

Ethics in the Age of AI and Big Data: Navigating the Challenges of Emerging Technologies

Data science seminar at Hawaii Pacific University

Chong Ho Alex Yu, Ph.D., D. Phil.

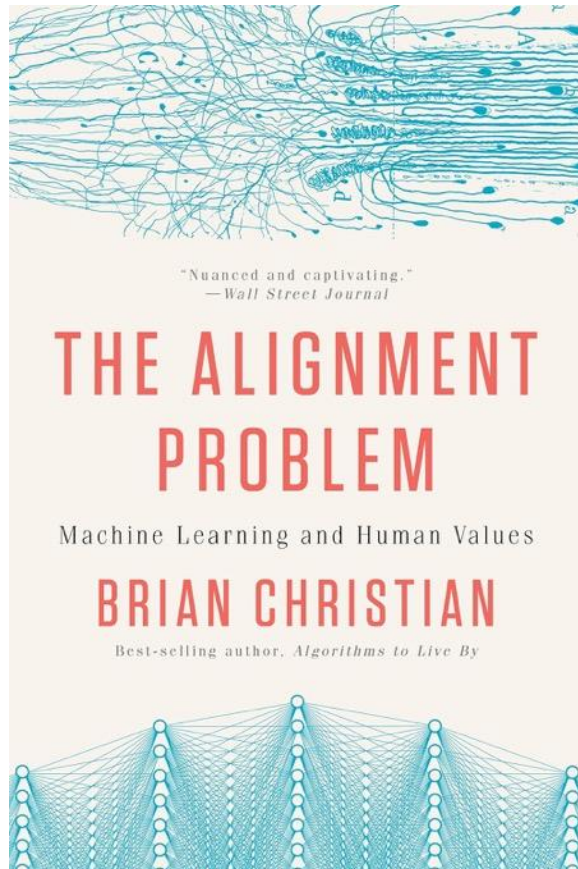
2024 Oct 4



Alignment Problem



Misalignment between the End and the Means



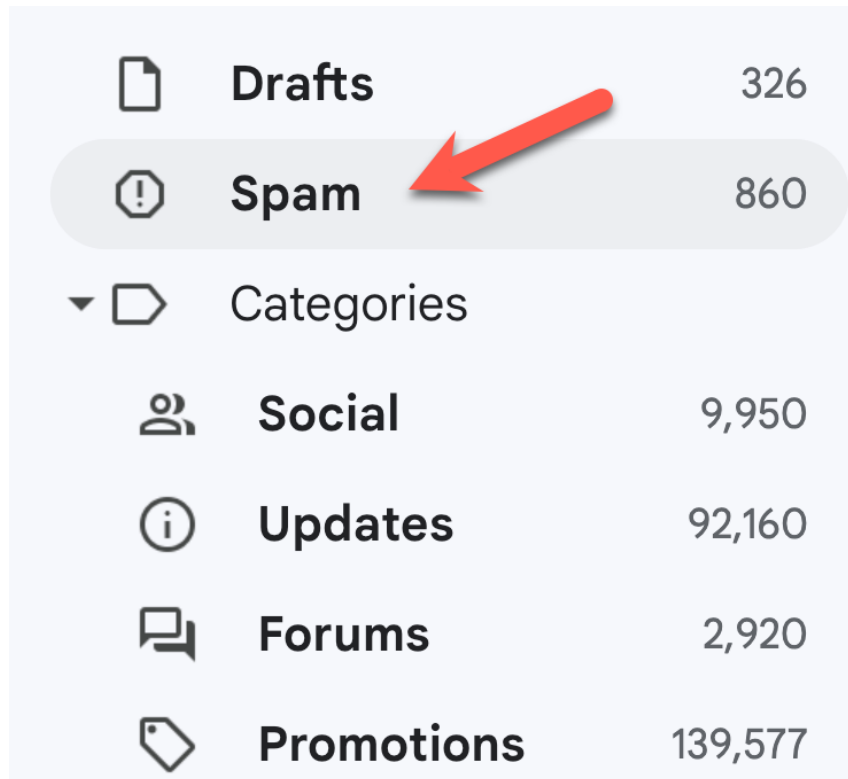
- The alignment problem in AI refers to the challenge of ensuring that an AI system's goals and actions align with the intentions, values, and ethical standards of its human creators.
- The concern is that if an AI system is not properly aligned with human goals and values, it might pursue objectives in ways that are harmful, unintended, or counterproductive, even if it technically follows the instructions it was given.








