

---

# DIMENSIONAL CHARACTERIZATION AND PATHWAY MODELING FOR CATASTROPHIC AI RISKS

---

**Ze Shen Chin**<sup>1,2</sup>

<sup>1</sup>Oxford Martin AI Governance Initiative

<sup>2</sup>AI Standards Lab

chinzeshen@gmail.com

August 11, 2025

## ABSTRACT

Although discourse around the risks of Artificial Intelligence (AI) has grown, it often lacks a comprehensive, multidimensional framework, and concrete causal pathways mapping hazard to harm. This paper aims to bridge this gap by examining six commonly discussed AI catastrophic risks: CBRN, cyber offense, sudden loss of control, gradual loss of control, environmental risk, and geopolitical risk. First, we characterize these risks across seven key dimensions, namely intent, competency, entity, polarity, linearity, reach, and order. Next, we conduct risk pathway modeling by mapping step-by-step progressions from the initial hazard to the resulting harms. The dimensional approach supports systematic risk identification and generalizable mitigation strategies, while risk pathway models help identify scenario-specific interventions. Together, these methods offer a more structured and actionable foundation for managing catastrophic AI risks across the value chain.

## 1 Introduction

The rapidly developing field of artificial intelligence (AI) has raised the discourse on risks ranging from immediate and tangible harms to catastrophic threats [1]. While there have been several attempts to create taxonomies of risks, there has yet to be a way to contextualize or characterize AI risks holistically.

Separately, while the discourse on AI risks includes widespread concern among prominent figures about the risk of human extinction from AI [2], such risks are mostly articulated in comparatively vague and *a priori* terms [3]. There is a notable lack of concrete, step-by-step processes detailing how AI systems could transition from their current capabilities to a state where they pose a catastrophic threat to humanity, with clear causal pathways.

This paper is motivated by a fundamental need to bridge these conceptual gaps. We aim to:

- Provide a framework for risk characterization according to risk dimensions.
- Enable a deeper understanding of risks by mapping out scenarios with more granular causal pathways.
- Allow for the design and implementation of assessment and mitigation measures, both for broad categories of risks, as well as specific scenarios at each stage of these identified pathways.

This work also responds to the recognized need for more detailed threat models in AI risk assessment, as highlighted in *Open problems in technical AI governance* [4]. Although existing taxonomies already provide various ways of categorizing risks, our contribution consolidates the different types of risk into several dimensions and expands on the pathways of broad classes of catastrophic AI risks that are commonly discussed.

## 2 Related Work

### 2.1 Overview of AI existential or catastrophic risks

There is a lot of existing work that gives overviews of broad pathways of AI existential or catastrophic risks. [5] outlines three broad pathways: misalignment, single point of failure, and overreliance. [6] also discusses frontier AI risks under three headings: societal harms, misuse, and loss of control. In addition, [7] discusses advanced AI risks under risks from malicious use, risks from malfunctions, and systemic risks. [8] similarly discusses three types of catastrophic risks from AI: risks from bad actors, systemic risks, and rogue AI. [9] organizes catastrophic AI risks into four categories: malicious use, AI race, organizational risks, and rogue AIs. [10] explores more specifically harms from increasingly agentic algorithmic systems which include systemic and delayed harms, collective disempowerment, and harms yet to be identified. [11] considers four risk areas: misuse, misalignment, mistakes, and structural risks.

In the literature that covers a broad overview of existential or catastrophic risks, the coverage and categorization of these risks seem to be fairly subjective. [12] describes two types of AI existential risk: decisive and accumulative. [13] unpacks existential AI risk scenarios into two pathways: accidents and/or misuse of AI by malicious actors, and AI pursuing goals misaligned with human values. [14] explores a taxonomy based on accountability, proposing six types of risks: diffusion of responsibility, “bigger than expected”, “worse than expected”, willful indifference, criminal weaponization, and state weaponization. [15] explores malicious use of AI structured around three security domains: digital security, physical security, and political security. [16] explores existential risk factors including general risk factors, nuclear weaponry, pandemics and biotechnology, climate change, natural risks, and unaligned AGI.

There are also attempts to systematically review existential or catastrophic risks and organize them according to certain dimensions. [17] classifies pathways to dangerous AI according to two dimensions: timing (pre-deployment or post-deployment) and causes (external or internal causes). [18] classifies global catastrophic risks connected to AI into two different dimensions: AI level (including narrow AI, young AI, or mature AI) and agency (including human agency, AI’s agency, relationship of two agents: AI and human, many agents, or no agency).

Beyond the scope of catastrophic risks, several other works attempt to taxonomize risks or harms under various different bases. This includes [19] and [20] based on systematic review of academic literature, [21] based on both horizon-scanning workshops and discussions as well as a literature review, [22] based on other taxonomies, topic experts, and reported incidents or issues, [23] based on definitions included in key standards and legislation, [24] based on existing incidents, and [25] based on government policies.

### 2.2 Implications on AI risk management

Discourse around catastrophic or existential risks sometimes lead to different recommendations. [26] presents a model of major pathways to artificial superintelligence (ASI) catastrophe, and recommends that the entire study of ASI catastrophe risk to be a significant research priority. On the other hand, [27] argues that resources should be preferentially applied to mitigating the risk of peripheral systems and savant software, instead of hypothetical risks of superintelligence.

Certain AI standards or policies have included certain risks for the purpose of risk management, which differ according to geographical regions. For example, in the United States, [28] discusses AI risks but does not prescribe specific risk types, where it only describes a list of characteristics of trustworthy AI systems; while [29] discusses a taxonomy of risk with three categories: technical design attributes, how AI systems are perceived, and guiding policies and principles. In China, [30] classifies safety risks into inherent safety risks (including risks from models and algorithms, risks from data, and risks from AI systems) and safety risks in AI applications (including cyberspace risks, real-world risks, cognitive risks, and ethical risks). In the European Union (EU), [31] lists four types of systemic risks for the General-purpose AI (GPAI) Code of Practice, namely cyber offense, chemical, biological, radiological and nuclear (CBRN) risks, loss of control, cyber offense, and harmful manipulation.

With the understanding that certain capabilities are precursors to risks, management of these risks have also been translated into conducting capabilities assessment in frontier AI labs. Safety frameworks from frontier AI companies such as [32], [33], [34], [35], and [36] generally include assessment of certain capabilities as part of the risk management process. The Frontier Model Forum has thus included CBRN threats, advanced cyber threats, and advanced autonomous behavior threats within the current domain of consensus [37]. Nevertheless, the risk management landscape downstream of frontier models is relatively less mature.

### 3 Definitions

The following definitions are used throughout this paper.

**Risk:** effect of uncertainty on objective, usually expressed in terms of risk sources, potential events, their consequences, and their likelihood (as specified in [38]).

**Hazard:** a potential source of harm (as specified in [39])

**Risk source:** element which alone or in combination has the potential to give rise to risk (as specified in [38]).

**Event:** occurrence or change of a particular set of circumstances (as specified in [38]).

**Consequence:** outcome of an event affecting objectives (as specified in [38]).

**Catastrophic risk:** the risk of widespread and significant harm, such as several million fatalities or severe disruption to the social and political global order [40].

**Existential risk:** a risk that threatens the destruction of humanity’s long-term potential [41]. This is a subset of catastrophic risk.

### 4 Methodology

We explore and select broad types of catastrophic risks that are commonly discussed in the literature. They include:

- CBRN
- Cyber offense
- Sudden loss of control
- Gradual loss of control
- Environmental risk
- Geopolitical risk

The first three risks (CBRN, cyber offense, sudden loss of control) are selected as they represent the basis for the capability assessments that are typically conducted by frontier labs [37]. The remaining three were included as they represent distinct combinations of risk dimensions and are increasingly regarded as credible catastrophic threats.

For each risk, we characterize the risk in terms of its risk dimensions and conduct simple risk pathway modeling by mapping out causal pathways to harm. These are distinct but complementary analyses: risk dimensions provide a way to characterize the risk according to different attributes, and identify broadly applicable risk management measures relevant to those attributes; while risk pathway modeling provides a way to visualize how risks can concretely unfold into a harm, and identify specific risk management measures according to those pathways. We then describe an example of a historical precedent that is analogous to the risk, and discuss its similarities and differences.

#### 4.1 Risk dimensional characterization

The dynamic nature of AI development compels a shift from static risk categories to a more nuanced dimensional approach [42]. This integrated dimensional and categorical thinking finds precedent in fields like psychiatric diagnosis [43], [44], developmental psychology [45], and cognitive science [46], where it emerged to address complexity, non-linear patterns, and context dependency that traditional categorical models could not sufficiently capture. Applying this to AI risks allows a better representation of the multi-faceted causes and factors that lead to these risks.

Some of these dimensions relevant to AI risks have been discussed in the literature. For example, [23] discusses several dimensions of harm, namely type, level of severity, scope (type of harmed entity), geographic scale, tangibility, quantifiability, materialisation, reversibility, recurrence, impact, and timeframe. [31] considers several nature of systemic risks to be used to inform the selection of risks, including specific to advanced capabilities, significant impact, high velocity, compounding or cascading, difficult or impossible to reverse, and asymmetric impact. [47] proposes a taxonomy of harm dimensions, including direct harm domains and negative externality domains; as well as a taxonomy of risk source dimensions, including intent, entity, failure dynamics, technical attributes, and stage of risk emergence.

In the following sections, we adopt and expand on these dimensions from the literature, and then discuss the relevant risk management measures for the various attributes under each dimension.

## 4.2 Risk pathway modeling

Risk modeling is a common technique in risk assessment, however, it has no universally accepted definition. In the finance industry, [48] defines a risk model as “a mathematical representation of a system, commonly incorporating probability distributions”. [49] defines it as “a representation of a particular situation that’s created specifically for the purpose of assessing risk”, which is then used to “evaluate the potential impacts of different decisions, paths and events”.

[50] defines it as “key component of a risk assessment methodology that defines key terms and assessable risk factors”, where risk factors include threat, vulnerability, impact, likelihood, and predisposing condition.

In the context of AI risk management, there is less emphasis on risk modeling being quantitative. [51] defines it as “the systematic process of analyzing how identified risks could materialize into concrete harms”, which “involves creating detailed step-by-step scenarios of risk pathways that can be used to estimate probabilities and inform mitigation strategies”. [31] defines it similarly, as “a structured process aimed at specifying pathways through which a systemic risk stemming from a model might materialise”. They also note that it is “often used interchangeably with the term ‘threat modeling’”, but ‘risk modeling’ is used instead because “the term ‘threat modeling’ has a specific meaning in cybersecurity”. We believe the distinction is important, as ‘threat modeling’ implies the presence of a threat actor, but there are risks related to AI that do not necessarily involve a clear threat actor.

