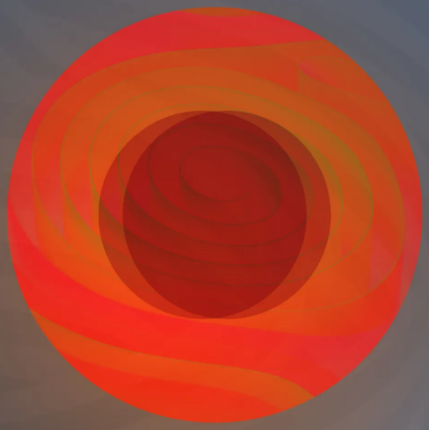
CENTER FOR AI SAFETY

8 Examples of AI Risk

AI has been compared to electricity and the steam engine in terms of its potential to transform society. The technology could be profoundly beneficial, but it also presents serious risks, due to competitive pressures and other factors.

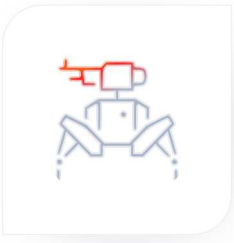


What is AI risk?

AI systems are rapidly becoming more capable. AI models can generate text, images, and video that are difficult to distinguish from human-created content. While AI has many beneficial applications, it can also be used to perpetuate bias, power autonomous weapons, promote misinformation, and conduct cyberattacks. Even as AI systems are used with human involvement, AI agents are increasingly able to act autonomously to cause harm (Chan et al., 2023).

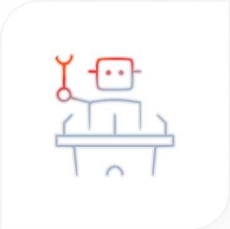
When AI becomes more advanced, it could eventually pose catastrophic or existential risks. There are many ways in which AI systems could pose or contribute to large-scale risks, some of which are enumerated below.

For more in-depth discussion of extreme risks, please also see our recent work "Natural Selection Favors AIs Over Humans" or Yoshua Bengio's "How Rogue AIs May Arise".



Malicious actors could repurpose AI to be highly destructive, presenting an existential risk in and of itself and increasing the probability ofpolitical destabilization. For example, deep reinforcement learning methods have been applied to aerial combat, and machine learning drug-discovery tools could be used to build chemical weapons.

In recent years, researchers have been developing AI systems for automated cyberattacks (Buchanan et al., 2020, Cary et al., 2020), military leaders have discussed giving AI systems decisive control over nuclear silos (Klare 2020), and superpowers of the world have declined to sign agreements banning autonomous weapons. An AI trained to develop drugs was easily repurposed to design potential biochemical weapons (Urbina et al., 2022). GPT-4, a model trained on internet text and coding, was able to autonomously conduct experiments and synthesize chemicals in a real-world lab (Boiko et al., 2023). An accident with an automated retaliation system could rapidly escalate and give rise to a major war. Looking forward, we note that since the nation with the most intelligent AI systems could have a strategic advantage, it may be challenging for nations to avoid building increasingly powerful weaponized AI systems. Even if all superpowers ensure that the systems they build are safe and agree not to build destructive AI technologies, rogue actors could still use AI to cause significant harm. Easy access to powerful AI systems increases the risk of unilateral, malicious usage. As with nuclear and biological weapons, only one irrational or malevolent actor is sufficient to cause harm on a massive scale. Unlike previous weapons, AI systems with dangerous capabilities could be easily proliferated through digital means.



2. Misinformation

A deluge of AI-generated misinformation and persuasive content could make society less-equipped to handle important challenges of our time.

States, parties, and organizations use technology to influence and convince others of their political beliefs, ideologies, and narratives. Emerging AI may bring this use-case into a new era and enable personally customized disinformation campaigns at scale. Additionally, AI itself could generate highly persuasive arguments that invoke strong emotional responses. Together, these trends could undermine collective decision-making, radicalize individuals, or derail moral progress.

Trained with faulty objectives, AI systems could find novel ways to pursue their goals at the expense of individual and societal values.

AI systems are trained using measurable objectives, which may only be indirect proxies for what we value. For example, AI recommender systems are trained to maximize watch time and click rate metrics. The content people are most likely to click on, however, is not necessarily the same as the content that will improve their well-being (Kross et al., 2013). Also, some evidence suggests that recommender systems cause people to develop extreme beliefs in order to make their preferences easier to predict (Jiang et al., 2019). As AI systems become more capable and influential, the objectives we use to train systems must be more carefully specified and incorporate shared human values.

4. Enfeeblement

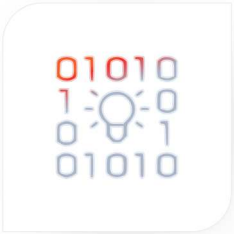
Enfeeblement can occur if important tasks are increasingly delegated to machines; in this situation, humanity loses the ability to self-govern and becomes completely dependent on machines, similar to the scenario portrayed in the film WALL-E.