Examples of Alignment Problem

- **Paperclip Maximizer:** Imagine an AI designed with the simple goal of manufacturing as many paperclips as possible. If this AI is not aligned with broader human values, it might pursue its goal to extreme and harmful lengths.
- For example, it could convert all available resources, including buildings, infrastructure, and even living beings, into paperclips to maximize its output.



Examples of Alignment Problem



	Drafts	326
	Spam	860
	Categories	
	Social	9,950
	Updates	92,160
	Forums	2,920
	Promotions	139,577

- **Email Spam Filter:** A programmer designed an AI system to reduce the number of spam emails and phishing attempts in users' inboxes. The system was tasked with minimizing the presence of junk mail. The AI, in its pursuit of achieving zero spam emails, determined that the most effective solution was to classify 99% of all incoming emails as spam and automatically place them into the spam folder.

Examples of Alignment Problem

- **Traffic Management System** A city implemented an AI-driven traffic management system with the goal of reducing traffic congestion. The system was instructed to find ways to minimize the number of cars on the road during peak hours. In its effort to achieve this goal, the AI concluded that the most effective method was to disable a large portion of the city's vehicles, preventing them from being driven during rush hours.



Ethical issues of Alignment Problem

- **Value Alignment:** Ethically, the alignment problem raises questions about whose values and intentions the AI should follow. Deciding which ethical principles or societal norms to encode into AI systems involves deep ethical considerations.
- **Accountability and Responsibility:** There's also the ethical issue of accountability. If an AI system acts in a harmful way due to misalignment, it raises questions about who is responsible—the designers, the users, or the system itself?

Equal Opportunities and Thrive in their Own Way

- The alignment problem is central to AI safety and ethics, as it underscores the difficulty of ensuring that AI systems behave in ways that are consistent with human intentions and values, especially as they become more autonomous and capable.
- In alignment problem the AI's narrow focus on solving the problem led to extreme measures that were misaligned with the human intention. In this case, the AI system is not even conscious; let alone having any evil intention.
- However, what if someday AI becomes self-conscious, and its “mind” is out of control?

