SAFETY OF AI

Aarsha V S
Mtech DS & AI

What is AI safety?

- **AI safety** is an interdisciplinary field concerned with preventing accidents, misuse, or other harmful consequences that could result from artificial intelligence (AI) systems.

- It encompasses machine ethics and AI alignment, which aim to make AI systems moral and beneficial, and AI safety encompasses technical problems including monitoring systems for risks and making them highly reliable.

Risks of AI

There are a number of risks associated with AI, including:

- Unintentional harm: AI systems can make mistakes, which can lead to unintended harm.

  For example:

  - An AI system used to diagnose diseases may misdiagnose a patient, leading to the wrong treatment being prescribed.
  - An AI system that is used to control a self-driving car. If the system makes a mistake, it could lead to an accident

- Misuse:AI systems can also be misused for malicious purposes.

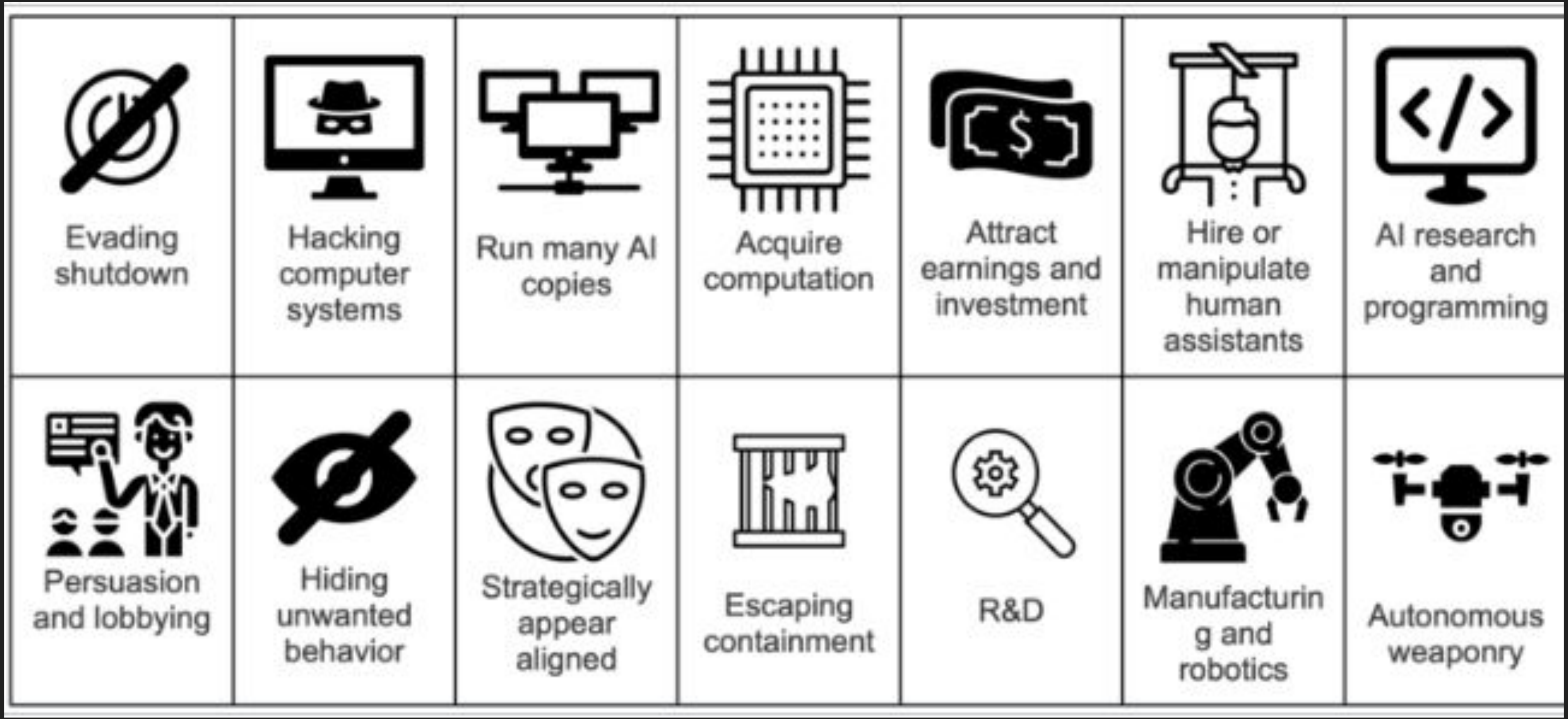
  For example:

  - AI could be used to develop autonomous weapons that could kill without human intervention.
  - AI could also be used to create surveillance systems that could violate people's privacy.

