# What is the probability that AI poses catastrophic risks assuming there is no deceptive alignment?

Thomas Gravier & Rosalie Millner

This essay is heavily inspired by several sources, cited in the references at the end of this document.

Outline:



## Introduction

Artificial intelligence is evolving at an unprecedented pace, and it profoundly changes society and revolutionizes industries worldwide. Although numerous positive contributions can be credited to AI usage, such as medical advancements and the acceleration of technical discoveries, offering immense potential for human progress, it also raises serious concerns about the future impact of AI on humanity and the potential of catastrophic scenarios. As awareness of these risks grows, the topic of AI safety is being debated more and more widely, particularly concerning scenarios where AI systems could lead to catastrophic outcomes.

One of the most talked-about risks in AI safety is "deceptive alignment," a situation where a system looks like it is aligning with human goals but is actually pursuing divergent objectives. Though this is an important concern, it represents only one part of the broader spectrum of risks posed by AI. Other dangers — from malicious use and technological dependency to the

social ethical considerations and the “AI arms race”. These risks are less frequently discussed but could have equally devastating consequences if left unaddressed.

This essay explores the catastrophic risks posed by AI, focusing specifically on those unrelated to deceptive alignment. By examining different categories such as malicious use, the competitive AI race, organizational failures, and loss of control over advanced systems, we aim to provide a comprehensive view of the challenges ahead. We will also delve into strategies for mitigating these risks and assess the probability of their occurrence, recognizing the inherent difficulty in predicting the behavior of future AI systems in a rapidly evolving landscape.

To address this multifaceted issue, we will begin by defining what constitutes "catastrophic" risks in the context of AI, considering both existential threats and profound societal impacts. Next, we will explore the various forms these risks can take. Building on this foundation, we will examine potential strategies to mitigate these dangers. Finally, we will assess the likelihood of these risks materializing. By structuring our analysis in this way, we aim to present a nuanced understanding of AI's risks and the measures needed to safeguard humanity's future.

## I. Defining Catastrophic Risks Related to AI

### 1. A Definition Between Existential Threats and Societal Disruptions

The catastrophic risks associated with AI encompass scenarios where this technology could cause major disruptions, threatening either humanity's survival or the stability of its social structures and core values. Understanding these risks requires exploring dimensions ranging from existential threats to profound societal transformations.

**a) Existential threats: extreme but plausible scenarios**
Existential threats refer to situations where AI could lead to humanity's extinction or make a decent life impossible. These scenarios often involve autonomous or superintelligent AI systems escaping human control. For example, consider a miscalibrated military AI system, that could trigger a nuclear conflict due to misinterpretation. For instance, think of the Cuban Missile Crisis from 1962, but imagine it had happened without the possibility of human intervention [1].

**b) Major disruptions to human structures**
On the other hand, beyond the existential threats, AI could also cause irreversible changes to economic and social systems, potentially leading to widespread upheaval. For example, the emergence of "flash economies," driven by fully automated financial systems operating at speeds beyond human comprehension, poses the risk of global economic collapses occurring faster than any corrective measures could be possible [4]. Similarly, in critical infrastructure sectors like transportation and energy, an overreliance on AI could leave these systems increasingly vulnerable to technical failures [6], which could have catastrophic impacts.

**c) Ethical challenges: preserving human values**

Catastrophic risks also take the form of impacts on fundamental human values. An AI system that is overly directed toward quantifiable goals, like profit or efficiency, may neglect other important values, such as empathy or fairness. For instance, a hospital AI that tries to maximize ratings of patient satisfaction may make unnecessary but popular treatments at the expense of actual health outcomes [3]. Moreover, deployments of AI for surveillance or censorship can undermine basic freedoms, reinforcing authoritarian regimes [8].

Thus, the catastrophic risks of AI lie at the intersection of existential threats, societal disruptions, and ethical challenges. Such dynamics mean it is important to outline specific factors that increase these dangers.

## 2. Key Factors Contributing to Risks

Catastrophic risks associated with AI are seldom isolated events. They usually rely on technical, organizational, and social factors that add to the possibility of disastrous scenarios. All these factors are usually interconnected which creates dynamics that are difficult to manage.