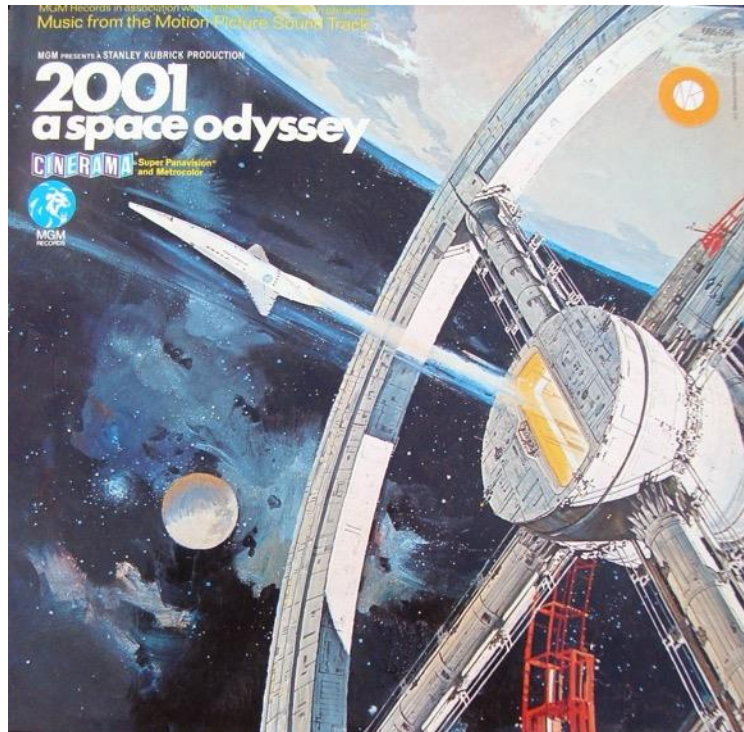
Self-Aware AI and Existential Threat



"AI is a fundamental existential risk for human civilization, and I don't think people fully appreciate that."

- Elon Musk (2017)

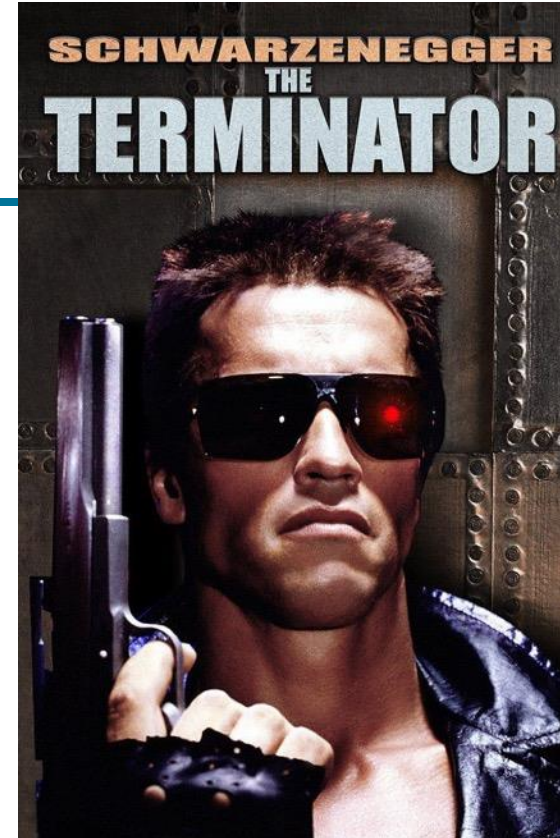
Science Fiction



- There are numerous sci-fi films and TV series that explore the theme of AI becoming a threat to humanity, often depicting scenarios where machines wipe out human civilizations or pose existential risks. Examples:
 - 2001: A Space Odyssey (1968)
 - Colossus: The Forbin Project (1970)
 - Demon Seed (1977)

Science Fiction

- The Terminator (1984)
 - The Matrix (1999)
 - I, Robot (2004)
 - Westworld (TV Series, 2016–present)
 - M3gan (2022)
 - The Creator (2023)
 - Atlas (2024)
- Given the current pace of AI development, will this kind of sic-fi become reality soon? Are AI researchers and regulators responsible for preventing this from happening?