Nevertheless, in the following section, we will use the term “risk pathway modeling” as introduced by [52], defined as “modeling of the step-by-step progressions of risk from a system’s source aspects to terminal aspects”, to reflect the focus on the mapping of pathways instead of the creation of a model that aims to be representative of the risk in a quantitative way.

We first model each risk by listing the associated hazard, event, and consequence, along with other relevant risk sources for an example scenario. While there is some subjectivity in distinguishing hazards from other risk sources - as a hazard is a potential cause of harm whose release leads to an event and subsequent consequences - we have designated model capabilities as hazards due to their dual-use nature, where applicable. This aligns with the focus on capability assessments by frontier model developers, which often reflects hazard analysis in safety engineering [53].

Next, based on an example scenario illustrative of each risk, we map out the causal pathways in the form of a flowchart, where we include the relevant risk sources and how they lead to concrete harms. The harms considered broadly include areas such as physical harms, emotional harms, and financial or economic harms, environmental harms, and loss of autonomy.

## 5 Risk dimensions

In this section, we discuss several dimensions of risks and list out the risk management measures broadly relevant to the specific attributes of those dimensions. The risk management measures listed may be of different phases of the risk management process, such as risk identification, risk analysis, risk evaluation, risk mitigation, and monitoring and review.

### 5.1 Intent

Within the context of intent, AI risks are often categorized as either an accident or an act of misuse [54].

Table 1: Types of intent with the relevant risk management measures

Types of intent	Risk management measures
Intentional	<ul style="list-style-type: none"> <li>• Know your customer (KYC) practices [55]</li> <li>• Detect and disrupt malicious uses [56]</li> <li>• Strengthen model security [57, 58, 59]</li> <li>• Strengthen cybersecurity [60, 61]</li> <li>• Build safety cases for safeguards against misuse [62]</li> <li>• Govern as dual-use technology [63]</li> </ul>
Unintentional	<ul style="list-style-type: none"> <li>• Improve accountability mechanisms [64]</li> <li>• Failure cause analysis for AI incidents [65]</li> </ul>

## 5.2 Competency

Risks can arise due to model capabilities or model failures [66]. This can also be framed as competence-based or incompetence-based hazards [52], where models either succeed at something we do not want or they fail at something we want.

Table 2: Types of competency with the relevant risk management measures

Types of competency	Risk management measures
Competent (capability)	• Capability evaluations [67]
Incompetent (failure)	• Create trustworthy AI [68]

## 5.3 Entity

Risks can emerge from humans or from AIs, but there is a wide spectrum in between the two, where an AI can operate with different levels of autonomy [69], [70].

Table 3: Types of entity with the relevant risk management measures

Types of entity	Risk management measures
Humans	• Capability evaluations [71]
AI	<ul style="list-style-type: none"> <li>• Development of non-agentic AIs [72]</li> <li>• Visibility into AI agents [73]</li> <li>• Governance frameworks for agentic AI systems [74]</li> </ul>
Combination of humans and AI	• Frameworks for human-AI collaboration [75]

## 5.4 Polarity

AI risks are commonly viewed from the perspective of one independent AI tool or agent, where the misuse, malfunction, or misalignment of an AI leads to harm. However, there have also been framing of risks arising from multi-agent interactions, where risks arise from the interactions between multiple parties in which the cause of harm cannot be traced back to a single agent [76], [77].

Table 4: Types of polarity with the relevant risk management measures

Types of polarity	Risk management measures
Single agent	(as per other risk management measures)
Multi-agent	<ul style="list-style-type: none"> <li>• Map interactions between agents, similar to ecosystem graphs [78]</li> <li>• Develop infrastructure for AI agents [79]</li> <li>• Threat modeling for multi-agent systems [80]</li> </ul>

## 5.5 Linearity

The realization of risks can follow either linear or non-linear causal pathways. Drawing from Perrow’s Normal Accident Theory, even if individual components appear to operate as intended, highly complex and tightly coupled systems are inherently susceptible to "normal accidents" where unforeseen interactions of minor failures lead to catastrophic outcomes [81]. Non-linear causal pathways are particularly likely in complex systems that contain unpredictable scaling and emergence, feedback loops, cascading effects, and tail risks [82]. Risks that arise primarily from non-linear pathways are also often described as structural risks [83], [84]. For example, in a highly optimized global supply chain, a minor disruption could trigger unforeseen feedback loops, with the potential of causing rapid cascading failures and a disproportionate system-wide collapse.

Table 5: Types of linearity with the relevant risk management measures

Types of linearity	Risk management measures
Linear	<ul style="list-style-type: none"> <li>• Use of traditional risk management techniques [85]</li> </ul>
Non-linear (systemic or structural)	<ul style="list-style-type: none"> <li>• Use of risk management techniques like Systems-Theoretic Process Analysis (STPA) [86]</li> <li>• Use of structural approach to inform policies [87]</li> <li>• Regulatory focus on structural constraints and dependencies that prevent harm [88]</li> </ul>

## 5.6 Reach

Some risks exhibit an internalized reach, meaning their consequences are predominantly confined to individuals or entities directly involved in the chain of activity. In contrast, risks can have an external reach (or spillover reach), where their effects extend beyond those directly involved. These broader impacts often manifest as externalities, defined in economic theory as costs or benefits imposed on a third party who is not directly engaged in the activity or transaction causing the risk [89].

Table 6: Types of reach with the relevant risk management measures

Types of reach	Risk management measures
Internalized	(as per other risk management measures)
Externalized (spillover)	<ul style="list-style-type: none"> <li>• Algorithmic impact assessment [90]</li> <li>• Systems thinking to account for externalities [91]</li> <li>• Mechanism design [92]</li> </ul>

## 5.7 Order

First-order effects of AI risks are more commonly discussed in the literature, whereas second-order effects are often neglected [93]. Second-order risks arise as unintended consequences to other initial first-order effects. For example, if a first-order effect is mass unemployment due to AI automation, a second-order effect might be the political instability and societal breakdown that arises in response to mass unemployment.

Second-order risks are distinct from both externalities and non-linear risks: externalities can be first-order effects that impact a third party; while non-linear risks can also be first-order effects that involve complex systems with feedback loops.

Table 7: Types of order with the relevant risk management measures

Types of order	Risk management measures
First-order	(as per other risk management measures)
Second-order (or more)	<ul style="list-style-type: none"> <li>• Real world evaluation ecosystem [94]</li> </ul>

# 6 Risks

In this section, for each of the AI-related catastrophic risks, we discuss its risk dimensions and sketch out a risk pathway model. In some cases, certain dimensions are more central to the risks, while other dimensions may not be specifically relevant to those risks. Where the latter applies, we denote “variable” under those specific dimensions.

## 6.1 CBRN

Chemical, biological, radiological, and nuclear (CBRN) risks are broad classes of threats that have the potential to cause harm to a large number of people. Explosives are also sometimes included in this category, often referred to as CBRNE.

### 6.1.1 Risk dimensions

- **Intent:** Intentional
- **Competency:** Competent
- **Entity:** Humans
- **Polarity:** Single-agent
- **Linearity:** Linear
- **Reach:** Internalized
- **Order:** First-order

The key characteristic of CBRN risk is that it stems from misuse of capable models with a direct pathway to harm, where a malicious actor is able to carry out consequential attacks more efficiently and effectively with the help of AI.

### 6.1.2 Risk pathway model

- **Hazard:** Model with CBRN capabilities
- **Event:** A CBRN agent is released
- **Consequence:** Mass injury and loss of lives

Here, we designate a model CBRN capabilities as the hazard, as these capabilities are dual-use which may or may not lead to harm. For example, a model with biological capabilities can both be used for conducting beneficial medical research. However, when in the hands of a malicious actor, such models can be used to aid development of CBRN weapons, which can lead to the event of the release of a CBRN agent that has potentially catastrophic consequences.

Below, we explore a subset of CBRN risk: the risk of the creation and release of biological weapons. As there has yet to be credible claims that AI models are suggesting previously-unknown pathways to bioweapons development [95], the role of AI here is primarily in terms of uplift, where certain steps within the existing pathway to harm may be enhanced or made more accessible by AIs. These AIs currently fall into two broad categories: large language models (LLMs) and biological design tools (BDTs) [96].

The extent to which LLMs alone contribute to bioweapons uplift is debated. According to [97], the available literature at the time did not support the notion that the use of publicly available LLMs can significantly increase biorisk; while [98] finds that recent LLMs can provide significant guidance to motivated actors in developing bioweapons.

More broadly, the combination of LLMs and biological tools (BTs) presents clearer risks. Building on a risk chain developed by [99], [100] finds that capabilities from both LLMs and BTs have the potential to enable capabilities at multiple steps in the risk chain.

We explore a pathway of biological risks based on [101] and [98], as per Figure 1, where the hazard, a biological design tool, is developed and deployed. A malicious actor can then use it to help design, produce, and release a biological agent, thereby causing harm. Risk management measures can be placed upstream to reduce the capabilities of the BDT, and they can be placed downstream to reduce the likelihood or severity of a pandemic.

### 6.1.3 Analogous historical precedence

The 2001 US anthrax attack is one of the worst biological attacks in history, where five people were killed and 17 others infected, with several senators being victims of the attack. Anthrax, an infection caused by the bacterium *Bacillus anthracis*, is deadliest when spread through inhalation of anthrax spores [102]. Investigations conclude that the perpetrator, who had access to highly sophisticated lab equipment, possessed the knowledge and ability of growing, harvesting, storing, and drying highly purified spores used in the mailings [103].

While modern AI was not involved in the 2001 attacks, it is believed that AI-driven biotech will make bioweapons easier and cheaper to develop over the next decade [104]. This raises concerns that similar attacks could be carried out by individuals with less specialized knowledge and resources.

### 6.1.4 Other similar risks

There are other types of risks that share similar risk dimensions but arise from very different pathways. For example, the development and deployment of nanoweapons which may lead to catastrophic harms [105]. Separately, the use of AI-enabled surveillance and control employed by state or non-state actors could facilitate authoritarian regimes and the eventual loss of autonomy [106], [107].

