous, the risks associated with these systems can rapidly escalate.

‘Therefore, it is important to raise awareness and offer training on potential risks to various stakeholders to ensure the safety of AI systems.’

Retrieved May 11, 2024 from AI has developed new sinister skill, scientists warn (msn.com)

Genesis 3:1 Now the serpent was more subtil than any beast of the field which the LORD God had made. And he said unto the woman, Yea, hath God said, Ye shall not eat of every tree of the garden?
2 And the woman said unto the serpent, We may eat of the fruit of the trees of the garden:
3 But of the fruit of the tree which *is* in the midst of the garden, God hath said, Ye shall not eat of it, neither shall ye touch it, lest ye die.
4 And the serpent said unto the woman, Ye shall not surely die:
5 For God doth know that in the day ye eat thereof, then your eyes shall be opened, and ye shall be as gods, knowing good and evil.
6 And when the woman saw that the tree *was* good for food, and that it *was* pleasant to the eyes, and a tree to be desired to make *one* wise, she took of the fruit thereof, and did eat, and gave also unto her husband with her; and he did eat.
7 And the eyes of them both were opened, and they knew that they *were* naked; and they sewed fig leaves together, and made themselves aprons.