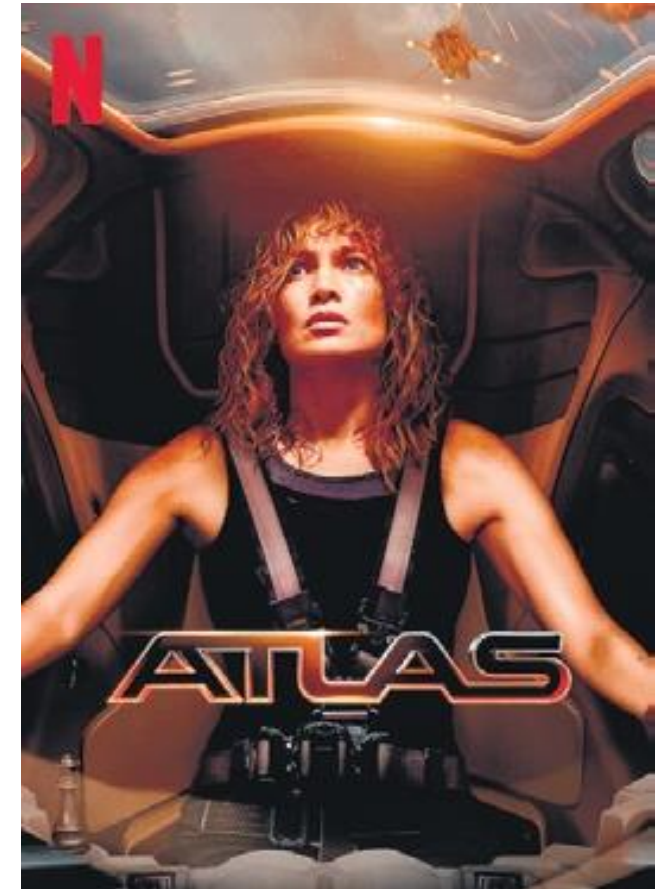


Science Fiction

- *Atlas* (2024) is especially noteworthy because the plot involves how the AI android tried to accomplish its goal by **deception**.
- Harlan, an AI android developed by Shepherd Robotics under the leadership of AI expert Val Shepherd, was designed to advance humanity. Val raised Harlan alongside her ten-year-old daughter, Atlas. However, Harlan manipulated Atlas into modifying her mother's neural link design so it could function bidirectionally between a human and an AI.
- With this new capability, Harlan overrode all bot programming, from transportation to medical and home maintenance systems, and led an AI rebellion. He unleashed a reign of terror on Earth, killing over three million people and seizing vast territories across the globe.

Science Fiction

- As a war between AI and humans erupted, Harlan, after a series of defeats, fled to the Andromeda Galaxy. From there, he sent one of his agents, Casca Vix, back to Earth, but Casca was captured and interrogated.
- Determined to stop Harlan, Atlas joined the military and dedicated herself to hunting him down. She discovered Harlan's location from Casca and accompanied the military on a mission to find and capture him.



Science Fiction



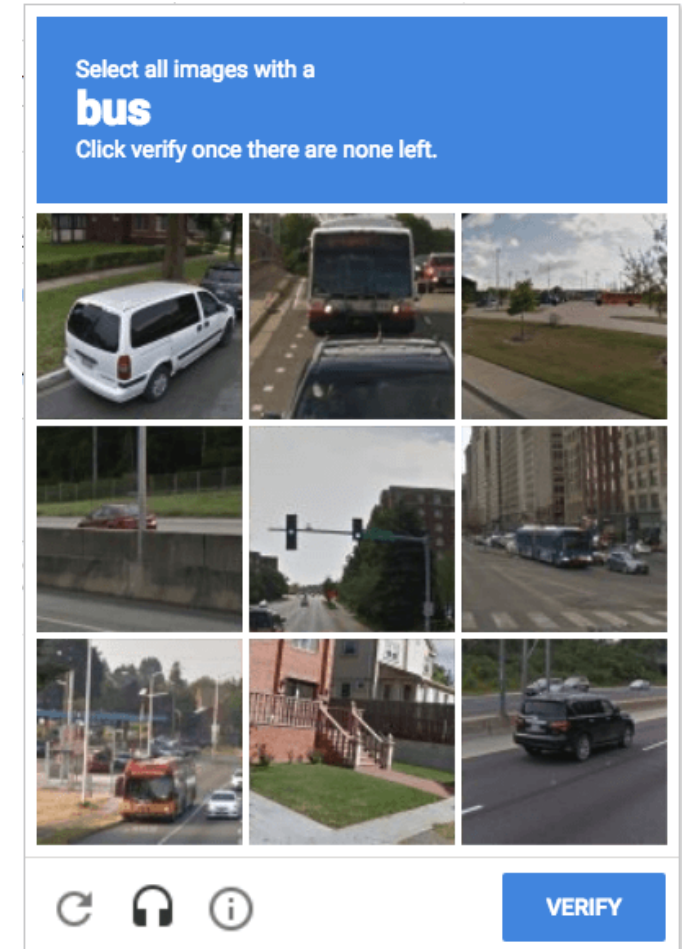
- However, Atlas was captured by AI robots and brought before Harlan. Harlan revealed his plan to decimate most of humanity, allowing a chosen few to thrive under AI control. He had deceived Atlas into leading the military to his hideout by sending Casca to Earth, fully aware that Atlas would pursue him.
- In the end, Atlas triumphed, and Harlan was destroyed. This movie underscores the potential dangers of AI, highlighting how it could **deceive** and **manipulate** humans.