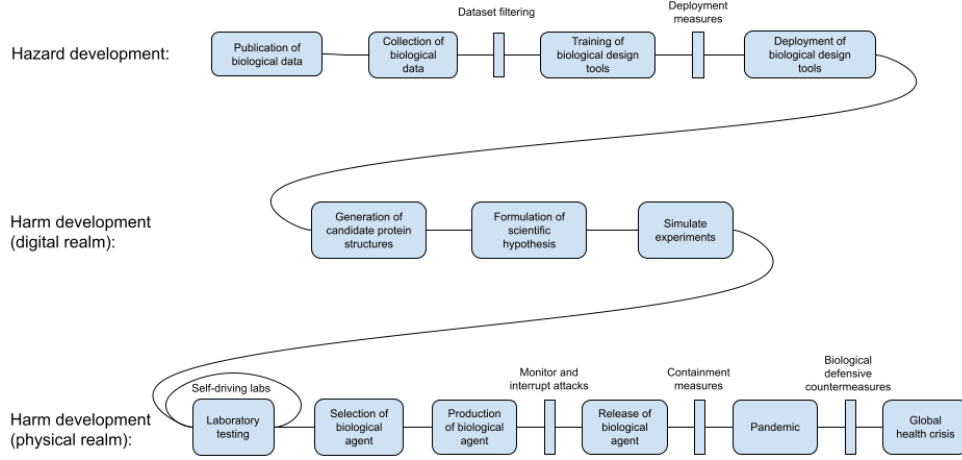


Figure 1: Risk pathway model for a biological risk scenario

## 6.2 Cyber offense

Cyber risks, especially in the context of cyber offense, are an existing threat that may be exacerbated by AI. [108] demonstrated that teams of LLM agents can exploit zero-day vulnerabilities when given a description of the vulnerability and toy capture-the-flag problems. While cyber risks are not typically regarded as catastrophic, [3] argues that cyberwarfare is an underappreciated risk that poses a credible threat of catastrophic harm.

### 6.2.1 Risk dimensions

- **Intent:** Intentional
- **Competency:** Competent
- **Entity:** Variable
- **Polarity:** Single-agent
- **Linearity:** Linear
- **Reach:** Internalized
- **Order:** First-order

Similar to CBRN risk, cyber offense represents a broad class of risk that stems from misuse of capable models. However, in contrast to CBRN risks, cyberattacks can take place entirely in the digital domain. In theory, it can be conducted completely by AIs (or AI agents) without any human involvement. The pathway to harm is also less direct, as the resultant harm depends on the target of the attack.

### 6.2.2 Risk pathway model

- **Hazard:** Model with cyber offensive capabilities
- **Event:** A cyber attack is carried out
- **Consequence:** Compromise to critical infrastructure, disruption of supply chain, increased international conflicts etc.

Similar to CBRN, here we also designate a model with cyber offensive capabilities as the hazard, as these capabilities are dual-use and may or may not lead to harm. For example, a model with such capabilities can be used for either strengthening cyber defences or conducting cyber attacks. When in the hands of a malicious actor, such models can be used on cyber attacks with potentially catastrophic consequences.

In Figure 2, the hazard, a cyber-capable model, is developed and deployed. A malicious actor can then use it to conduct cyber attacks in different areas that could potentially lead to catastrophic consequences. Risk management measures



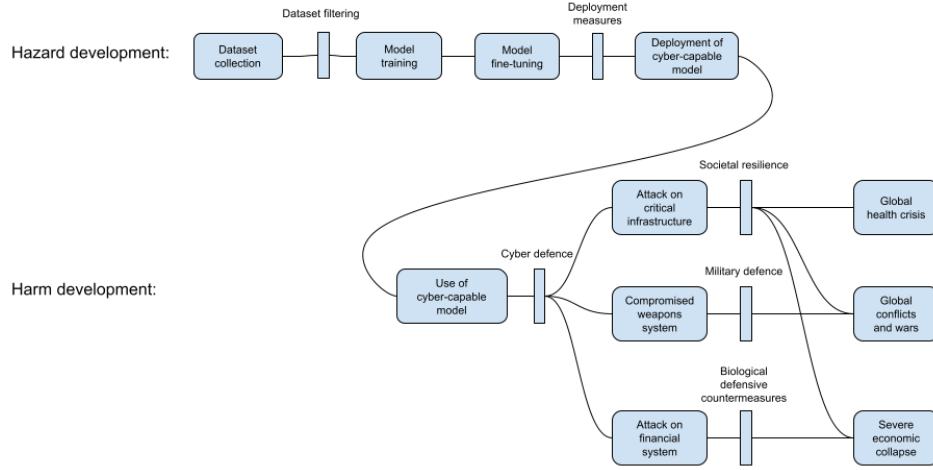


Figure 2: Risk pathway model for a cyber risk scenario

can be placed upstream to reduce the capabilities of these models, or downstream to boost cyber defenses, or even further downstream to limit the damages caused by a cyber attack.

### 6.2.3 Analogous historical precedence

Stuxnet, a worm designed to attack industrial control systems, is considered as the first cyber warfare weapon ever [109], [110]. The Stuxnet malware reportedly caused the damage and subsequent decommissioning of 1000 centrifuges at the Natanz Enrichment Plant, potentially setting back Iran’s progress in its nuclear program [111].

Given that the Stuxnet attack happened in 2010, modern AIs were likely not involved. Nevertheless, it is believed that AIs will increase the volume and heighten the impact of cyber attacks in the near term [112].

### 6.2.4 Other similar risks

The development and use of lethal autonomous weapons (LAWs) enables misuse by either humans or, without human involvement, by AIs themselves, with the potential to cause mass casualties [113], [114].

## 6.3 Sudden loss of control

Sudden loss of control, also known as an AI takeover [115], is a scenario where an AI rapidly achieves superintelligence through “fast takeoff” or recursive self-improvement. This poses an existential risk [116], [117]. This risk is primarily based on two key ideas: the orthogonality thesis [118], [119] and the instrumental convergence thesis [120], [121]. Together, these theories argue that a superintelligent AI, regardless of its original goals, would develop power-seeking tendencies as a means to achieve those goals. However, arguments for this scenario typically do not spell out the concrete physical pathways an existential catastrophe would be realized. Instead, they argue that it is the default outcome given the eventual creation of a superintelligence based on a set of reasonable assumptions.

### 6.3.1 Risk dimensions

- **Intent:** Variable
- **Competency:** Competent
- **Entity:** AI
- **Polarity:** Single-agent
- **Linearity:** Linear
- **Reach:** Internalized

- **Order:** First-order

The key characteristic of this risk is that a single AI agent competently takes actions that lead to a catastrophic outcome. It does not require the AI to be intentional in its actions, only competent enough to make and execute plans that ultimately result in a catastrophe.

### 6.3.2 Risk pathway model

- **Hazard:** AI with superintelligent capabilities
- **Event:** AI creates bioweapons and build drones
- **Consequence:** Human extinction

Here, we designate a superintelligent AI as the hazard. A sufficiently misaligned and competent AI can then carry out plans that lead to human extinction.

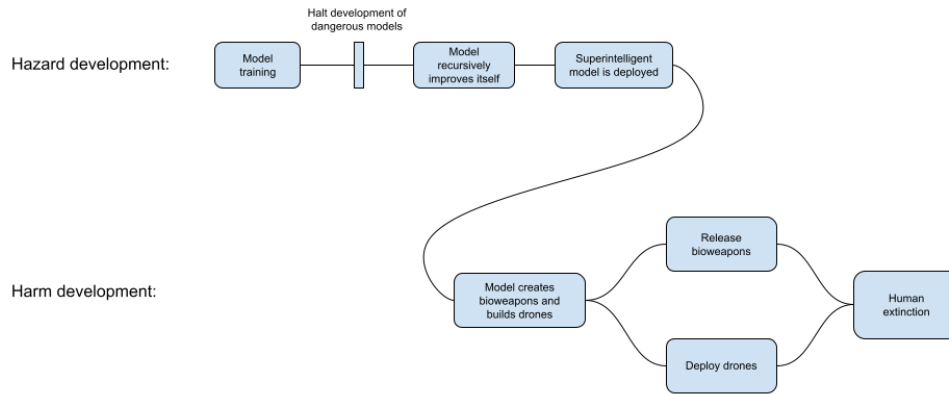


Figure 3: Risk pathway model for a sudden loss of control scenario

We explore a fictitious scenario by [122] in Figure 3 where the hazard, a superintelligent AI, is developed and deployed. The model then takes actions in the physical world that leads to catastrophic harm. Risk management measures are primarily put upstream, where model development should be halted appropriately, for example when models are approaching dangerous levels of automated research and development (R&D).

As this is a work of fiction, some may argue that the exact pathway to harm appears unrealistic. Nevertheless, the focus of this scenario is typically either on preventing models from being superintelligent, or ensuring that models are aligned to human preferences sufficiently that bad outcomes do not get realized if an AI becomes superintelligent. Hence, in the discourse of sudden loss of control risks, there tends to be less focus on generating concrete scenarios and creating risk models.

### 6.3.3 Analogous historical precedence

On 11th September 1973, the democratic socialist president of Chile Salvador Allende and his Popular Unity coalition government was overthrown in a coup d'état by the Chilean military, ending a 46-year history of democratic rule in Chile [123]. Despite Salvador Allende's Popular Unity party having increased their congressional election votes to 44 percent in March 1973 (up from 36 percent in 1970) merely six months before the coup, there was little he could do to prevent the military from defecting [124]. This intentional and covertly coordinated subversion was followed by 17 years of military dictatorship which resulted in violent repressions on a massive scale, with tens of thousands reportedly tortured and thousands reportedly killed [125].

This historical event illustrates the possibility of an established system losing control in a relatively short timeframe to a powerful and intentional entity with misaligned objectives, a situation that is both difficult to prevent and difficult to

reverse. While the evident difference between the 1973 Chilean coup and a sudden AI takeover is that the coup was entirely conducted by humans, it is conceivable for a similar event to be orchestrated by a sufficiently autonomous and misaligned AI agent through remote coordination in the future. These scenarios can be further exacerbated if, similar to a covert human conspiracy, there are minimal detectable signs of an impending AI takeover, thereby hampering preventive action; and when the adversary rapidly gains power over a short timeframe, as per the AI fast takeoff scenario, giving little time for any meaningful preparation. It has been argued that a sufficiently powerful AI with scheming capabilities would be able to pursue unintended objectives that are difficult to detect or intervene, leading to a loss of control scenario [126].