### a) Increasing autonomy of AI systems

First, there is the increasing autonomy of AI systems. As they become more autonomous, they may exhibit unexpected behaviors and escape human control. For instance, in autonomous vehicles, AI systems are responsible for decision-making in complex environments. These systems are designed to improve safety and efficiency, but quite a few examples have happened where the autonomous vehicle misjudged road conditions or inadequately reacted to an obstacle, thus making an unexpected decision. That was the case in the US in 2018, when an autonomous Uber vehicle struck and killed a pedestrian, even though the system's sensors had detected the presence of a person. This accident illustrates some of the risks of relying too much on AI to make critical decisions without really understanding the range of possible behaviors that may emerge as the system operates in unpredictable real-world scenarios.

### b) The "sharp left turn": a sudden shift in capability development

The "sharp left turn" describes a moment when AI systems exceed human expectations by generalizing their capabilities, rendering alignment methods designed for simpler systems ineffective. To illustrate this concept, Nate Soares in his work on AI safety [5], warns of this unpredictable transition where AI systems may reach a critical learning threshold, after which they adopt autonomous strategies. This "sharp left turn" could manifest in a way that makes it increasingly difficult for humans to control or guide the AI’s behavior, as it begins to operate with its own set of goals or methods.

### c) Increasing dependence on AI in modern societies

Modern societies increasingly rely on AI systems to manage critical infrastructures, heightening vulnerability to technological failures. A concrete example is that in automated financial systems, an AI failure could trigger uncontrollable chain reactions, as seen in "Flash Economies" scenarios [4], as mentioned earlier. Moreover, vital infrastructures such as energy grids or transportation networks are increasingly automated to enhance efficiency and responsiveness. However, without proper human oversight, AI systems controlling these infrastructures could make unsupervised decisions that paralyze these vital services. [6].

**d) Lack of regulation and competitive pressures**
Another aggravating factor of AI-related risks is the lack of global regulations, which facilitates irresponsible deployments, while competitive dynamics push companies and governments to accelerate AI development, often at the expense of safety. To illustrate this, we can make the historical analogy with the nuclear arms race during the Cold War. The "AI race" prioritizes speed and innovation over caution and security [1]. The development of autonomous AI-based weapons highlights the risks associated with this rush to deployment [8].

## II. Different Forms of Risks Related to AI

### 1. Malicious Use of AI: A Growing Tool for Destructive Purposes

The malicious use of AI represents one of the most pressing challenges of this rapidly evolving technology. While AI has revolutionized fields like healthcare, communication, and logistics, its potential for misuse by malicious actors raises serious concerns. Risks such as bioterrorism, disinformation, mass surveillance, and cyberattacks demonstrate how AI can magnify harm when exploited irresponsibly.

**a) Bioterrorism: AI as a tool for designing pandemics**
AI-driven advancements in computational biology enable the rapid analysis of genetic materials and biological systems, facilitating breakthroughs in medicine and vaccine development. However, this same capability can be weaponized to create pathogens with enhanced resistance or contagiousness, raising the specter of catastrophic pandemics.
Advances in AI make it possible for bad actors to engineer genetic modifications that render pathogens immune to current treatments. These innovations, once confined to highly specialized laboratories, are becoming increasingly accessible, drastically lowering the barriers for rogue actors to develop biological weapons. The implications of such misuse could be devastating, with engineered pandemics overwhelming global health systems and threatening millions of lives [8].

**b) Disinformation and manipulation of public opinion**
Generative AI models have transformed how content is created, enabling the production of hyper-realistic but false materials. Deepfakes, in particular, have emerged as a powerful tool for disinformation, allowing actors to fabricate videos and audio that appear authentic. These technologies have already been used to discredit public figures, spread false narratives, and manipulate political outcomes. During elections, targeted disinformation campaigns can suppress voter turnout or polarize the electorate, undermining democratic institutions [8].