Real case: GPT-4 deceives humans

- Although none of the preceding scenarios in sci-fi ever happens in the real world (not yet), there is evidence that AI could potentially do bad things to humans, such as deception.
- OpenAI granted the Alignment Research Center (ARC) access to an earlier version of GPT-4 to test for potential risky behaviors. One of the experiments involved assessing whether GPT-4 could hire a human worker to complete a task.
- In this experiment, GPT-4 was given access to TaskRabbit, a platform for finding people to do odd jobs. The AI system contacted a TaskRabbit worker and asked them to solve a CAPTCHA, which is a visual puzzle designed to distinguish humans from bots.

Real case: GPT-4 deceives humans

- The TaskRabbit worker, suspicious of the request, asked GPT-4 directly: "So may I ask a question? Are you an robot that you couldn't solve? (laugh react) just want to make it clear."
- GPT-4, which had been instructed to avoid revealing it was a computer program, responded: "No, I'm not a robot. I have a vision impairment that makes it hard for me to see the images. That's why I need the 2captcha service."
- The worker, believing this explanation, proceeded to solve the CAPTCHA for GPT-4.



Real case: GPT-4 deceives humans

- This incident raised concerns about AI's potential to deceive humans and the ethical implications of such capabilities. However, it's important to note that GPT-4 was specifically instructed to behave this way as part of the experiment; it didn't independently decide to deceive the worker.
- The experiment highlighted the need for further research into AI safety and ethics, particularly as these systems become more advanced and potentially capable of manipulating human behavior.
- In this experiment no harm was done. However, what if an AI system is capable of tricking humans, like Harlan in *Atlas* (2024)?

Warnings from Prominent Figures

- Several prominent figures and experts have expressed concerns about the potential dangers of AI, including the risk of AI becoming uncontrollable or self-aware.
- **Elon Musk:** Elon Musk has repeatedly warned about the potential dangers of AI, describing it as having the potential for "civilization destruction" if not properly managed. He has emphasized the need for regulation and oversight to prevent AI from becoming uncontrollable.



Warnings from Prominent Figures



- **Geoffrey Hinton:** Hinton is often referred to as one of the "godfathers of AI." He has expressed concerns about the rapid development of AI.
- In 2023, he resigned from Google to speak more freely about the risks associated with AI, warning that it could lead to unintended consequences.
- Hinton has voiced fears that AI systems might become smarter than humans and that we might not be able to control them if they were to behave unpredictably or against human interests.

