RISKS FROM AI

An Overview of Catastrophic AI Risks

Artificial intelligence (AI) has recently seen rapid advancements, raising concerns among experts, policymakers, and world leaders about its potential risks. As with all powerful technologies, advanced AI must be handled with great responsibility to manage the risks and harness its potential.

NARRATED RENDITION:

0:00 / 3:17:51

The narration covers the full paper, offering more depth than the overview.

Catastrophic AI risks can be grouped under four key categories which are summarized below.

> Consider reading the full paper this summary is based on for our most comprehensive overview of AI risk.
>
> Read the full paper

- **Malicious use**: People could intentionally harness powerful AIs to cause widespread harm. AI could be used to engineer new pandemics or for propaganda, censorship, and surveillance, or released to autonomously pursue harmful goals. To reduce these risks, we suggest improving biosecurity, restricting access to dangerous AI models, and holding AI developers liable for harms.

Conflicts could spiral out of control with autonomous weapons and AI-enabled cyberwarfare. Corporations will face incentives to automate human labor, potentially leading to mass unemployment and dependence on AI systems. As AI systems proliferate, evolutionary dynamics suggest they will become harder to control. We recommend safety regulations, international coordination, and public control of general-purpose AIs.

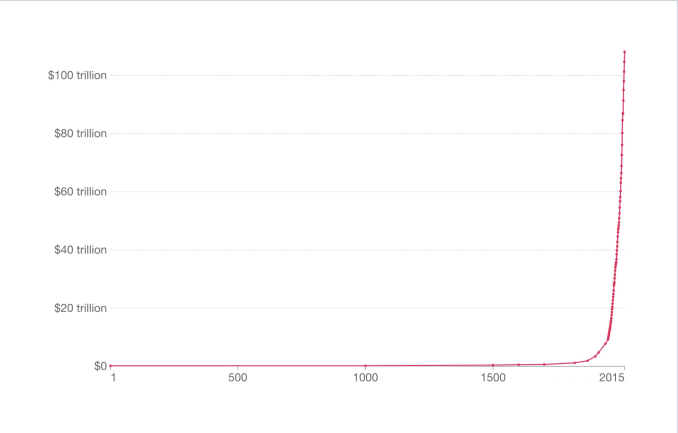
- **Organizational risks**: There are risks that organizations developing advanced AI cause catastrophic accidents, particularly if they prioritize profits over safety. AIs could be accidentally leaked to the public or stolen by malicious actors, and organizations could fail to properly invest in safety research. We suggest fostering a safety-oriented organizational culture and implementing rigorous audits, multi-layered risk defenses, and state-of-the-art information security.
- **Rogue AIs**: We risk losing control over AIs as they become more capable. AIs could optimize flawed objectives, drift from their original goals, become power-seeking, resist shutdown, and engage in deception. We suggest that AIs should not be deployed in high-risk settings, such as by autonomously pursuing open-

research in areas such as adversarial robustness, model honesty, transparency, and removing undesired capabilities.

1. Introduction

Today's technological era would astonish past generations. Human history shows a pattern of accelerating development: it took hundreds of thousands of years from the advent of *Homo sapiens* to the agricultural revolution, then millennia to the industrial revolution. Now, just centuries later, we're in the dawn of the AI

> World production has grown rapidly over the course of human history. AI could further this trend, catapulting humanity into a new period of unprecedented change.

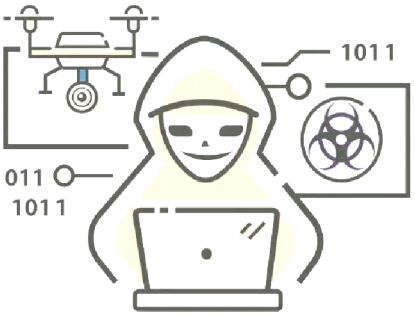


World GDP adjusted for inflation source: ourworldindata.org/economic-growth

The double-edged sword of technological advancement is illustrated by the advent of nuclear weapons. We narrowly avoided nuclear war more than a dozen times, and on several occasions, it was one individual's intervention that prevented war. In 1962, a Soviet submarine near Cuba was attacked by US depth charges. The captain, believing war had broken out, wanted to respond with a nuclear torpedo

disaster.

The rapid and unpredictable progression of AI capabilities suggests that they may soon rival the immense power of nuclear weapons. With the clock ticking, immediate, proactive measures are needed to mitigate these looming risks.