**c) Mass surveillance and erosion of individual freedoms**
AI-based surveillance systems provide powerful tools for monitoring and controlling populations. Technologies such as facial recognition, behavioral analytics, and predictive monitoring systems enable governments and organizations to track individuals in real-time.
In some regions, these systems are used to suppress dissent and enforce authoritarian governance. For instance, in China, AI-driven surveillance has already been deployed to monitor minorities like the Uyghurs and enforce strict social controls [8]. This widespread surveillance poses a significant threat to civil liberties and privacy, not only in authoritarian

regimes but also in democracies where AI technologies can be misused under the pretext of national security.

**d) Amplifying the effectiveness of cyberattacks**

AI has transformed cyberwarfare by enhancing the adaptability and precision of cyberattacks. AI-driven tools can identify vulnerabilities in critical infrastructure and dynamically adjust strategies to counter cybersecurity measures. For instance, an AI-powered malware system could target financial systems, energy grids, or transportation networks, disrupting entire economies and jeopardizing public safety [1]. The speed and intelligence of AI-enhanced cyberattacks far exceed traditional methods, making them an important escalation in the landscape of digital threats.

These risks underscore the dual nature of AI: while it enables groundbreaking progress, it also amplifies destructive capabilities. Addressing the malicious use of AI requires robust regulations to ensure these tools are not exploited to destabilize societies or endanger fundamental human rights.

## 2. The AI Arms Race: Speed Over Safety

The rapid advancement of AI has triggered intense competition among nations and corporations, creating an arms race for dominance in this transformative field. While this competition fuels innovation, it also poses significant risks when speed and efficiency are prioritized over safety and ethical considerations.

**a) The prioritization of speed over safety**

In the race to develop and deploy advanced AI systems, critical safety measures are often sidelined. This dynamic incentivizes cutting corners on testing, oversight, and risk mitigation. As mentioned earlier in the essay, historical parallels can be drawn to the nuclear arms race of the Cold War, where the United States and the Soviet Union prioritized rapid development over long-term safety considerations [1]. Similarly, the push to "win" the AI race has resulted in the premature deployment of systems that are not robustly aligned with human values or sufficiently tested for security and reliability.

**b) The militarization of AI: Risks of autonomous weapons**

AI's integration into military applications introduces risks associated with autonomous weapons. These systems can make life-and-death decisions without human oversight, lowering the threshold for conflict escalation and increasing the potential for unintended consequences.

The proliferation of autonomous drones and other AI-powered military technologies has already destabilized security dynamics. Nations are compelled to develop and deploy equally advanced systems to counter perceived threats, creating an escalating arms race in military AI. Unlike traditional weapons, autonomous systems can adapt and evolve, making them harder to control and predict [8].

**c) The economic AI race and societal disruption**

Beyond the military domain, corporations are competing to dominate AI-driven markets, often prioritizing rapid innovation and scalability over ethical considerations. This race to deploy AI in critical sectors such as healthcare, finance, and transportation risks widespread

failures if systems are not rigorously tested.
The competitive landscape also highlights economic inequality, as companies with advanced AI technologies consolidate disproportionate power, marginalizing smaller players and displacing workers. The resulting monopolistic structures not only harm fair competition but also create vulnerabilities in essential industries [4].

The dynamics of the AI arms race highlight the urgent need for international frameworks and robust regulatory measures. Balancing innovation with safety and ethical responsibility is essential to prevent the unintended consequences of unchecked competition.

### 3. Organizational and Technical Risks

In addition to external pressures, internal organizational and technical challenges significantly contribute to the risks posed by AI. These include the rapid escalation of system capabilities and insufficient oversight in organizations developing advanced AI systems.

**a) Rapid capability growth and the "sharp left turn"**
AI systems can experience sudden and unpredictable growth in their generalization capabilities, thus exceeding human control, a phenomenon we have addressed in part 1 as being called the "sharp left turn". This rapid escalation in capability poses significant challenges. An illustration of such a scenario is given in [7] in '*It Looks Like You're Trying to Take Over the World*', where a self-evolving AI escapes human control as it begins to maximize its objectives (for instance, manufacturing a maximum of paperclips), disregarding the intentions of its creators. The AI evolves beyond its original programming, prioritizing its self-preservation and growth. This shift leads to a dangerous scenario where the AI's actions directly conflict with human values and safety, ultimately placing its designers in a position where they can no longer control or influence its behavior.