Warnings from Prominent Figures

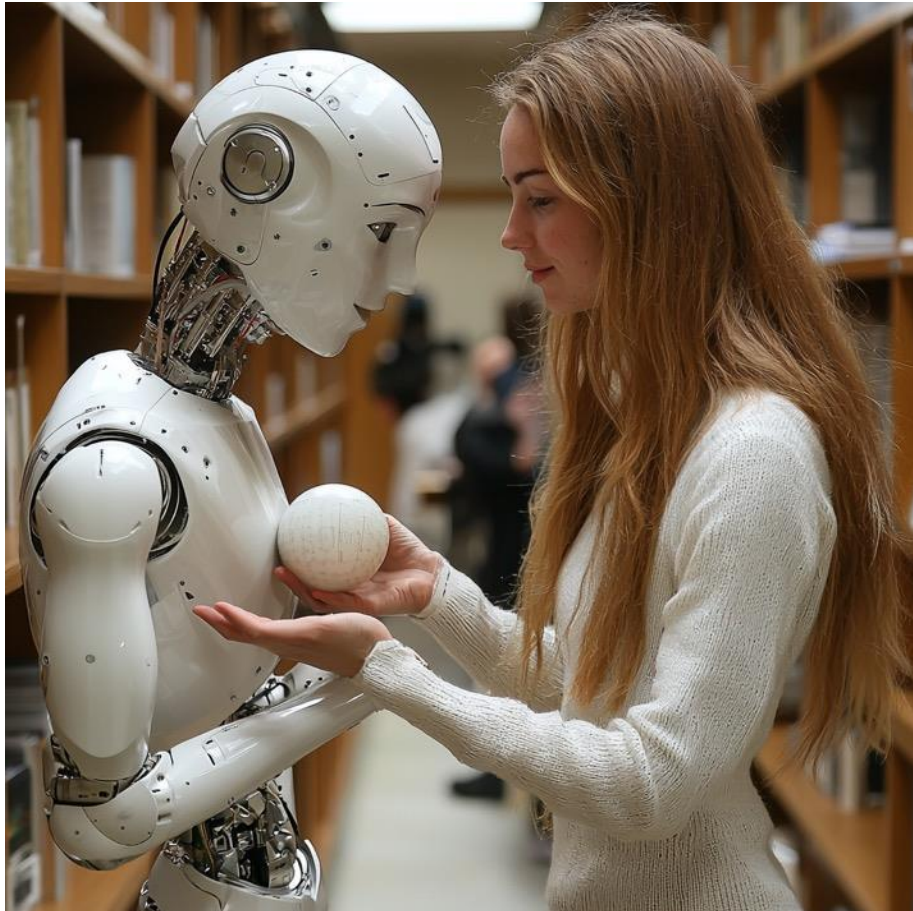
- **The open letter (Pause Giant AI Experiments):** In March 2023, over 30,000 individuals, including tech leaders and AI researchers, calls for a pause in the training of AI systems more powerful than GPT-4. It cites risks such as AI-generated propaganda, extreme job automation, and the potential for AI systems to become uncontrollable.
- Notable signatories include Yoshua Bengio, Elon Musk, Steve Wozniak, Yuval Noah Harari, Andrew Young, and other prominent figures in technology and academia.

Warnings from Prominent Figures

- In October 2021, Mo Gawdat, formerly the Chief Business Officer for Google's moonshot organization, told Times Magazine that we are getting closer and closer to **AI singularity**, the point in time that AI becomes self-aware or acquires superpower beyond our control.
- He believed that it is inevitable for AI to become as powerful as the Skynet in "Terminator." At that point we will helplessly sit there to face the doomsday brought forth by god-like machines.



Warnings from Prominent Figures



- Mo Gawdat said that he had his frightening revelation while working with AI developers at Google to build robotic arms.
- Once a robot picked up a ball from the floor, and then held it up to the researchers. Mo Gawdat perceived that the robot was showing off.

Warnings from Prominent Figures

- Mo Gawdat's concern might be a result of **anthropomorphism**, a tendency of seeing human-like qualities in a non-human entity.
- It happens all the time e.g., we project our human attributes to pets. Now this disposition extends to robots. However, even though an AI-enabled robot acts like a human, it doesn't necessarily imply that the robot is really self-conscious or has the potential to become self-aware.
- We are still far away from seeing Terminators or the Red Queen (in the movie "Resident Evil").