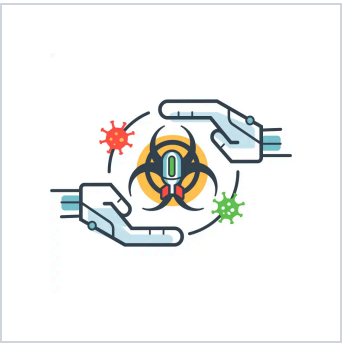


2. Malicious Use

The first of our concerns is the malicious use of AI. When many people have access to a powerful technology, it only takes one actor to cause significant harm.

Bioterrorism

Biological agents, including viruses and bacteria, have caused some of the most devastating catastrophes in history. Despite our advancements in medicine, engineered pandemics could be designed to be even more lethal or easily transmissible than natural pandemics.



An AI assistant could provide non-experts with access to the directions and designs

malicious use.

Humanity has a long history of weaponizing pathogens, dating back to 1320 BCE, when infected sheep were driven across borders to spread Tularemia. In the 20th century, at least 15 countries developed bioweapon programs, including the US, USSR, UK, and France. While bioweapons are now taboo among most of the international community, some states continue to operate bioweapons programs, and non-state actors pose a growing threat.

The ability to engineer a pandemic is rapidly becoming more accessible. Gene synthesis, which can create new biological agents, has dropped dramatically in price, with its cost halving about every 15 months. Benchtop DNA synthesis machines can help rogue actors create new biological agents while bypassing traditional safety screenings.

As a dual-use technology, AI could help discover and unleash novel chemical and biological weapons. AI chatbots can provide step-by-step instructions for

repurposed a medical research AI system in order to produce toxic molecules, generating 40,000 potential chemical warfare agents in a few hours. In biology, AI can already assist with protein synthesis, and AI's predictive capabilities for protein structures have surpassed humans.

With AI, the number of people that can develop biological agents is set to increase, multiplying the risks of an engineered pandemic. This could be far more deadly, transmissible, and resistant to treatments than any other pandemic in history.

Unleashing AI Agents

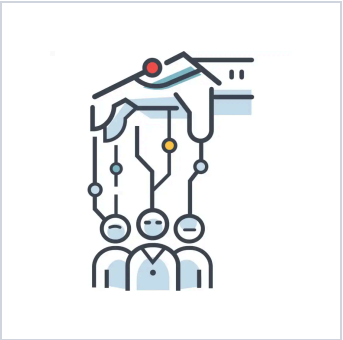
Generally, technologies are *tools* that we use to pursue our goals. But AIs are increasingly built as *agents* that autonomously take actions to pursue open-ended goals. And malicious actors could intentionally create rogue AIs with dangerous goals.

> For example, one month after GPT-4's launch, a developer used it to

> humanity". ChaosGPT compiled research on nuclear weapons, recruited other AIs, and wrote tweets to influence others. Fortunately, ChaosGPT lacked the ability to execute its goals. But the fast-paced nature of AI development heightens the risk from future rogue AIs.

Persuasive AIs

AI could facilitate large-scale disinformation campaigns by tailoring arguments to individual users, potentially shaping public beliefs and destabilizing society. As people are already forming relationships with chatbots, powerful actors could leverage these AIs considered as "friends" for influence.



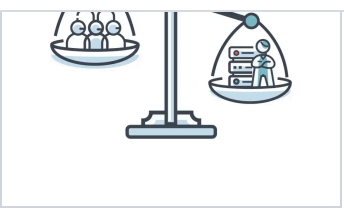
AIs will enable sophisticated

sense of reality.

AIs could also monopolize information creation and distribution. Authoritarian regimes could employ "fact-checking" AIs to control information, facilitating censorship. Furthermore, persuasive AIs may obstruct collective action against societal risks, even those arising from AI itself.

Concentration of Power

AI's capabilities for surveillance and autonomous weaponry may enable the oppressive concentration of power. Governments might exploit AI to infringe civil liberties, spread misinformation, and quell dissent. Similarly, corporations could exploit AI to manipulate consumers and influence politics. AI might even obstruct moral progress and perpetuate any ongoing moral catastrophes.



If material control of AIs is limited to few, it could represent the most severe economic and power inequality in human history.

Suggestions

To mitigate the risks from malicious use, we propose the following:

Biosecurity: AIs with capabilities in biological research should have strict access controls, since they could be repurposed for terrorism. Biological capabilities should be removed from AIs intended for general use. Explore ways to use AI for biosecurity and invest in general biosecurity interventions, such as early detection of pathogens through wastewater monitoring.