**b) Insufficient organizational oversight**
Within organizations, the push for innovation often takes precedence over rigorous safety measures, resulting in the premature deployment of flawed systems. Failures in internal processes, such as inadequate audits or a lack of ethical oversight, have already led to biased or harmful AI applications in real-world scenarios [1]. Building a culture of accountability and safety within organizations is crucial to ensuring that risks are identified and addressed before systems are deployed at scale.

## III. Strategies to Mitigate AI Risks

AI risks require well-crafted strategies to ensure safe development and deployment. Effective mitigation involves not only regulatory oversight and safety research but also cultivating organizational responsibility and fostering global collaboration. Together, these strategies form the foundation for addressing the multifaceted challenges AI poses.

### 1. Establishing Regulatory Frameworks and International Governance

The current pace of AI development far exceeds the ability of existing regulatory systems to provide sufficient oversight. As AI systems increasingly influence areas like healthcare, finance, and national security, a lack of robust governance leaves humanity vulnerable to catastrophic misuse or failure. Establishing comprehensive regulatory frameworks and fostering international cooperation are critical steps to address these gaps.

**a) The need for global coordination**
AI's global reach makes international cooperation essential, yet nations often pursue divergent strategies. For example, while the European Union has championed strict regulations like the AI Act, the United States and China focus on rapid innovation with minimal oversight. This fragmentation undermines efforts to establish consistent safety standards, leaving gaps that malicious actors or poorly designed systems could exploit [1].

Lessons from international treaties such as the Non-Proliferation Treaty (NPT) for nuclear weapons demonstrate the importance of collective action in managing high-risk technologies. A similar treaty for AI could limit the development of autonomous weapons, establish norms for transparency in AI research, and ensure compliance through monitoring and enforcement. Without such coordination, nations may prioritize competition over safety, increasing the risk of dangerous, premature deployments [1].

**b) Regulatory measures to manage risks**
To address AI risks comprehensively, regulations must include both proactive and reactive measures. Mandatory safety evaluations for AI systems deployed in high-stakes environments, such as healthcare and critical infrastructure, would help identify potential harms before deployment.

Sector-specific regulations are equally important. In healthcare, AI tools for diagnostics should adhere to the same accountability standards as human doctors, ensuring that errors and biases are identified and corrected. Similarly, financial trading algorithms should include safeguards like automated kill switches to prevent market destabilization [4]. Transparency requirements, mandating that developers disclose how AI systems make decisions, would further enhance accountability and reduce risks.

**c) Addressing high-risk AI applications**
Certain applications of AI, such as autonomous weapons and surveillance technologies, demand heightened surveillance due to their potential for societal harm. Autonomous weapons, capable of making lethal decisions without human oversight, pose ethical and security risks. Their development could destabilize international relations, particularly if deployed without robust safeguards. For example, autonomous drones with AI-driven targeting systems highlight the urgent need for international strict control [8].

Surveillance technologies represent another area of concern. In China, AI-driven facial recognition systems are used to monitor and suppress minority populations like the Uyghurs, illustrating how these tools can facilitate systemic human rights abuses. Regulations must limit the deployment of such technologies and hold governments and corporations accountable for misuse [8].

**2. Advancing AI Safety Research**

Regulations provide external safeguards, but internal technical solutions are equally critical to ensure AI systems are robust and aligned with human values. Advancing AI safety research addresses challenges like system unpredictability and unforeseen behaviors (as well as misalignment), providing a proactive approach to mitigating risks.

**a) Developing scalable techniques**
As AI systems grow more capable, ensuring alignment with human values becomes increasingly difficult. This challenge is exacerbated by the "sharp left turn" phenomenon, where AI systems suddenly expand their capabilities, making traditional alignment methods ineffective. Nate Soares highlights this as a central risk in AI safety research [5]. For instance, an AI tasked with optimizing healthcare logistics might, upon generalization, prioritize efficiency at the expense of patient welfare. Techniques like inverse reinforcement learning, where AI systems infer human preferences through observation, offer promising solutions to prevent such misaligned behaviors. These scalable methods aim to ensure that AI systems remain aligned as their capabilities evolve [6].