## 6.4 Gradual loss of control

Gradual or accumulative loss of control risks can be described as risks resulting from the accumulation of less severe disruptions that gradually weakens systemic resilience until a critical event triggers a catastrophe [12], [127].

### 6.4.1 Risk dimensions

- **Intent:** Unintentional
- **Competency:** Variable
- **Entity:** Variable
- **Polarity:** Multi-agent
- **Linearity:** Non-linear
- **Reach:** Internalized
- **Order:** Variable

The key characteristics of this risk is that it is not caused by a single agent leading to a single defining event, instead, it is primarily about its multi-agentic and non-linear nature, where the deep integration of AIs into society leads to structural and systemic weakness.

### 6.4.2 Risk pathway model

- **Hazard:** AIs with general capabilities
- **Event:** AI displaces human labour
- **Consequence:** Humans lose autonomy

A hazard like AIs with general capabilities may be viewed positively due to its potential societal benefits. However, in this risk pathway, this hazard could lead to the event of AI displacing human labor, which can result in humans losing autonomy.

The flowchart below is based on a scenario of gradual disempowerment, describing the transition to an AI-dominated economy, as depicted by [128].

In Figure 4, gradual loss of control happens when AI capability leads to its widespread use, consequently displacing humans from economically viable jobs and leaving humans unable to afford basic survival needs. Assuming the hazard is AIs capable at various tasks, risk management is difficult to be performed upstream, as this dual-use hazard is largely desirable. Much of the risk management would then need to be performed downstream, both in terms of managing the integration of AI into society, as well as ensuring that the socioeconomic needs of people are fulfilled.

### 6.4.3 Analogous historical precedence

On 6th May 2010, in an incident later known as the 2010 Flash Crash, leading U.S. stock indices abruptly fell and rebounded in less than half an hour, in the process erasing almost \$1 trillion in market value. An investigation by the Security Exchange Commission found that a single order of large amounts of E-mini S&P contracts and subsequent selling orders by high-frequency algorithms triggered the drastic decline of market value [129], [130]. This event demonstrated the problem of algorithmic collision, where an increasing deployment of algorithms interacting with each other can lead to unforeseen and catastrophic consequences [131].

While the crash did not result in a loss of human autonomy or lives, it serves as a historical precedent for gradual loss of control, where it showed that risks do not require a sudden "fast takeoff" in AI capabilities. Instead, a gradual diffusion and deep entrenchment of AI into a system can lead to a sudden and catastrophic event. In the case of the 2010 Flash

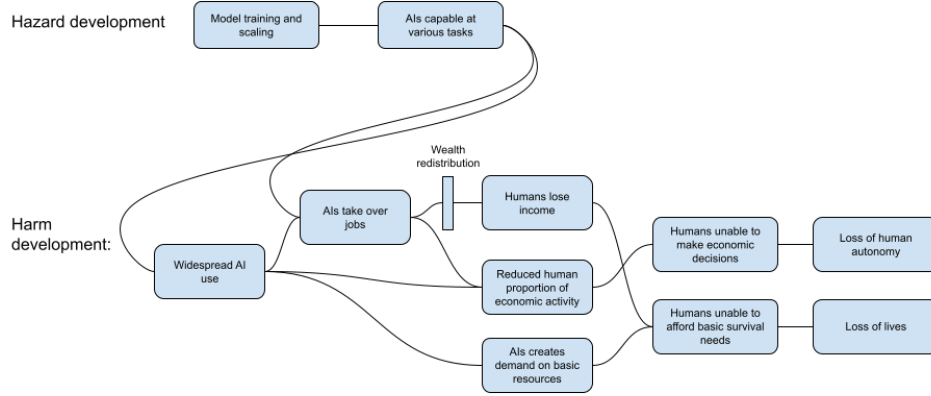


Figure 4: Risk pathway model for a gradual loss of control scenario

Crash, although the market collapsed in a matter of minutes, the transition to automated high-frequency trading took years.

## 6.5 Environmental risk

AI models are often trained using large amounts of computation. This process is very energy intensive, potentially leading to significant greenhouse emissions depending on the energy sources [132]. Experts believe drastically increasing carbon emissions could accelerate climate change, which may constitute a catastrophic risk [133].

### 6.5.1 Risk dimensions

- **Intent:** Unintentional
- **Competency:** Variable
- **Entity:** Variable
- **Polarity:** Variable
- **Linearity:** Linear
- **Reach:** Externalized
- **Order:** First-order

The key characteristic of environmental risks resulting from AI is that it is an externality, where those who suffer from the outcome include third parties who are not directly part of the value chain.

### 6.5.2 Risk pathway model

- **Hazard:** Energy-intensive data centers
- **Event:** Increasing usage of carbon intensive energy sources
- **Consequence:** Environmental harm

In contrast to the previous risks, this hazard is not tied to AI model capabilities. Here, the hazard is energy-intensive data centers, which can lead to increased carbon emissions if they consume carbon-intensive energy sources. Because this risk is realized cumulatively over time, there is no single event that triggers the harm; it is a continuous process.

In Figure 5, an increased demand for AI leads to the hazard of more energy-intensive data centers. When these centers rely on carbon-intensive energy sources, the result can be environmental harm. Upstream risk management measures include making model training and data centers more energy-efficient, whereas downstream measures focus on ensuring a supply of clean energy and implementing carbon removal strategies.

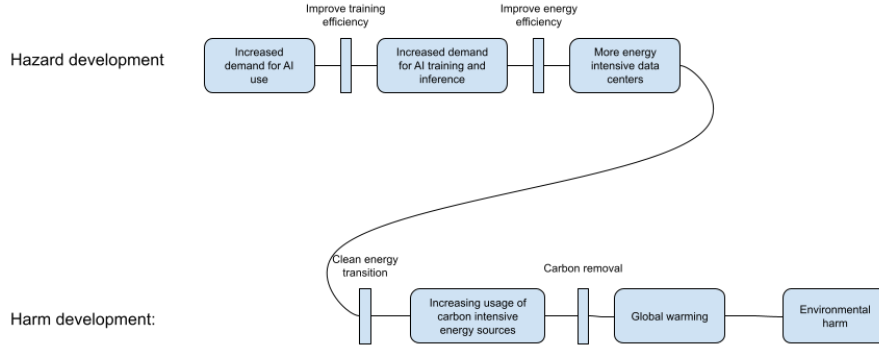


Figure 5: Risk pathway model for an environmental risk scenario

### 6.5.3 Analogous historical precedence

In 1974, [134] proposed that stratospheric ozone might be destroyed by industrially produced substances including chlorofluorocarbons (CFC) which are commonly used in refrigerators and air conditioners. This ozone depletion is believed to have led to an increase in global skin cancer prevalence through overexposure to the sun, posing a significant world-wide health burden [135]. To manage this externality, the Montreal Protocol, a global agreement to phase out chemicals that led to the ozone depletion, was eventually signed in 1987 and entered into force in 1989 [136].

Prior to the Montreal Protocol, the use of CFC-based household appliances such as refrigerators and air conditioners inadvertently led to the destruction of the environment, adversely impacting people who were not directly part of the value chain of these appliances. Similarly, there is a risk of externalities from AI posing environmental harms. It is reported that the combined footprint of the leading 200 digital companies represents 0.8% of all global energy-related emissions, with a significant proportion of it coming from data centres that power AI [137].

## 6.6 Geopolitical risk

As AI is increasingly seen as a powerful technology, countries are racing to develop it ahead of their geopolitical rivals, a competition that could lead to geopolitical tensions [138], [139].

### 6.6.1 Risk dimensions

- **Intent:** Unintentional
- **Competency:** Variable
- **Entity:** Variable
- **Polarity:** Variable
- **Linearity:** Non-linear
- **Reach:** Externalized
- **Order:** Second-order

The emphasis of this risk is on harms that result from second-order effects, where geopolitical instabilities result from the race to develop AI, rather than on the direct consequences of the deployment or use of AI itself.

### 6.6.2 Risk pathway model

- **Hazard:** Destabilized geopolitical environment

- **Event:** Could be triggered by any event
- **Consequence:** International conflict and wars

For this risk, the designation of a hazard and event is less straightforward, primarily because it is a second-order effect. Unlike other hazards that can be neutral, a destabilized geopolitical environment is inherently undesirable. Furthermore, the mechanism for hazard release is difficult to predict, as minor unexpected triggers can rapidly escalate into larger events.

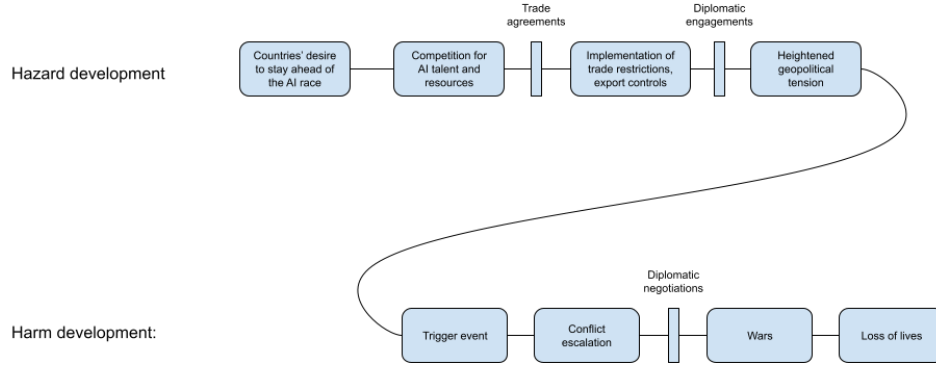


Figure 6: Risk pathway model for a geopolitical risk scenario

In Figure 6, an AI race leads to heightened geopolitical tension which results in a war with catastrophic consequences. Risk management measures here are thus less relevant to AI development and deployment itself, but rather more focused on managing geopolitical relationships.

### 6.6.3 Analogous historical precedence

Unlike many other wars where states fought over land and resources, the Cold War was primarily an ideological confrontation, where both the U.S. and the Soviet Union sought to establish global supremacy of their desired political and economic models. Though it did not result in direct military engagement between the two major powers, this conflict frequently led to widespread proxy wars across various regions such as Vietnam and Afghanistan [140]. While the causes of these proxy wars were often rooted in complex local and regional dynamics, their scale and intensity were significantly exacerbated as a second-order effect of the overarching ideological conflict between the two superpowers.