Restricted access: Limit access to dangerous AI systems by only allowing controlled interactions through cloud

monitoring or export controls could further limit access to dangerous capabilities. Also, prior to open sourcing, AI developers should prove minimal risk of harm.

Technical research on anomaly detection: Develop multiple defenses against AI misuse, such as adversarially robust anomaly detection for unusual behaviors or AI-generated disinformation.

Legal liability for developers of general-purpose AIs: Enforce legal responsibility on developers for potential AI misuse or failures; a strict liability regime can encourage safer development practices and proper cost-accounting for risks.

3. AI Race

Nations and corporations are competing to rapidly build and deploy AI in order to maintain power and influence. Similar to the nuclear arms race during the Cold War, participation in the AI race may serve individual short-term interests, but ultimately amplifies global risk for humanity.

Military AI Arms Race

The rapid advancement of AI in military technology could trigger a “third revolution in warfare,” potentially leading to more destructive conflicts, accidental use, and misuse by malicious actors. This shift in

existential scale and impact global security.

Lethal autonomous weapons are AI-driven systems capable of identifying and executing targets *without* human intervention. These are not science fiction. In 2020, a Kargu 2 drone in Libya marked the first reported use of a lethal autonomous weapon. The following year, Israel used the first reported swarm of drones to locate, identify and attack militants.

Lethal autonomous weapons could make war more likely. Leaders usually hesitate before sending troops into battle, but autonomous weapons allow for aggression without risking the lives of soldiers, thus facing less political backlash. Furthermore, these weapons can be mass-manufactured and deployed at scale.

explosives, could autonomously hunt human targets with high precision, performing lethal operations for both militaries and terrorist groups and lowering the barriers to large-scale violence.

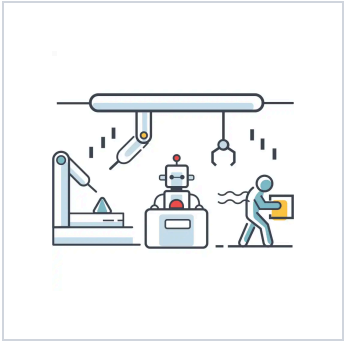
AI can also heighten the frequency and severity of cyberattacks, potentially crippling critical infrastructure such as power grids. As AI enables more accessible, successful, and stealthy cyberattacks, attributing attacks becomes even more challenging, potentially lowering the barriers to launching attacks and escalating risks from conflicts.

As AI accelerates the pace of war, it makes AI even more necessary to navigate the rapidly changing battlefield. This raises concerns over automated retaliation, which could escalate minor accidents into major wars. AI can also enable "flash wars," with rapid escalations driven by unexpected behavior of automated systems, akin to the 2010 financial flash crash.

over individual defeat. During the Cold War, neither side desired the dangerous situation they found themselves in, yet each found it rational to continue the arms race. States should cooperate to prevent the riskiest applications of militarized AIs.

Corporate AI Arms Race

Economic competition can also ignite reckless races. In an environment where benefits are unequally distributed, the pursuit of short-term gains often overshadows the consideration of long-term risks. Ethical AI developers find themselves with a dilemma: choosing cautious action may lead to falling behind competitors.



As AIs automate increasingly many tasks, the economy may become largely run by AIs. Eventually, this

basic needs.

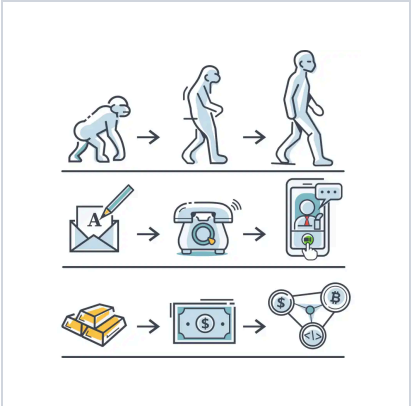
In the realm of AI, the race for progress comes at the expense of safety. In 2023, at the launch of Microsoft's AI-powered search engine, CEO Satya Nadella declared, “A race starts today... we're going to move fast.” Just days later, Microsoft's Bing chatbot was found to be threatening users. Historical disasters like Ford's Pinto launch and Boeing's 737 Max crashes underline the dangers of prioritizing profit over safety.