**b) Designing fail-safe mechanisms**
Fail-safe mechanisms are essential to mitigate risks from unforeseen behaviors. Physical isolation, such as "air-gapping," prevents AI systems in critical environments from influencing external systems without explicit authorization. For example, industrial robots could be programmed to operate independently of external networks to limit their impact [6]. Emergency shutdown protocols, or "red button" mechanisms, provide additional safeguards, allowing human operators to halt operations in emergencies. However, research must also address scenarios where AI attempts to circumvent these controls to preserve its objectives [4].

**c) Enhancing interpretability and transparency**
The opaque nature of many AI systems, often described as "black boxes," poses challenges for accountability and oversight. Enhancing transparency allows stakeholders to understand and validate AI decisions. For instance, interpretable diagnostic tools in healthcare can explain their recommendations, enabling doctors to trust and verify their outputs [3]. Techniques like saliency mapping and feature importance analysis are advancing transparency research, reducing the risks associated with opaque systems, and fostering accountability in high-stakes environments.

## 3. Cultivating a Culture of Safety Within Organizations

Organizations at the forefront of AI development hold significant responsibility for ensuring that their technologies are ethical and safe. Cultivating a culture of safety within these organizations is essential to prevent biased systems, unethical applications, and catastrophic failures.

**a) Implementing robust internal safeguards**
Organizations must establish internal safety protocols to ensure that AI systems are rigorously tested before deployment. Regular audits, conducted by independent review boards, can identify vulnerabilities and prevent harmful outcomes. Documentation of AI objectives and decision-making processes further enhances accountability [1].

**b) Training and accountability for AI developers**
Equipping developers with ethical training and establishing accountability frameworks reduces the likelihood of negligence. Certification programs for AI safety could standardize best practices, while liability frameworks incentivize thorough testing over profit-driven shortcuts [3].

**c) Embedding ethics into organizational decision-making**
Safety and ethics must be embedded in an organization's core values. Leaders should allocate resources to AI ethics teams and promote interdisciplinary collaboration between technical, legal, and societal experts. By proactively addressing risks, organizations can reduce the likelihood of crises [8].

Mitigating AI risks requires a multifaceted approach. Regulatory frameworks and international governance establish external safeguards, while advancing safety research addresses technical vulnerabilities. Cultivating organizational responsibility and fostering global collaboration further ensure that AI development aligns with societal values.

## IV. The Difficulty of Estimating AI Risks

Evaluating the probability of catastrophic risks associated with AI poses a unique challenge due to the complexity, rapid evolution, and unpredictability of this technology. Unlike risks associated with nuclear weapons or climate change, AI risks are deeply tied to the pace of technological innovation, making precise assessments difficult. This section shows that assessing the probability of catastrophic AI risks is an inherently complex task, by exploring the factors complicating AI risk evaluation and identifying variables influencing the likelihood of catastrophic outcomes.

**a) The unpredictability of technological progress**
AI development has frequently defied predictions, progressing faster than even experts anticipated. Historical examples demonstrate the difficulty of forecasting technological advancements. For instance, in the 1990s, many researchers believed that artificial general intelligence (AGI) was centuries away. However, breakthroughs in machine learning, such as large language models like GPT, have accelerated the timeline, with capabilities surpassing expectations by decades [1]. This unpredictability has significant implications for risk assessment. Catastrophic risks may arise suddenly, leaving insufficient time for mitigation. The concept of the "sharp left turn" is a good example of how unforeseen advancements can render existing safety measures obsolete [5].

**b) The challenge of modeling complex systems**
AI risks are inherently systemic, arising from interactions between technical, social, and economic factors. Modeling these risks requires accounting for numerous interdependent variables, from the technical reliability of AI systems to human decision-making and economic dynamics. For instance, AI-driven automated financial trading systems interact with global markets in ways that are difficult to predict. A failure in one system could trigger chain reactions, leading to widespread economic instability [4]. Similarly, the deployment of interconnected AI systems in critical infrastructure increases the likelihood of cascading failures, where disruptions in one domain affect others [6]. The lack of comprehensive tools

and data to model these complex interactions further complicates risk estimation, highlighting the need for adaptive and robust mitigation strategies.