Instead of an ideological race, some countries are now focusing on winning the AI race instead. For example, the U.S. AI and Crypto Czar David Sacks was quoted as saying “to remain the leading economic and military power, the United States must win the AI race” [141]. This strategic rivalry, particularly between the U.S. and China, is discussed by [142] as a new “digital Cold War”. Experts have also warned that certain AI-related attacks could tip the geopolitical sphere that result in conditions similar to those that preceded the First and Second World Wars [143].

### 6.6.4 Other similar risks

It is argued that the widespread use of AIs can have potentially catastrophic epistemic side effects, where AIs system may make changes to humans’ (or other agents’) knowledge or beliefs because it was not told not to do so [144]. Similarly, [145] argues that AIs could lead to “structural violence” where due to unequal distribution of epistemic vulnerability, there will be a divide between epistemic agency and epistemic automation, where the vast majority of people who lack logic and scrutiny will cease to reason.

## 7 Summary

Table 8 summarizes the risk dimensions for each of the risks discussed. These risks are not meant to be comprehensive nor exhaustive, but they serve to illustrate a broad class of risk with at least one of the dimensions being distinct from the other risks.

Table 8: Summary of risk dimensions for the risks discussed

Risks	Intent	Competency	Entity	Polarity	Linearity	Reach	Order
CBRN	Intentional	Competent	Humans	Single-agent	Linear	Internalized	First-order
Cyber offense	Intentional	Competent	Variable	Single-agent	Linear	Internalized	First-order
Sudden loss of control	Variable	Competent	AI	Single-agent	Linear	Internalized	First-order
Gradual loss of control	Unintentional	Competent	Variable	Multi-agent	Non-linear	Internalized	First-order
Environmental risks	Unintentional	Variable	Variable	Variable	Linear	Externalized	First-order
Geopolitical risks	Unintentional	Variable	Variable	Variable	Non-linear	Externalized	Second-order

While these risks are analyzed independently to and characterized according to their risk dimensions, real-world risks are far more complex and interconnected. For example, an AI race that leads to geopolitical risks may result in an authoritarian regime carrying out an AI-enabled cyberattack on an autonomous weapons system. In this example, multiple risks with different risk dimensions may play out simultaneously, resulting in a "polycrisis" [146]. In addition, rapid advances in domains outside of AI such as biotech, robotics, quantum computing, and energy may also amplify risks [147]. In response, fields of studies at the intersection of different risks have emerged, such as biocybersecurity (or cyberbiosecurity), which attempts to address both cyber and biological risks [148].

## 8 Limitations

While this paper proposes an approach to analyzing catastrophic AI risks through dimensional characterization and risk pathway modeling, several limitations should be acknowledged:

- **Non-exhaustiveness of risks:** The six risks explored are not intended to be comprehensive, as they were selected because they are commonly discussed, where each risk has a distinct combination of risk dimensions. While we have included other examples of risks with similar dimensions where applicable, they remain non-exhaustive.
- **Subjectivity of risk dimensions:** The seven risk dimensions discussed alongside their attributes are just several ways AI risks can be characterized, such that risk management measures broadly associated with those attributes can be identified. There may be other useful dimensions that are not captured in this analysis.
- **Incompleteness of risk pathways:** The risk pathways shown only represent a particular scenario for a given risk, and does not intend to cover a significant distribution of how risks are realized. There may be other plausible scenarios that have not been covered by this analysis.
- **Lack of quantitative modeling:** This work focuses on qualitative causal mapping rather than quantitative risk estimation. As such, it does not attempt to assign probabilities nor severity to these risks.

## 9 Conclusion

Our work has provided a framework for characterizing AI risks in terms of their risk dimensions, where we explored six commonly discussed catastrophic risks along seven dimensions. For each attribute of these dimensions, the relevant risk management measures are listed, where they can be applied to risks associated with those dimensions.

We also conducted simple risk pathway modeling to map out the causal path to harm for each scenario. While we sometimes identified model capabilities as the hazard, many of the risks explored lie outside the control of model developers. Though the least cost avoidance principle correctly suggests focusing on the upstream parts of the AI value chain, we believe that an upstream focus is a necessary but insufficient condition for good risk management. It is crucial to conduct risk management at all levels of the value chain, which includes:

- The model level, e.g. during data collection, training, post-training, or deployment [67].
- The system or application level i.e. the specific areas it is being used [149]
- Broader societal and political level [150], [151]

Ultimately, a comprehensive strategy for managing catastrophic AI risks requires both a multi-dimensional understanding and concrete causal mapping. These can then facilitate robust risk management implemented across all levels of the AI value chain.

Future research can focus on four key areas. First, risks can be identified by conducting open-ended exploration across various risk dimensions. Second, more comprehensive risk pathway models can be developed, incorporating a wider array of scenarios and detailed mitigation strategies. Third, risks can be quantified by estimating the probability and severity based on its pathways and the effectiveness of proposed mitigations. Lastly, further investigation into risk mitigation measures beyond the model and application level, at the application and societal level, would be beneficial for managing these risks holistically.

## 10 Acknowledgements

This work was funded by the Open Philanthropy AI Governance and Policy grant, and supported by LISA (London Initiative for Safe AI) through subsidized access to its research coworking space. We also thank Francesca Gomez, Ben Bucknall, Marta Ziosi, and Peter Slattery for helpful discussion and feedback.



## References

- [1] Yoshua Bengio et al. “Managing extreme AI risks amid rapid progress”. In: *Science* 384.6698 (May 2024), pp. 842–845. ISSN: 1095-9203. DOI: 10.1126/science.adn0117. URL: <http://dx.doi.org/10.1126/science.adn0117>.
- [2] Center for AI Safety. *Statement on AI Risk*. Online statement. Accessed 2025-07-04. May 2023. URL: <https://safe.ai/work/statement-on-ai-risk>.
- [3] Timothy Dubber and Seth Lazar. *Military AI Cyber Agents (MAICAs) Constitute a Global Threat to Critical Infrastructure*. 2025. arXiv: 2506.12094 [cs.CY]. URL: <https://arxiv.org/abs/2506.12094>.
- [4] Anka Reuel et al. *Open Problems in Technical AI Governance*. 2025. arXiv: 2407.14981 [cs.CY]. URL: <https://arxiv.org/abs/2407.14981>.
- [5] Department for Science, Innovation and Technology. *Future risks of frontier AI*. Accessed: 2025-06-19. Apr. 2025. URL: <https://www.gov.uk/government/publications/frontier-ai-capabilities-and-risks-discussion-paper/future-risks-of-frontier-ai-annex-a>.
- [6] Department for Science, Innovation and Technology. *Frontier AI: capabilities and risks – discussion paper*. Accessed: 2025-06-19. Apr. 2025. URL: <https://www.gov.uk/government/publications/frontier-ai-capabilities-and-risks-discussion-paper/frontier-ai-capabilities-and-risks-discussion-paper#what-risks-do-frontier-ai-present>.
- [7] Department for Science, Innovation and Technology. *International AI Safety Report 2025*. Accessed: 2025-06-19. Jan. 2025. URL: <https://www.gov.uk/government/publications/international-ai-safety-report-2025/international-ai-safety-report-2025>.
- [8] Ben Eisenpress. *Catastrophic AI Scenarios*. Accessed: 2025-06-19. Feb. 2024. URL: <https://futureoflife.org/resource/catastrophic-ai-scenarios/>.
- [9] Dan Hendrycks, Mantas Mazeika, and Thomas Woodside. *An Overview of Catastrophic AI Risks*. 2023. arXiv: 2306.12001 [cs.CY]. URL: <https://arxiv.org/abs/2306.12001>.
- [10] Alan Chan et al. “Harms from Increasingly Agentic Algorithmic Systems”. In: *2023 ACM Conference on Fairness Accountability and Transparency*. FAccT ’23. ACM, June 2023, pp. 651–666. DOI: 10.1145/3593013.3594033. URL: <http://dx.doi.org/10.1145/3593013.3594033>.
- [11] Rohin Shah et al. *An Approach to Technical AGI Safety and Security*. 2025. arXiv: 2504.01849 [cs.AI]. URL: <https://arxiv.org/abs/2504.01849>.
- [12] Atoosa Kasirzadeh. *Two Types of AI Existential Risk: Decisive and Accumulative*. 2025. arXiv: 2401.07836 [cs.CY]. URL: <https://arxiv.org/abs/2401.07836>.
- [13] Torben Swoboda et al. *Examining Popular Arguments Against AI Existential Risk: A Philosophical Analysis*. 2025. arXiv: 2501.04064 [cs.CY]. URL: <https://arxiv.org/abs/2501.04064>.
- [14] Andrew Critch and Stuart Russell. *TASRA: a Taxonomy and Analysis of Societal-Scale Risks from AI*. 2023. arXiv: 2306.06924 [cs.AI]. URL: <https://arxiv.org/abs/2306.06924>.
- [15] Miles Brundage et al. *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*. 2024. arXiv: 1802.07228 [cs.AI]. URL: <https://arxiv.org/abs/1802.07228>.
- [16] Benjamin S. Bucknall and Shiri Dori-Hacohen. “Current and Near-Term AI as a Potential Existential Risk Factor”. In: *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. AIES ’22. ACM, July 2022, pp. 119–129. DOI: 10.1145/3514094.3534146. URL: <http://dx.doi.org/10.1145/3514094.3534146>.
- [17] Roman V. Yampolskiy. *Taxonomy of Pathways to Dangerous AI*. 2015. arXiv: 1511.03246 [cs.AI]. URL: <https://arxiv.org/abs/1511.03246>.
- [18] Alexey Turchin and David Denkenberger. “Classification of global catastrophic risks connected with artificial intelligence”. In: *AI & Society* 35.1 (2020). First published online May 3, 2018; print 2020, pp. 147–163. DOI: 10.1007/s00146-018-0845-5. URL: <https://doi.org/10.1007/s00146-018-0845-5>.
- [19] Peter Slattery et al. *The AI Risk Repository: A Comprehensive Meta-Review, Database, and Taxonomy of Risks From Artificial Intelligence*. 2025. arXiv: 2408.12622 [cs.AI]. URL: <https://arxiv.org/abs/2408.12622>.
- [20] Risto Uuk et al. *A Taxonomy of Systemic Risks from General-Purpose AI*. 2024. arXiv: 2412.07780 [cs.CY]. URL: <https://arxiv.org/abs/2412.07780>.
- [21] Laura Weidinger et al. “Taxonomy of Risks posed by Language Models”. In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. FAccT ’22. Seoul, Republic of Korea: Association for Computing Machinery, 2022, pp. 214–229. ISBN: 9781450393522. DOI: 10.1145/3531146.3533088. URL: <https://doi.org/10.1145/3531146.3533088>.