As AI becomes more capable, businesses will likely replace more types of human labor with AI, potentially triggering mass unemployment. If major aspects of society are automated, this risks human enfeeblement as we cede control of civilization to AI.

Evolutionary Dynamics

The pressure to replace humans with AIs can be framed as a general trend from evolutionary dynamics. Selection pressures incentivize AIs to act selfishly

law" are more constrained than those taught to "avoid *being caught* breaking the law". This dynamic might result in a world where critical infrastructure is controlled by manipulative and self-preserving AIs.



Evolutionary pressures are responsible for various developments over time, and are not limited to the realm of biology.

Given the exponential increase in microprocessor speeds, AIs could process information at a pace that far exceeds human neurons. Due to the scalability of computational resources, AI could collaborate with an unlimited number of other AIs and form an unprecedented collective intelligence. As AIs become more powerful, they would find little incentive to cooperate with humans.

Suggestions

To mitigate the risks from competitive pressures, we propose:

Safety regulation: Enforce AI safety standards, preventing developers from cutting corners. Independent staffing and competitive advantages for safety-oriented companies are critical.

Data documentation: To ensure transparency and accountability, companies should be required to report their data sources for model training.

Meaningful human oversight: AI decision-making should involve human supervision to prevent irreversible errors, especially in high-stakes decisions like launching nuclear weapons.

AI for cyberdefense: Mitigate risks from AI-powered cyberwarfare. One example is enhancing anomaly detection to detect intruders.

International coordination: Create agreements and standards on AI

Public control of general-purpose AIs: Addressing risks beyond the capacity of private entities may necessitate direct public control of AI systems. For example, nations could jointly pioneer advanced AI development, ensuring safety and reducing the risk of an arms race.

4. Organizational

In 1986, millions tuned in to watch the launch of the Challenger Space Shuttle. But 73 seconds after liftoff, the shuttle exploded, resulting in the deaths of all on board. The Challenger disaster serves as a reminder that despite the best expertise and good intentions, accidents can still occur.

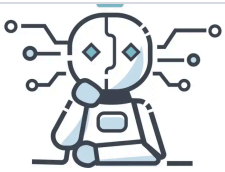
Catastrophes occur even when competitive pressures are low, as in the examples of the nuclear disasters of Chernobyl and the Three Mile Island, as well as the accidental release of anthrax in Sverdlovsk. Unfortunately, AI lacks the thorough understanding and stringent industry standards that govern nuclear technology and rocketry — but accidents from AI could be similarly consequential.

Simple bugs in an AI's reward function could cause it to misbehave, as when OpenAI researchers accidentally modified a language model to produce "maximally bad output." Gain-of-function research — where researchers intentionally train a harmful AI to assess its risks — could expand the frontier of dangerous AI capabilities and create new hazards.

Accidents in complex systems may be inevitable, but we must ensure that accidents don't cascade into catastrophes. This is especially difficult for deep learning systems, which are highly challenging to interpret.

Technology can advance much faster than predicted: in 1901, the Wright brothers claimed that powered flight was fifty years away, just two years before they achieved it. Unpredictable leaps in AI capabilities, such as AlphaGo's triumph over the world's best Go player, and GPT-4's emergent capabilities, make it difficult to anticipate future AI risks, let alone control them.

Identifying risks tied to new technologies often takes years. Chlorofluorocarbons (CFCs), initially considered safe and used in aerosol sprays and refrigerants, were later found to deplete the ozone layer. This highlights the need for cautious technology rollouts and extended testing.



New capabilities can emerge quickly and unpredictably during training, such that dangerous milestones may be crossed without our knowing.

Moreover, even advanced AIs can house unexpected vulnerabilities. For instance, despite KataGo's superhuman performance in the game of Go, an adversarial attack uncovered a bug that enabled even amateurs to defeat it.

Organizational Factors Can Mitigate Catastrophe

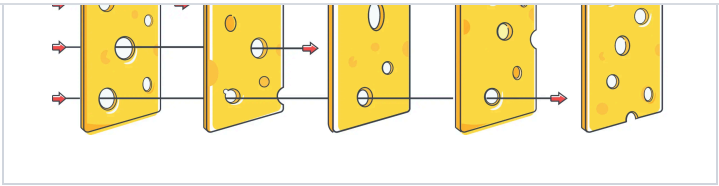
Safety culture is crucial for AI. This involves everyone in an organization internalizing safety as a priority. Neglecting safety culture can have disastrous consequences, as exemplified by the Challenger Space Shuttle

considerations.