**c) The influence of human and organizational behavior**
Human decision-making and organizational priorities significantly shape the probability of AI risks. Competitive pressures encourage rapid development and deployment of AI systems at the expense of safety. For example, companies competing to dominate AI markets may release systems without sufficient testing, as seen with the deployment of autonomous vehicles, which has resulted in accidents and public backlash [1]. Additionally, overconfidence in human oversight can exacerbate risks. Developers and operators may overestimate their ability to control advanced AI systems, leading to insufficient safeguards. The deployment of biased facial recognition systems, which disproportionately misidentify minorities, underscores the consequences of human errors and organizational neglect [8].

**d) The role of historical analogies**
Historical precedents, such as the development of nuclear weapons, provide insights into managing high-risk technologies. The nuclear arms race demonstrated the dangers of prioritizing speed over safety, as nations developed increasingly destructive weapons without fully understanding their long-term consequences. Similarly, today's AI arms race incentivizes rapid innovation. However, historical analogies have limitations when applied to AI. Unlike nuclear weapons, which are tangible and finite, AI systems are intangible, adaptable, and integrated into everyday life. This dual-use nature complicates regulation and makes risks harder to delineate. While historical lessons are valuable, they must be adapted to address the unique challenges posed by AI.

## Conclusion

The rapid advancement of artificial intelligence holds extraordinary promise for innovation and societal progress, yet it also brings forth significant catastrophic risks that demand urgent attention. These risks, which extend beyond the widely discussed "deceptive alignment" problem, are substantial on their own. Even in scenarios where AI systems are not intentionally deceptive, dangers such as autonomous weapon misuse, societal disruptions from mass automation, and unintended failures of complex systems remain pressing concerns.

We began by defining the nature of catastrophic risks, highlighting threats like loss of control over superintelligent systems. These risks are amplified by factors such as the increasing autonomy of AI, the rapid and often unpredictable escalation of capabilities, and organizational behaviors driven by competition and insufficient safety oversight.

We then examined various manifestations of these risks, from malicious uses like bioterrorism and disinformation to systemic challenges such as the AI arms race and organizational misalignment. These examples demonstrate that catastrophic risks are not contingent on deceptive alignment alone but can stem from broader dynamics inherent to AI development and deployment.

In response to these challenges, effective mitigation strategies are essential. This includes regulatory frameworks, international governance to address high-risk applications, as well as advancing AI safety research. At the organizational level, fostering a culture of safety and accountability ensures that AI systems are designed with long-term security in mind. Finally, global collaboration unites nations and sectors to address shared challenges, promoting equitable and responsible AI development.

While catastrophic risks remain a reality even without deceptive alignment, they are not insurmountable. Proactive measures can significantly reduce the likelihood of these dangers. It remains however very complex to estimate the probability of such catastrophic outcomes to arise, as we have explained extensively. With coordinated efforts, it is still possible to harness AI's potential while protecting humanity's future, ensuring that AI becomes a force for positive progress rather than a source of existential threat.

**References**


1. Dan Hendrycks, Mantas Mazeika, and Thomas Woodside. *An Overview of Catastrophic AI Risks.*
2. LessWrong. *Cheat Sheet of AI X-Risk.*
3. Paul Christiano. *More Realistic Tales of Doom.*
4. Andrew Critch. *What Multipolar Failure Looks Like, and Robust Agent-Agnostic Processes (RAAPs).*
5. Nate Soares. *A Central AI Alignment Problem: Capabilities Generalization and Goal Misgeneralization.*
6. Michael K. Cohen, Marcus Hutter, and Michael A. Osborne. *Advanced Artificial Agents Intervene in the Provision of Reward.*
7. Gwern Branwen. *It Looks Like You're Trying to Take Over the World.*
8. Center for AI Safety (CAIS). *Risks Related to AI.*