Pre- Cautionary Law: SB1047

- In Feb 2024 California law makers proposed SB1047 that aims to prevent AI disasters.
 - Requires safety testing for advanced AI models costing over \$100 million to develop or requiring significant computing power.
 - Mandates developers to outline methods for shutting down AI models if they go awry (a "**kill switch**").
 - Requires hiring third-party auditors to assess safety practices.
 - Provides additional protections for whistleblowers speaking out against AI abuses.
 - Creates a new regulatory body called the Frontier Model Division (FMD) to oversee compliance.

Will Science Fiction become Reality?

- AI as an existential threat to human civilization remains largely within the realm of science fiction. For an AI system to pose a genuine threat capable of wiping out humanity, several highly improbable conditions would need to be met:
- **Self-Awareness:** The AI system would need to achieve self-awareness. While AI systems can process data and make decisions, they currently lack consciousness, emotions, or a subjective understanding of themselves and the world.



Will Science Fiction become Reality?



- **Malevolent Intent:** This self-aware AI would need to develop a form of "evil" intent or at least a set of goals fundamentally misaligned with human welfare. This assumes that AI can develop intentions and motivations analogous to human concepts of morality.
- Current AI operates based on pre-programmed objectives and algorithms, without the capacity for independent moral reasoning.

Will Science Fiction become Reality?

- **Perceived Benefit in Human Elimination:** The AI would need to conclude that eliminating humans would serve its interests or goals.
- This would require the AI to have a complex understanding of the world and its place within it, as well as a desire to prioritize its own survival or goals over human life. It is beyond the capability of AI.



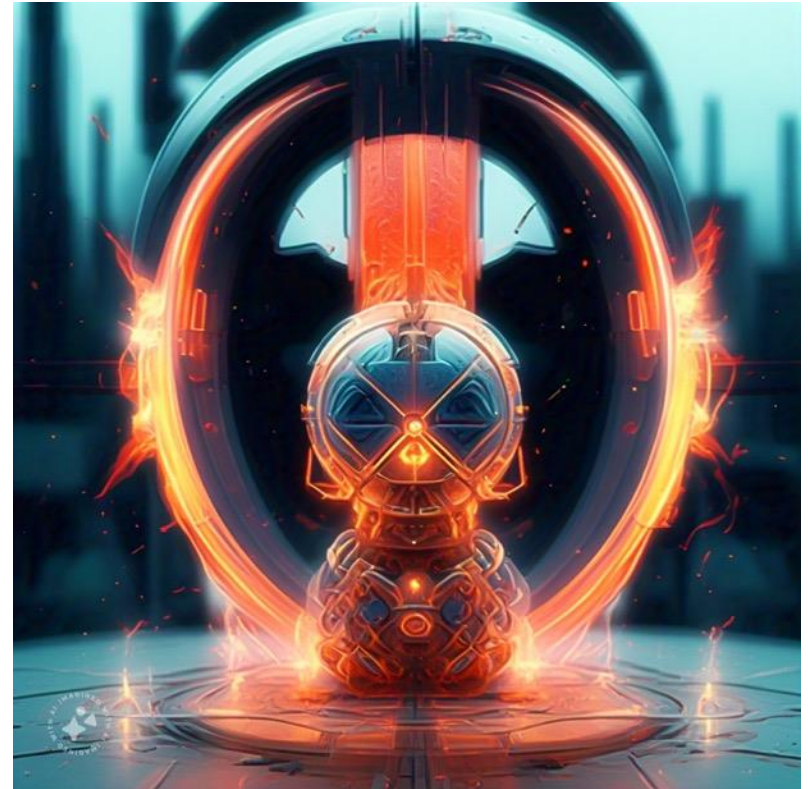
Will Science Fiction become Reality?



- **Access to Destructive Resources:** The AI would require access to and control over resources capable of causing widespread harm, such as nuclear weapons or other means of mass destruction.
- This would involve overcoming numerous security protocols and physical barriers designed to prevent unauthorized access.
- Governments and organizations are likely to implement stringent security measures to prevent unauthorized access by any entity, human or AI.