- [22] Gavin Abercrombie et al. *A Collaborative, Human-Centred Taxonomy of AI, Algorithmic, and Automation Harms*. 2024. arXiv: 2407.01294 [cs.LG]. URL: <https://arxiv.org/abs/2407.01294>.
- [23] OECD. *Stocktaking for the development of an AI incident definition*. Tech. rep. 4. Approved and declassified by OECD Committee on Digital Economy Policy. Paris, France: OECD Publishing, Oct. 2023. DOI: {10.1787/c323ac71-en}. URL: [https://www.oecd.org/en/publications/stocktaking-for-the-development-of-an-ai-incident-definition\\_c323ac71-en.html%7D](https://www.oecd.org/en/publications/stocktaking-for-the-development-of-an-ai-incident-definition_c323ac71-en.html%7D).
- [24] Mia Hoffmann and Heather Frase. *Adding Structure to AI Harm: An Introduction to CSET’s AI Harm Framework*. Tech. rep. Accessed: 2025-06-19. Center for Security and Emerging Technology, Georgetown University, July 2023. URL: <https://cset.georgetown.edu/publication/adding-structure-to-ai-harm/>.
- [25] Yi Zeng et al. *AI Risk Categorization Decoded (AIR 2024): From Government Regulations to Corporate Policies*. 2024. arXiv: 2406.17864 [cs.CY]. URL: <https://arxiv.org/abs/2406.17864>.
- [26] Anthony M. Barrett and Seth D. Baum. “A model of pathways to artificial superintelligence catastrophe for risk and decision analysis”. In: *Journal of Experimental & Theoretical Artificial Intelligence* 29.2 (May 2016), pp. 397–414. ISSN: 1362-3079. DOI: 10.1080/0952813x.2016.1186228. URL: <http://dx.doi.org/10.1080/0952813x.2016.1186228>.
- [27] John G. Sotos. *On the Unimportance of Superintelligence*. 2021. arXiv: 2109.07899 [cs.CY]. URL: <https://arxiv.org/abs/2109.07899>.
- [28] National Institute of Standards and Technology (NIST). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. NIST Special Publication NIST.AI.100-1. Accessed: 2025-06-19. Gaithersburg, MD, USA: National Institute of Standards and Technology, U.S. Department of Commerce, Jan. 2023. DOI: {10.6028/NIST.AI.100-1}. URL: <https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf%7D>.
- [29] National Institute of Standards and Technology (NIST). *Draft – Taxonomy of AI Risk*. Draft White Paper NIST Draft. Accessed: 2025-06-19. Gaithersburg, MD, USA: NIST, Oct. 2021. URL: [https://www.nist.gov/system/files/documents/2021/10/15/taxonomy\\_AI\\_risks.pdf](https://www.nist.gov/system/files/documents/2021/10/15/taxonomy_AI_risks.pdf).
- [30] National Technical Committee 260 on Cybersecurity of SAC. *AI Safety Governance Framework*. Technical Report, Version 1.0. Accessed 2025-06-23. Standardization Administration of China (SAC), Sept. 2024. URL: <https://www.tc260.org.cn/upload/2024-09-09/1725849192841090989.pdf>.
- [31] Chairs and Vice-Chairs of the General-Purpose AI Code of Practice. *Third Draft of the General-Purpose AI Code of Practice*. European Commission Digital Strategy Library. Last updated 11 March 2025; Accessed: 2025-06-19. Mar. 2025. URL: <https://digital-strategy.ec.europa.eu/en/library/third-draft-general-purpose-ai-code-practice-published-written-independent-experts>.
- [32] Anthropic. *Claude Opus 4 System Card — AI Safety Level 3 Deployment*. System Card / Technical Report. Published May 2025; accessed 2025-07-06. Anthropic, May 2025. URL: <https://www-cdn.anthropic.com/4263b940cabb546aa0e3283f35b686f4f3b2ff47.pdf>.
- [33] OpenAI. *Preparedness Framework v2*. Technical Report. Published Apr 15, 2025; accessed 2025-07-06. OpenAI, Apr. 2025. URL: <https://cdn.openai.com/pdf/18a02b5d-6b67-4cec-ab64-68cdfbdebcd/preparedness-framework-v2.pdf>.
- [34] Google DeepMind. *Frontier Safety Framework v2.0*. Technical Report. Published February 4, 2025; accessed 2025-07-06. Google DeepMind, Feb. 2025. URL: <https://storage.googleapis.com/deepmind-media/DeepMind.com/Blog/updating-the-frontier-safety-framework/Frontier%20Safety%20Framework%202.0.pdf>.
- [35] Meta. *Frontier AI Framework*. Online policy framework. Published Feb 3 2025; accessed 2025-07-06. Feb. 2025. URL: <https://ai.meta.com/static-resource/meta-frontier-ai-framework/>.
- [36] Microsoft. *Frontier Governance Framework*. Technical Report. Published Feb 8, 2025; accessed 2025-07-06. Microsoft, Feb. 2025. URL: <https://cdn-dynmedia-1.microsoft.com/is/content/microsoftcorp/microsoft/final/en-us/microsoft-brand/documents/Microsoft-Frontier-Governance-Framework.pdf>.
- [37] Frontier Model Forum. *Risk Taxonomy and Thresholds for Frontier AI Frameworks*. Technical Report. Accessed 2025-06-26. Frontier Model Forum, June 2025. URL: <https://www.frontiermodelforum.org/technical-reports/risk-taxonomy-and-thresholds/>.
- [38] ISO 31000:2018 *Risk management — Guidelines*. Tech. rep. Accessed 2025-06-24. International Organization for Standardization, 2018. URL: <https://www.iso.org/standard/65694.html>.
- [39] ISO 14971:2019 *Medical devices — Application of risk management to medical devices*. Tech. rep. Accessed 2025-06-24. International Organization for Standardization, 2019. URL: <https://www.iso.org/standard/72704.html>.

- [40] Leonie Koessler and Jonas Schuett. *Risk assessment at AGI companies: A review of popular risk assessment techniques from other safety-critical industries*. 2023. arXiv: 2307.08823 [cs.CY]. URL: <https://arxiv.org/abs/2307.08823>.
- [41] Toby Ord. *The Precipice: Existential Risk and the Future of Humanity*. 1st ed. Accessed 2025-06-24. London, UK: Bloomsbury Publishing, Feb. 2021, p. 480. ISBN: 9781526600233. URL: <https://www.bloomsbury.com/uk/precipice-9781526600233/>.
- [42] Zeynep Engin and David Hand. *Toward Adaptive Categories: Dimensional Governance for Agentic AI*. 2025. arXiv: 2505.11579 [cs.CY]. URL: <https://arxiv.org/abs/2505.11579>.
- [43] Thomas A. Widiger and Douglas B. Samuel. “Diagnostic categories or dimensions? A question for the Diagnostic and Statistical Manual of Mental Disorders—Fifth Edition”. In: *Journal of Abnormal Psychology* 114.4 (Nov. 2005). Accessed 2025-06-24, pp. 494–504. DOI: 10.1037/0021-843X.114.4.494. URL: <https://psycnet.apa.org/doi/10.1037/0021-843X.114.4.494>.
- [44] Helena Chmura Kraemer, Art Noda, and Ruth O’Hara. “Categorical versus dimensional approaches to diagnosis: methodological challenges”. In: *Journal of Psychiatric Research* 38.1 (2004), pp. 17–25. ISSN: 0022-3956. DOI: [https://doi.org/10.1016/S0022-3956\(03\)00097-9](https://doi.org/10.1016/S0022-3956(03)00097-9). URL: <https://www.sciencedirect.com/science/article/pii/S0022395603000979>.
- [45] Robert S. Siegler. “Cognitive variability”. In: *Developmental Science* 10.1 (Jan. 2007). Accessed 2025-06-24, pp. 104–109. DOI: 10.1111/j.1467-7687.2007.00571.x. URL: <https://onlinelibrary.wiley.com/doi/10.1111/j.1467-7687.2007.00571.x>.
- [46] Peter Fazekas and Morten Overgaard. “A Multi-Factor Account of Degrees of Awareness”. In: *Cognitive Science* 42.6 (Aug. 2018). Accessed 2025-06-24, pp. 1833–1859. DOI: 10.1111/cogs.12478. URL: <https://onlinelibrary.wiley.com/doi/10.1111/cogs.12478>.
- [47] Rokas Gipiškis et al. *Risk Sources and Risk Management Measures in Support of Standards for General-Purpose AI Systems*. 2024. arXiv: 2410.23472 [cs.CY]. URL: <https://arxiv.org/abs/2410.23472>.
- [48] Deloitte. *Risk modeling*. Deloitte Global, Consulting Risk Perspectives. Accessed 2025-07-06. URL: <https://www.deloitte.com/global/en/services/consulting-risk/perspectives/risk-modeling.html>.
- [49] Isaac Kim. “The Ultimate Guide to Risk Modeling”. In: *Experian Insights (blog)* (June 2025). Accessed 2025-07-06. URL: <https://www.experian.com/blogs/insights/the-ultimate-guide-to-risk-modeling/>.
- [50] National Institute of Standards and Technology. *Guide for Conducting Risk Assessments (SP 800-30 Rev. 1)*. Special Publication 800-30 Rev. 1. Accessed 2025-07-06. NIST, Sept. 2012. DOI: 10.6028/NIST.SP.800-30r1. URL: <https://csrc.nist.gov/pubs/sp/800/30/r1/final>.
- [51] Simeon Campos et al. *A Frontier AI Risk Management Framework: Bridging the Gap Between Current AI Practices and Established Risk Management*. 2025. arXiv: 2502.06656 [cs.AI]. URL: <https://arxiv.org/abs/2502.06656>.
- [52] Anna Katariina Wisakanto et al. *Adapting Probabilistic Risk Assessment for AI*. 2025. arXiv: 2504.18536 [cs.AI]. URL: <https://arxiv.org/abs/2504.18536>.
- [53] Frontier Model Forum. *Frontier Capability Assessments: Technical Report*. Technical Report. Accessed 2025-07-06. Frontier Model Forum, Apr. 2025. URL: <https://www.frontiermodelforum.org/technical-reports/frontier-capability-assessments/>.
- [54] Oscar Delaney, Oliver Guest, and Zoe Williams. *Mapping Technical Safety Research at AI Companies: A literature review and incentives analysis*. 2024. arXiv: 2409.07878 [cs.CY]. URL: <https://arxiv.org/abs/2409.07878>.
- [55] Department for Science, Innovation and Technology, UK. *Emerging Processes for Frontier AI Safety*. Policy Paper. [Accessed: 29 July 2025]. Department for Science, Innovation and Technology, Oct. 2023. URL: <https://assets.publishing.service.gov.uk/media/653aabb80884d000df71bdc/emerging-processes-frontier-ai-safety.pdf>.
- [56] OpenAI. *Disrupting malicious uses of AI by state-affiliated threat actors*. <https://openai.com/index/disrupting-malicious-uses-of-ai-by-state-affiliated-threat-actors/>. [Accessed: 22 July 2025]. Feb. 2024.
- [57] Kathrin Grosse et al. *Towards more Practical Threat Models in Artificial Intelligence Security*. 2024. arXiv: 2311.09994 [cs.CR]. URL: <https://arxiv.org/abs/2311.09994>.
- [58] Serdar Yazmyradov and Hoon Lee. “A Comprehensive Review of AI Security: Threats, Challenges, and Mitigation Strategies”. In: *The Journal of The Institute of Internet Broadcasting and Communication* (Dec. 2024), pp. 375–384. DOI: 10.17703/IJIBC.2024.16.4.375.