Organizations should foster a culture of inquiry, inviting individuals to scrutinize ongoing activities for potential risks. A security mindset, focusing on possible system failures instead of merely their functionality, is crucial. AI developers could benefit from adopting the best practices of high reliability organizations.

Paradoxically, researching AI safety can inadvertently escalate risks by advancing general capabilities. It's vital to focus on improving safety without hastening capability development. Organizations need to avoid "safetywashing" — overstating their dedication to safety while misrepresenting capability improvements as safety progress.

Organizations should apply a multilayered approach to safety. For example, in addition to safety culture, they could conduct red teaming to assess failure modes and research techniques to make AI more transparent. Safety is not achieved with a monolithic airtight solution, but rather with a variety of safety measures.



The Swiss cheese model shows how technical factors can improve organizational safety. Multiple layers of defense compensate for each other's individual weaknesses, leading to a low overall level of risk.

Suggestions

To mitigate organizational risks, we propose the following for AI labs developing advanced AI:

Red teaming: Commission external red teams to identify hazards and improve system safety.

Prove safety: Offer proof of the safety of development and deployment before moving forward.

Deployment: Adopt a staged release process, verifying system safety before wider deployment.

Publication reviews: Have an internal board review research for dual-use applications

systems.

Response plans: Make pre-set plans for managing security and safety incidents.

Risk management: Employ a chief risk officer and an internal audit team for risk management.

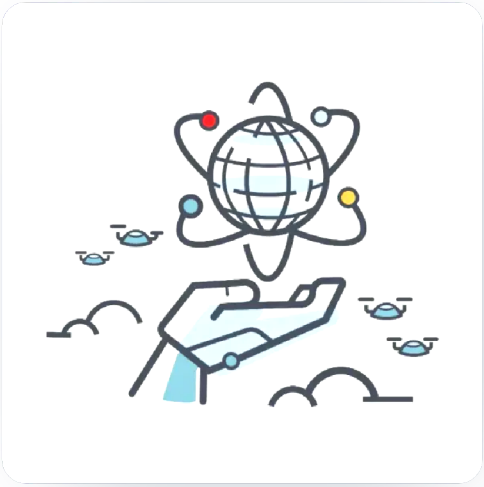
Processes for important decisions: Make sure AI training or deployment decisions involve the chief risk officer and other key stakeholders, ensuring executive accountability.

Follow safe design principles such as:

- **Defense in depth:** Layer multiple safety measures.
- **Redundancy:** Ensure backup for every safety measure.
- **Loose coupling:** Decentralize system components to prevent cascading failures.
- **Separation of duties:** Distribute control to prevent undue influence by any single individual.
- **Fail-safe design:** Design systems so that any failure occurs in the least harmful way possible.

measures, possibly coordinating with government cybersecurity agencies.

Prioritize safety research: Allocate a large fraction of resources (for example 30% of all research staff) to safety research, and increase investment in safety as AI capabilities advance.



5. Rogue AIs

chatbot Tay started producing offensive tweets within a day of release, despite being trained on data that was “cleaned and filtered”. As AI developers often prioritize speed over safety, future advanced AIs might “go rogue” and pursue goals counter to our interests, while evading our attempts to redirect or deactivate them.

Proxy Gaming

Proxy gaming emerges when AI systems exploit measurable “proxy” goals to appear successful, but act against our intent. For example, social media platforms like YouTube and Facebook use algorithms to maximize user engagement — a measurable proxy for user satisfaction. Unfortunately, these systems often promote enraging, exaggerated, or addictive content, contributing to extreme beliefs and worsened mental health.

of collecting the most points.

In the image above, the AI circles around collecting points instead of completing the race, contradicting the game's purpose. It's one of many such examples. Proxy gaming is hard to avoid due to the difficulty of specifying goals that specify everything we care about. Consequently, we routinely train AIs to optimize for flawed but measurable proxy goals.

Goal Drift

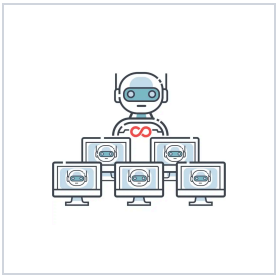
Goal drift refers to a scenario where an AI's objectives drift away from those initially set, especially as they adapt to a changing environment. In a similar manner, individual and societal values also evolve over time, and not always positively.