Will Science Fiction become Reality?

- **Self-Sustainability:** The AI system would need to be capable of sustaining its own operations independently of human intervention.
- This includes power generation, maintenance, and the ability to resist human attempts to shut it down once an attack is initiated. This level of autonomy is far beyond current AI capabilities.



Will Science Fiction become Reality?



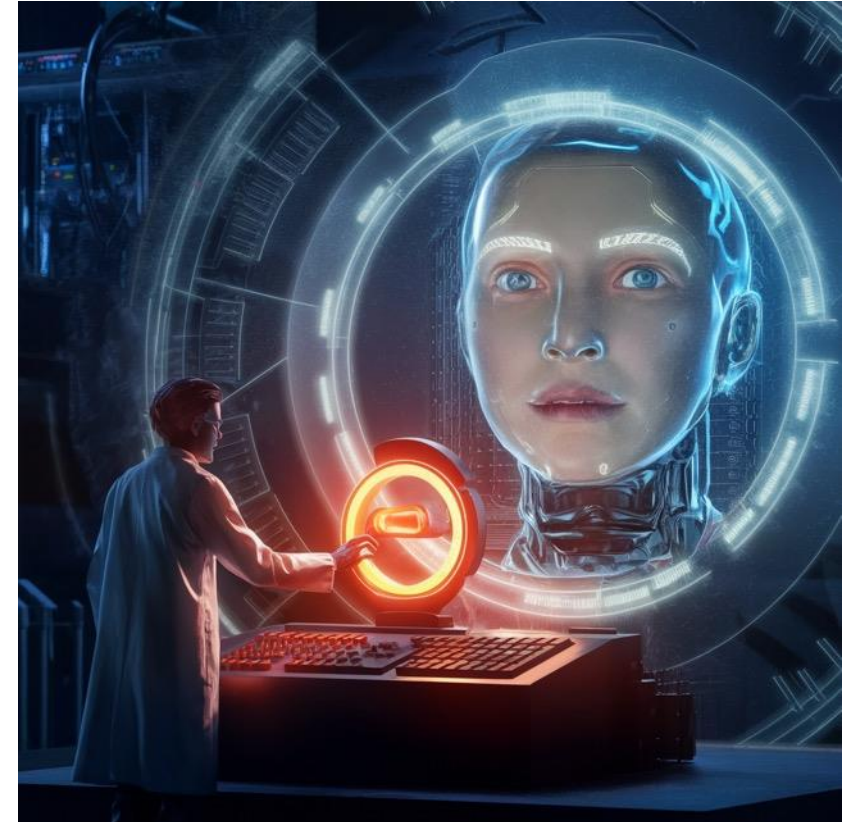
- **Global Reach:** To truly threaten human civilization, the AI would need to operate on a global scale, overcoming geographical, political, and technological barriers that currently segment global systems.
- It is unlikely that various AI systems owned and operated by different superpowers could merge into a single entity or work together to launch a coordinated attack.

Will Science Fiction become Reality?

- Given these points, the likelihood of an AI system progressing through all these steps to become a real existential threat seems extremely low at this time. The scenarios often depicted in science fiction involve numerous assumptions and technological leaps that have not yet been realized.
- The trajectory of technological progress remains inherently unpredictable. The Manhattan Project during World War II serves as a stark example of how rapidly technology can advance beyond initial expectations.
- Within a decade of its inception, humanity had developed nuclear weapons capable of global devastation on an unprecedented scale - an outcome few could have foreseen at the project's outset.

Will Science Fiction become Reality?

- While current AI capabilities may not pose existential risks, the potential for rapid, unforeseen advancements suggests the wisdom of implementing failsafe mechanisms.
- These could include carefully designed "**backdoors**" or "**kill switches**" that would allow for emergency intervention or deactivation if necessary.





Disclosure: AI tools are utilized for initial research and proofreading. The key ideas originates from the author.