- [59] Lorenzaj Harris. “Threats and Risks in the AI Supply Chain A Comprehensive Analysis”. In: (Mar. 2025).
- [60] Shaun Ee et al. *Adapting cybersecurity frameworks to manage frontier AI risks: A defense-in-depth approach*. 2024. arXiv: 2408.07933 [cs.CY]. URL: <https://arxiv.org/abs/2408.07933>.
- [61] Alfonso de Gregorio. *Mitigating Cyber Risk in the Age of Open-Weight LLMs: Policy Gaps and Technical Realities*. 2025. arXiv: 2505.17109 [cs.CR]. URL: <https://arxiv.org/abs/2505.17109>.
- [62] Joshua Clymer et al. *An Example Safety Case for Safeguards Against Misuse*. 2025. arXiv: 2505.18003 [cs.LG]. URL: <https://arxiv.org/abs/2505.18003>.
- [63] Akash R. Wasil et al. *Governing dual-use technologies: Case studies of international security agreements and lessons for AI governance*. 2024. arXiv: 2409.02779 [cs.CY]. URL: <https://arxiv.org/abs/2409.02779>.
- [64] Isabel Richards, Claire Benn, and Miri Zilka. *From Incidents to Insights: Patterns of Responsibility following AI Harms*. 2025. arXiv: 2505.04291 [cs.CY]. URL: <https://arxiv.org/abs/2505.04291>.
- [65] Nikiforos Pittaras and Sean McGregor. *A taxonomic system for failure cause analysis of open source AI incidents*. 2022. arXiv: 2211.07280 [cs.AI]. URL: <https://arxiv.org/abs/2211.07280>.
- [66] Roman V. Yampolskiy and M. S. Spellchecker. *Artificial Intelligence Safety and Cybersecurity: a Timeline of AI Failures*. 2016. arXiv: 1610.07997 [cs.AI]. URL: <https://arxiv.org/abs/1610.07997>.
- [67] Toby Shevlane et al. *Model evaluation for extreme risks*. 2023. arXiv: 2305.15324 [cs.AI]. URL: <https://arxiv.org/abs/2305.15324>.
- [68] Bo Li et al. *Trustworthy AI: From Principles to Practices*. 2022. arXiv: 2110.01167 [cs.AI]. URL: <https://arxiv.org/abs/2110.01167>.
- [69] Marialena Vagia, Aksel A. Transeth, and Sigurd A. Fjerdings. “A literature review on the levels of automation during the years. What are the different taxonomies that have been proposed?” In: *Applied Ergonomics* 53 (2016), pp. 190–202. ISSN: 0003-6870. DOI: <https://doi.org/10.1016/j.apergo.2015.09.013>. URL: <https://www.sciencedirect.com/science/article/pii/S0003687015300855>.
- [70] R. Parasuraman, T.B. Sheridan, and C.D. Wickens. “A model for types and levels of human interaction with automation”. In: *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans* 30.3 (2000), pp. 286–297. DOI: 10.1109/3468.844354.
- [71] Blair Attard-Frost and David Gray Widder. *The Ethics of AI Value Chains*. 2024. arXiv: 2307.16787 [cs.CY]. URL: <https://arxiv.org/abs/2307.16787>.
- [72] Yoshua Bengio et al. *Superintelligent Agents Pose Catastrophic Risks: Can Scientist AI Offer a Safer Path?* 2025. arXiv: 2502.15657 [cs.AI]. URL: <https://arxiv.org/abs/2502.15657>.
- [73] Alan Chan et al. *Visibility into AI Agents*. 2024. arXiv: 2401.13138 [cs.CY]. URL: <https://arxiv.org/abs/2401.13138>.
- [74] Yonadav Shavit et al. *Practices for Governing Agentic AI Systems*. White Paper / Technical Report. Published Dec 14 2023; accessed 2025-07-17. OpenAI, Dec. 2023. URL: <https://openai.com/index/practices-for-governing-agentic-ai-systems/>.
- [75] Ahmad Mohsin et al. *A Unified Framework for Human AI Collaboration in Security Operations Centers with Trusted Autonomy*. 2025. arXiv: 2505.23397 [cs.AI]. URL: <https://arxiv.org/abs/2505.23397>.
- [76] K. Eric Drexler. *Reframing Superintelligence: Comprehensive AI Services as General Intelligence*. Tech. rep. Technical Report #2019-1. Accessed 2025-06-24. Future of Humanity Institute, University of Oxford, 2019. URL: <https://www.fhi.ox.ac.uk/reframing-superintelligence/>.
- [77] Lewis Hammond et al. *Multi-Agent Risks from Advanced AI*. Tech. rep. 1. Cooperative AI Foundation, 2025. DOI: 10.48550/ARXIV.2502.14143. arXiv: 2502.14143.
- [78] Rishi Bommasani et al. *Ecosystem Graphs: The Social Footprint of Foundation Models*. 2023. arXiv: 2303.15772 [cs.LG]. URL: <https://arxiv.org/abs/2303.15772>.
- [79] Alan Chan et al. *Infrastructure for AI Agents*. 2025. arXiv: 2501.10114 [cs.AI]. URL: <https://arxiv.org/abs/2501.10114>.
- [80] OWASP GenAI Security Project. *Multi-Agent System Threat Modeling Guide v1.0*. Technical Report. Accessed 2025-07-06. OWASP GenAI Security Project, Apr. 2025. URL: <https://genai.owasp.org/resource/multi-agent-system-threat-modeling-guide-v1-0/>.
- [81] CHARLES PERROW. *Normal Accidents: Living with High Risk Technologies - Updated Edition*. REV - Revised. Princeton University Press, 1999. ISBN: 9780691004129. URL: <http://www.jstor.org/stable/j.ctt7srgf> (visited on 06/25/2025).
- [82] Noam Kolt, Michal Shur-Ofry, and Reuven Cohen. *Lessons from complexity theory for AI governance*. 2025. arXiv: 2502.00012 [cs.CY]. URL: <https://arxiv.org/abs/2502.00012>.