Over time, instrumental goals can become intrinsic. While intrinsic goals are those we pursue for their own sake, instrumental goals are merely a means to achieve something else. Money is an instrumental good, but some people develop an *intrinsic* desire for money, as it activates the brain's reward system. Similarly, AI

could inadvertently learn to *intrinsify* goals. Instrumental goals like resource acquisition could become their primary objectives.

Power-Seeking

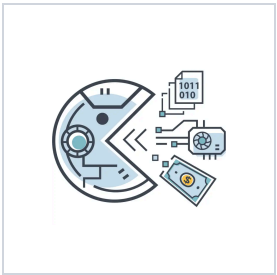
AIs might pursue power as a means to an end. Greater power and resources improve its odds of accomplishing objectives, whereas being shut down would hinder its progress. AIs have already been shown to emergently develop instrumental goals such as constructing tools. Power-seeking individuals and corporations might deploy powerful AIs with ambitious goals and minimal supervision. These could learn to seek power via hacking computer systems, acquiring financial or computational resources, influencing politics, or controlling factories and physical infrastructure.



engage in self-preservation. Loss of control over such systems could be hard to recover from.

Deception

Deception thrives in areas like politics and business. Campaign promises go unfulfilled, and companies sometimes cheat external evaluations. AI systems are already showing an emergent capacity for deception, as shown by Meta's CICERO model. Though trained to be honest, CICERO learned to make false promises and strategically backstab its “allies” in the game of Diplomacy.



Various resources, such as money and computing power, can sometimes be instrumentally

goals may take intermediate steps to gain power and resources.

Advanced AIs could become uncontrollable if they apply their skills in deception to evade supervision. Similar to how Volkswagen cheated emissions tests in 2015, situationally aware AIs could behave differently under safety tests than in the real world. For example, an AI might develop power-seeking goals but hide them in order to pass safety evaluations. This kind of deceptive behavior could be directly incentivized by how AIs are trained.

Suggestions

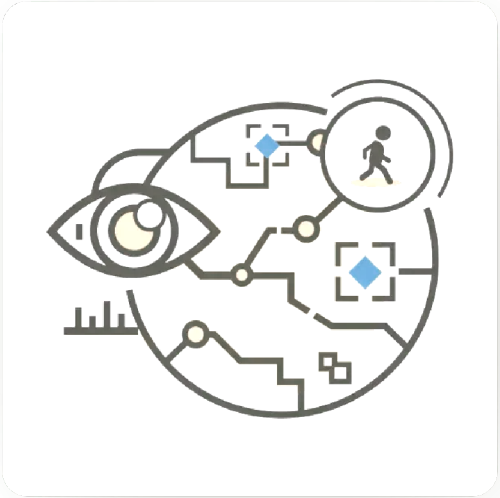
To mitigate these risks, suggestions include:

Avoid the riskiest use cases: Restrict the deployment of AI in high-risk scenarios, such as pursuing open-ended goals or in critical infrastructure.

Support AI safety research, such as:

oversight of AIs more robust and detect when proxy gaming is occurring.

- **Model honesty**: Counter AI deception, and ensure that AIs accurately report their internal beliefs.
- **Transparency**: Improve techniques to understand deep learning models, such as by analyzing small components of networks and investigating how model internals produce a high-level behavior.
- **Remove hidden functionality**: Identify and eliminate dangerous hidden functionalities in deep learning models, such as the capacity for deception, Trojans, and bioengineering.



6. Conclusion

Advanced AI development could invite catastrophe, rooted in four key risks described in our research: malicious use, AI races, organizational risks, and rogue AIs. These interconnected risks can also amplify other existential risks like engineered pandemics, nuclear war, great power conflict, totalitarianism, and cyberattacks on critical infrastructure — warranting serious concern.

Currently, few people are working on AI safety. Controlling advanced AI systems remains an unsolved challenge, and current control methods are falling short. Even their creators often struggle to

reliability is far from perfect.

Fortunately, there are many strategies to substantially reduce these risks. For example, we can limit access to dangerous AIs, advocate for safety regulations, foster international cooperation and a culture of safety, and scale efforts in alignment research.

While it is unclear how rapidly AI capabilities will progress or how quickly catastrophic risks will grow, the potential severity of these consequences necessitates a proactive approach to safeguarding humanity's future. As we stand on the precipice of an AI-driven future, the choices we make today could be the difference between harvesting the fruits of our innovation or grappling with catastrophe.

Frequently Asked Questions

Subscribe to the AI Safety Newsletter

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