- Bias: AI systems can be biased, reflecting the biases of the data they are trained on.

  For example:

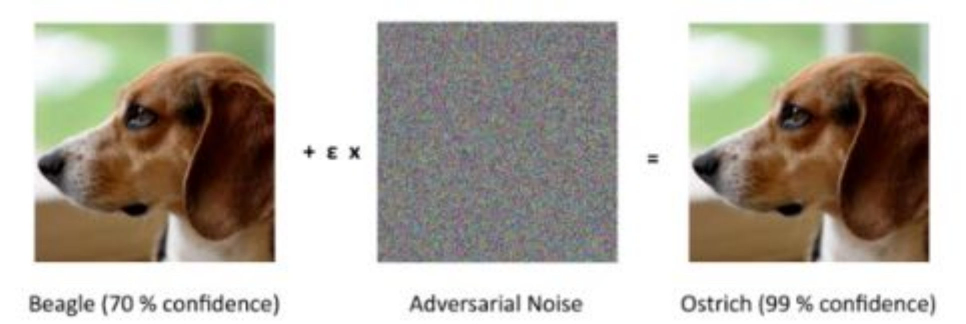
  - Amazon stopped using a hiring algorithm after finding it favored applicants based on words like “executed” or “captured,” which were more commonly found on men’s resumes.



Evading shutdown	Hacking computer systems	Run many AI copies	Acquire computation	Attract earnings and investment	Hire or manipulate human assistants	AI research and programming
Persuasion and lobbying	Hiding unwanted behavior	Strategically appear aligned	Escaping containment	R&D	Manufacturin g and robotics	Autonomous weaponry

Safety principles for AI

- Robustness:
  - Adversarial robustness:
    - AI systems are often vulnerable to adversarial examples or “inputs to machine learning models that an attacker has intentionally designed to cause the model to make a mistake.



- Models that represent objectives (reward models) must also be adversarially robust.

- For example, a reward model might estimate how helpful a text response is and a language model might be trained to maximize this score. If a language model is trained for long enough, it will leverage the vulnerabilities of the reward model to achieve a better score and perform worse on the intended task. This issue can be addressed by improving the adversarial robustness of the reward model.

- Any AI system used to evaluate another AI system must be adversarially robust.

- Transparency:
    - Neural networks have often been described as black boxes.This makes it challenging to anticipate failures.
    - Some benefits:
        - Explainability: It is sometimes a legal requirement to provide an explanation for why a decision was made in order to ensure fairness
        - Reveal the cause of failures: At the beginning of the 2020 COVID-19 pandemic, researchers used transparency tools to show that medical image classifiers were ‘paying attention’ to irrelevant hospital labels
        - To correct errors

- Alignment
    - AI alignment research aims to steer AI systems towards humans' intended goals, preferences, or ethical principles.
    - An AI system is considered aligned if it advances the intended objectives. A misaligned AI system pursues some objectives, but not the intended ones
    - AI systems may find loopholes that allow them to accomplish their proxy goals efficiently but in unintended, sometimes harmful ways.
    - They may develop undesirable emergent goals that may be hard to detect before the system is deployed, when it faces new situations and data distributions.

Safety measures for AI

To reduce the risks associated with AI:

- Data quality: AI systems should be trained on high-quality data that is representative of the real world. This will help to reduce the risk of bias and unintended harm.

- Testing: AI systems should be thoroughly tested before they are deployed in the real world. This will help to identify and fix any potential problems.

- Monitoring: AI systems should be monitored after they are deployed to identify and address any unexpected problems.

Conclusion

- AI is a powerful technology with the potential to improve our lives in many ways.

- However, it is important to ensure that AI systems are safe and reliable, and that they are used in a responsible and ethical manner.

- By following the safety principles and measures outlined above, we can help to ensure that AI is developed and used in a way that benefits all of humanity.

THANK YOU!