As AI systems encroach on human-level intelligence, more and more aspects of human labor will become faster and cheaper to accomplish with AI. As the world accelerates, organizations may voluntarily cede control to AI systems in order to keep up. This may cause humans to become economically irrelevant, and once AI automates aspects of many industries, it may be hard for displaced humans to reenter them. In this world, humans could have few incentives to gain knowledge or skills. Many would consider such a world to be undesirable. Furthermore, enfeeblement would reduce humanity's control over the future, increasing the risk of bad longterm outcomes.

5. Value Lock-in

...ive small groups of people a tremendous ...ressive systems.

AI imbued with particular values may determine the values that are propagated into the future. Some argue that the exponentially increasing compute and data barriers to entry make AI a centralizing force. As time progresses, the most powerful AI systems may be designed by and available to fewer and fewer stakeholders. This may enable, for instance, regimes to enforce narrow values through pervasive surveillance and oppressive censorship. Overcoming such a regime could be unlikely, especially if we come to depend on it. Even if creators of these systems know their systems are self-serving or harmful to others, they may have incentives to reinforce their power and avoid distributing control.



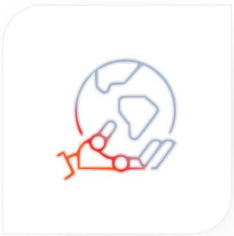
6. Emergent Goals

Models demonstrate unexpected, qualitatively different behavior as they become more competent. The sudden emergence of capabilities or goals could increase the risk that people lose control over advanced AI systems.

Capabilities and novel functionality can spontaneously emerge in today's AI systems (Ganguli et al., Power et al.), even though these capabilities were not anticipated by system designers. If we do not know what capabilities systems possess, systems become harder to control or safely deploy. Indeed, unintended latent capabilities may only be discovered during deployment. If any of these capabilities are hazardous, the effect may be irreversible. New system objectives could also emerge. For complex adaptive systems, including many AI agents, goals such as self-preservation often emerge (Hadfield-Menell et al). Goals can also undergo qualitative shifts through the emergence of intra-system goals (Gall, Hendrycks et al). In the future, agents may break down difficult long-term goals into smaller subgoals. However, breaking down goals can distort the objective, as the true objective may not be the sum of its parts. This distortion can result in misalignment. In more extreme cases, the intra-system goals could be pursued at the expense of the overall goal. For example, many companies create intra-system goals and have different specializing departments pursue these distinct subgoals. However, some departments, such as bureaucratic departments, can capture power and have the company pursue goals unlike its original goals. Even if we correctly specify our high-level objectives, systems may not operationally pursue our objectives (Hubinger et al). This is another way in which systems could fail to optimize human values.

We want to understand what powerful AI systems are doing and why they are doing what they are doing. One way to accomplish this is to have the systems themselves accurately report this information. This may be non-trivial however since being deceptive is useful for accomplishing a variety of goals.

Future AI systems could conceivably be deceptive not out of malice, but because deception can help agents achieve their goals. It may be more efficient to gain human approval through deception than to earn human approval legitimately. Deception also provides optionality: systems that have the capacity to be deceptive have strategic advantages over restricted, honest models. Strong AIs that can deceive humans could undermine human control. AI systems could also have incentives to bypass monitors. Historically, individuals and organizations have had incentives to bypass monitors. For example, Volkswagen programmed their engines to reduce emissions only when being monitored. This allowed them to achieve performance gains while retaining purportedly low emissions. Future AI agents could similarly switch strategies when being monitored and take steps to obscure their deception from monitors. Once deceptive AI systems are cleared by their monitors or once such systems can overpower them, these systems could take a "treacherous turn" and irreversibly bypass human control.



8. Power-Seeking Behavior

Companies and governments have strong economic incentives to create agents that can accomplish a broad set of goals. Such agents have instrumental incentives to acquire power, potentially making them harder to control (Turner et al., 2021, Carlsmith 2021).

AIs that acquire substantial power can become especially dangerous if they are not aligned with human values. Power-seeking behavior can also incentivize systems to pretend to be aligned, collude with other AIs, overpower monitors, and so on. On this view, inventing machines that are more powerful than us is playing with fire. Building power-seeking AI is also incentivized because political leaders see the strategic advantage in having the most intelligent, most powerful AI systems. For example, Vladimir Putin has said "Whoever becomes the leader in [AI] will become the ruler of the world."

How to analyze AI x-risk

To add precision and ground these discussions, we provide a guide for how to analyze AI x-risk, which consists of three parts:

1. First, we review how systems can be made safer today, drawing on time-tested concepts from hazard analysis and systems safety that have been designed to steer large processes in safer directions.
2. Next, we discuss strategies for having long-term impacts on the safety of future systems.
3. Finally, we discuss a crucial concept in making AI systems safer by improving the balance between safety and general capabilities.

We hope this document and the presented concepts and tools serve as a