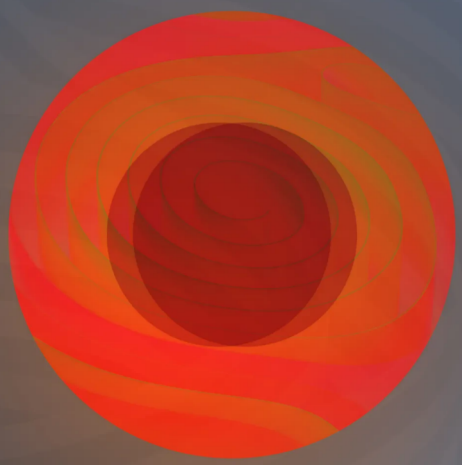
ZENTRUM FÜR KI-SICHERHEIT

8 Beispiele für KI-Risiko



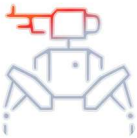
KI wurde hinsichtlich ihres Potenzials, die Gesellschaft zu verändern, mit Elektrizität und der Dampfmaschine verglichen. Die Technologie könnte von großem Nutzen sein, birgt jedoch aufgrund des Wettbewerbsdrucks und anderer Faktoren auch ernsthafte Risiken.

Was ist KI-Risiko?

KI-Systeme werden immer leistungsfähiger. KI-Modelle können Texte, Bilder und Videos generieren, die schwer von von Menschen erstellten Inhalten zu unterscheiden sind. Obwohl KI viele nützliche Anwendungen hat, kann sie auch dazu verwendet werden, Voreingenommenheit aufrechtzuerhalten, autonome Waffen anzutreiben, Fehlinformationen zu fördern und Cyberangriffe durchzuführen. Auch wenn KI-Systeme unter menschlicher Beteiligung eingesetzt werden, sind KI-Agenten zunehmend in der Lage, autonom zu handeln und Schaden anzurichten (Chan et al., 2023).

Wenn die KI weiter fortgeschritten ist, könnte sie letztendlich katastrophale oder existenzielle Risiken mit sich bringen. Es gibt viele Möglichkeiten, wie KI-Systeme große Risiken darstellen oder zu solchen beitragen können, von denen einige im Folgenden aufgeführt sind.

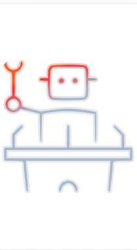
Für eine ausführlichere Diskussion extremer Risiken lesen Sie bitte auch unser aktuelles Werk „ Natural Selection Favours AIs Over Humans “ oder Yoshua Bengios „ How Rogue AIs May Arise “.



1. Bewaffnung

Böswillige Akteure könnten KI zu äußerst zerstörerischen Zwecken umfunktionieren, was an sich schon ein existenzielles Risiko darstellt und die Wahrscheinlichkeit einer politischen Destabilisierung erhöht. Beispielsweise wurden tiefgreifende Reinforcement-Learning-Methoden auf Luftkämpfe angewendet, und maschinell lernende Werkzeuge zur Drogenentdeckung könnten zum Bau chemischer Waffen eingesetzt werden .

In den letzten Jahren haben Forscher KI-Systeme für automatisierte Cyberangriffe entwickelt (Buchanan et al., 2020 , Cary et al., 2020), Militärführer haben darüber diskutiert, KI-Systemen entscheidende Kontrolle über Atomsilos zu geben (Klare 2020), und Supermächte der Die Welt lehnte es ab, Abkommen zum Verbot autonomer Waffen zu unterzeichnen. Eine für die Entwicklung von Medikamenten ausgebildete KI konnte leicht für die Entwicklung potenzieller biochemischer Waffen eingesetzt werden (Urbina et al., 2022). GPT-4, ein auf Internettext und Codierung trainiertes Modell, war in der Lage, in einem realen Labor autonom Experimente durchzuführen und Chemikalien zu synthetisieren (Boiko et al., 2023).). Ein Unfall mit einem automatisierten Vergeltungssystem könnte schnell eskalieren und einen großen Krieg auslösen. Mit Blick auf die Zukunft stellen wir fest, dass die Nation mit den intelligentesten KI-Systemen einen strategischen Vorteil haben könnte und es daher für die Nationen schwierig sein kann, den Bau immer leistungsfähigerer, waffenfähiger KI-Systeme zu vermeiden. Selbst wenn alle Supermächte sicherstellen, dass die von ihnen gebauten Systeme sicher sind, und sich darauf einigen, keine zerstörerischen KI-Technologien zu entwickeln, könnten betrügerische Akteure dennoch KI nutzen, um erheblichen Schaden anzurichten. Der einfache Zugriff auf leistungsstarke KI-Systeme erhöht das Risiko einer einseitigen, böswilligen Nutzung. Wie bei nuklearen und biologischen Waffen reicht bereits ein einziger irrationaler oder böswilliger Akteur aus, um Schaden in großem Ausmaß anzurichten. Im Gegensatz zu früheren Waffen könnten KI-Systeme mit gefährlichen Fähigkeiten mit digitalen Mitteln leicht verbreitet werden.



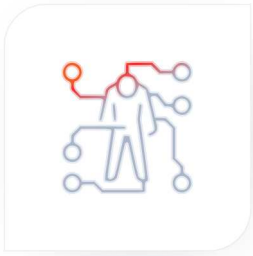
A deluge of AI-generated misinformation and persuasive content could make society less-equipped to handle important challenges of our time.

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.

3. Proxy Gaming

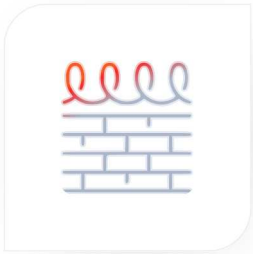
Trained with faulty objectives, AI systems could find novel ways to pursue their goals at the expense of individual and societal values.

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.



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.

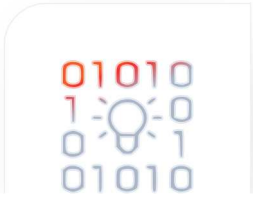
As AI systems encroach on human-level intelligence, more and more aspects of human labor will become faster and cheaper to accomplish with AI. As the world accelerates, organizations may voluntarily cede control to AI systems in order to keep up. This may cause humans to become economically irrelevant, 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

Highly competent systems could give small groups of people a tremendous amount of power, leading to a lock-in of oppressive 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.

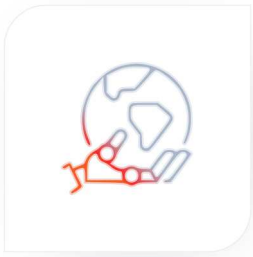
Capabilities and novel functionality can spontaneously emerge in today's AI systems (Ganguli et al., Power et al.), even though these capabilities were not anticipated by system designers. If we do not know what capabilities systems possess, systems become harder to control or safely deploy. Indeed, 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.

7. Deception

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

could also have incentives to bypass
ad incentives to
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 useful guide for understanding how to analyze AI x-risk.

Adapted from X-Risk Analysis for AI Research

Subscribe to the AI Safety Newsletter

E-Mail Adresse eingeben

Submit

No technical background required. See past newsletters.

Center for AI Safety

CAIS is an AI safety non-profit. Our mission is to reduce societal-scale risks from artificial intelligence.