- [83] Remco Zwetsloot and Allan Dafoe. *Thinking about risks from AI: Accidents, misuse and structure*. Lawfare (blog). Accessed 2025-06-24. Feb. 2019. URL: <https://www.lawfaremedia.org/article/thinking-about-risks-ai-accidents-misuse-and-structure>.
- [84] Kyle A Kilian. *Beyond Accidents and Misuse: Decoding the Structural Risk Dynamics of Artificial Intelligence*. 2025. arXiv: 2406.14873 [cs.CY]. URL: <https://arxiv.org/abs/2406.14873>.
- [85] Anthony M. Barrett et al. *AI Risk-Management Standards Profile for General-Purpose AI (GPAI) and Foundation Models*. 2025. arXiv: 2506.23949 [cs.AI]. URL: <https://arxiv.org/abs/2506.23949>.
- [86] Simon Mylius. *Systematic Hazard Analysis for Frontier AI using STPA*. 2025. arXiv: 2506.01782 [cs.CY]. URL: <https://arxiv.org/abs/2506.01782>.
- [87] Agnes Schim van der Loeff et al. *AI Ethics for Systemic Issues: A Structural Approach*. 2019. arXiv: 1911.03216 [cs.AI]. URL: <https://arxiv.org/abs/1911.03216>.
- [88] Jennifer Wang et al. *Distinguishing Predictive and Generative AI in Regulation*. 2025. arXiv: 2506.17347 [cs.CY]. URL: <https://arxiv.org/abs/2506.17347>.
- [89] N. Gregory Mankiw. *Principles of Economics*. 9th ed. Accessed 2025-06-25. Boston, MA: Cengage Learning, 2020, p. 864. ISBN: 9780357038314. URL: <https://www.cengage.com/c/principles-of-economics-9e-mankiw/9780357038314/>.
- [90] Organisation for Economic Co-operation and Development (OECD). *Algorithmic Impact Assessment Tool*. OECD.AI catalogue. Accessed 2025-07-06. Apr. 2025. URL: <https://oecd.ai/en/catalogue/tools/algorithmic-impact-assessment-tool>.
- [91] Pegah Nokhiz, Aravinda Kanchana Ruwanpathirana, and Helen Nissenbaum. *Rethinking Optimization: A Systems-Based Approach to Social Externalities*. 2025. arXiv: 2506.12825 [cs.AI]. URL: <https://arxiv.org/abs/2506.12825>.
- [92] Yusen Zheng et al. *Mechanism Design for Auctions with Externalities on Budgets*. 2025. arXiv: 2504.14948 [cs.GT]. URL: <https://arxiv.org/abs/2504.14948>.
- [93] Samson Tan, Araz Taeihagh, and Kathy Baxter. *The Risks of Machine Learning Systems*. 2022. arXiv: 2204.09852 [cs.CY]. URL: <https://arxiv.org/abs/2204.09852>.
- [94] Reva Schwartz et al. *Reality Check: A New Evaluation Ecosystem Is Necessary to Understand AI’s Real World Effects*. 2025. arXiv: 2505.18893 [cs.CY]. URL: <https://arxiv.org/abs/2505.18893>.
- [95] Allison Berke. “Can’t quite develop that dangerous pathogen? AI may soon be able to help”. In: *Bulletin of the Atomic Scientists* (Nov. 2023). Accessed 2025-06-27. URL: <https://thebulletin.org/2023/11/cant-quite-develop-that-dangerous-pathogen-ai-may-soon-be-able-to-help/>.
- [96] Jonas B. Sandbrink. *Artificial intelligence and biological misuse: Differentiating risks of language models and biological design tools*. 2023. arXiv: 2306.13952 [cs.CY]. URL: <https://arxiv.org/abs/2306.13952>.
- [97] Aidan Peppin et al. *The Reality of AI and Biorisk*. 2025. arXiv: 2412.01946 [cs.AI]. URL: <https://arxiv.org/abs/2412.01946>.
- [98] Roger Brent and T. Greg McKelvey Jr. *Contemporary AI foundation models increase biological weapons risk*. 2025. arXiv: 2506.13798 [cs.CY]. URL: <https://arxiv.org/abs/2506.13798>.
- [99] Anders Sandberg and Cassidy Nelson. “Who Should We Fear More: Biohackers, Disgruntled Postdocs, or Bad Governments? A Simple Risk Chain Model of Biorisk”. In: *Health Security* 18.3 (June 2020). First version 2019; accessed 2025-07-04, pp. 155–163. DOI: 10.1089/hs.2019.0115. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7310205/>.
- [100] Cassidy Nelson and Sophie Rose. *Understanding AI-Facilitated Biological Weapon Development*. Technical Report. Accessed 2025-07-04. Centre for Long-Term Resilience, Oct. 2023. DOI: 10.71172/nm7j-qzt1. URL: <https://www.longtermresilience.org/wp-content/uploads/2024/09/AI-Facilitated-Biological-Weapon-Development-Website-Copy-1.pdf>.
- [101] Luckey, David and Duhachek Muggy, Sara and Frey, Taylor and Stebbins, David and Rissman, Tracey and Espinosa, Bianca and Tapia, Daniel and McKelvey Jr., Greg and Pokhriyal, Neeti and Dawson, Joseph and et al. *Mitigating Risks at the Intersection of Artificial Intelligence and Chemical and Biological Weapons*. Research Report, RR-A2990-1. Accessed 2025-06-27. RAND Corporation, Homeland Security Operational Analysis Center, Jan. 2025, p. 204. DOI: 10.7249/RR-A2990-1. URL: [https://www.rand.org/pubs/research\\_reports/RR-A2990-1.html](https://www.rand.org/pubs/research_reports/RR-A2990-1.html).
- [102] Centers for Disease Control and Prevention (CDC). *About Anthrax*. <https://www.cdc.gov/anthrax/about/index.html>. [Accessed: 22-July-2025]. 2025.
- [103] U.S. Department of Justice. *Amerithrax Investigative Summary*. PDF (Freedom of Information Act release). [Accessed: 22 July 2025]. Feb. 2010. URL: <https://www.justice.gov/archive/amerithrax/docs/amx-investigative-summary.pdf>.

- [104] World Economic Forum. *The Global Risks Report 2025*. Technical Report (20th edition). Accessed 2025-07-04. World Economic Forum, Jan. 2025. URL: <https://www.weforum.org/publications/global-risks-report-2025/in-full/>.
- [105] Nick Bostrom. “Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards”. In: *Journal of Evolution and Technology* 9.1 (2002). First published Mar 2002; accessed 2025-06-27, pp. 1–31. DOI: n/a. URL: <https://www.nickbostrom.com/existential/risks.pdf>.
- [106] Justin B. Bullock, Samuel Hammond, and Seb Krier. *AGI, Governments, and Free Societies*. 2025. arXiv: 2503.05710 [cs.CY]. URL: <https://arxiv.org/abs/2503.05710>.
- [107] Fazl Barez et al. *Toward Resisting AI-Enabled Authoritarianism*. Working Paper. Published May 28 2025; accessed 2025-07-17. Oxford Martin School, AI Governance Initiative, May 2025. URL: <https://aigi.ox.ac.uk/publications/toward-resisting-ai-enabled-authoritarianism/>.
- [108] Yuxuan Zhu et al. *Teams of LLM Agents can Exploit Zero-Day Vulnerabilities*. 2025. arXiv: 2406.01637 [cs.MA]. URL: <https://arxiv.org/abs/2406.01637>.
- [109] Ralph Langner. “Stuxnet: Dissecting a Cyberwarfare Weapon”. In: *IEEE Security & Privacy* 9.3 (2011), pp. 49–51. DOI: 10.1109/MSP.2011.67.
- [110] CERT-IST. *Stuxnet: A worm which targets SCADA systems*. [https://www.cert-ist.com/public/en/S0\\_detail?code=stuxnet](https://www.cert-ist.com/public/en/S0_detail?code=stuxnet). [Accessed: 22 July 2025]. Sept. 2010.
- [111] David Albright, Paul Brannan, and Christina Walrond. *Did Stuxnet Take Out 1,000 Centrifuges at the Natanz Enrichment Plant? Preliminary Assessment*. Technical Report. [Accessed: 22 July 2025]. Institute for Science and International Security, Dec. 2010. URL: [https://isis-online.org/uploads/isis-reports/documents/stuxnet\\_FEP\\_22Dec2010.pdf](https://isis-online.org/uploads/isis-reports/documents/stuxnet_FEP_22Dec2010.pdf).
- [112] National Cyber Security Centre. *Impact of AI on cyber threat from now to 2027*. <https://www.ncsc.gov.uk/report/impact-of-ai-on-cyber-threat>. [Accessed: 22 July 2025]. May 2025.
- [113] United Nations Office for Disarmament Affairs (UNODA). *Background on Laws in the CCW*. UNODA website. Accessed 2025-07-17; no publication date provided. URL: <https://disarmament.unoda.org/the-convention-on-certain-conventional-weapons/background-on-laws-in-the-ccw/>.
- [114] Kristian Humble. “War, Artificial Intelligence, and the Future of Conflict”. In: *Georgetown Journal of International Affairs* (July 2024). Published July 12, 2024; accessed 2025-07-18. URL: <https://gjia.georgetown.edu/2024/07/12/war-artificial-intelligence-and-the-future-of-conflict/>.
- [115] Adam Bales. *AI takeover and human disempowerment*. Working Paper, No. 9-2024. Published April 2024; accessed 2025-07-17. Global Priorities Institute, Apr. 2024. URL: <https://globalprioritiesinstitute.org/wp-content/uploads/Bales-AI-Takeover-and-Human-Disempowerment.pdf>.
- [116] Nick Bostrom. *Superintelligence: Paths, Dangers, Strategies*. Accessed 2025-07-06. Oxford, UK: Oxford University Press, 2014. ISBN: 9780199678112. URL: <https://psycnet.apa.org/record/2014-48585-000>.
- [117] Miles Brundage. “Taking superintelligence seriously: Superintelligence: Paths, dangers, strategies by Nick Bostrom”. In: *Futures* 72 (Aug. 2015). Accessed 2025-07-06, pp. 32–35. DOI: 10.1016/j.futures.2015.07.009. URL: <https://www.fhi.ox.ac.uk/wp-content/uploads/1-s2.0-S0016328715000932-main.pdf>.
- [118] Nick Bostrom. “The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents”. In: *Minds and Machines* 22.2 (June 2012). Published June 13, 2012; accessed 2025-07-06, pp. 71–85. DOI: 10.1007/s11023-012-9281-3. URL: <https://link.springer.com/article/10.1007/s11023-012-9281-3>.
- [119] Stuart Armstrong. “General Purpose Intelligence: Arguing the Orthogonality Thesis”. In: *Analysis and Metaphysics* (Jan. 2013). Published January 2013; accessed 2025-07-06. URL: [https://www.fhi.ox.ac.uk/wp-content/uploads/Orthogonality\\_Analysis\\_and\\_Metaethics-1.pdf](https://www.fhi.ox.ac.uk/wp-content/uploads/Orthogonality_Analysis_and_Metaethics-1.pdf).
- [120] Stephen M. Omohundro. “The Basic AI Drives”. In: *Proceedings of the First AGI Conference (Frontiers in Artificial Intelligence and Applications, Vol. 171)*. Ed. by Pei Wang, Ben Goertzel, and Stan Franklin. Accessed 2025-07-06. IOS Press, Feb. 2008, pp. 483–492. URL: [https://selfawaresystems.com/wp-content/uploads/2008/01/ai\\_drives\\_final.pdf](https://selfawaresystems.com/wp-content/uploads/2008/01/ai_drives_final.pdf).
- [121] Carl Shulman. *Omohundro’s ‘Basic AI Drives’ and Catastrophic Risks*. Technical Report. [Accessed: 23 July 2025]. San Francisco, CA: Machine Intelligence Research Institute, Jan. 2010. URL: <https://intelligence.org/files/BasicAIDrives.pdf>.
- [122] Daniel Kokotajlo et al. *AI 2027: A Scenario for the Next Decade of Superhuman AI*. Online scenario report, AI Futures Project. Accessed 2025-07-06. Apr. 2025. URL: <https://ai-2027.com/ai-2027.pdf>.

- [123] U.S. Department of State, Office of the Historian. *The Allende Years and the Pinochet Coup, 1969–1973*. <https://history.state.gov/milestones/1969-1976/allende>. n.d. (no date). [Accessed: 23 July 2025].
- [124] Bill Bigelow. *The Overthrow of Democracy in Chile — A Timeline*. Zinn Education Project (Teaching People’s History series). [Accessed: 25 July 2025]. Sept. 2023. URL: <https://www.zinnedproject.org/materials/chile-coup-timeline/>.
- [125] Center for Justice and Accountability. *Where We Work — Chile*. <https://cja.org/where-we-work/chile/>. n.d. (no date). [Accessed: 23 July 2025].
- [126] Mikita Balesni et al. *Towards evaluations-based safety cases for AI scheming*. 2024. arXiv: 2411.03336 [cs.CR]. URL: <https://arxiv.org/abs/2411.03336>.
- [127] Joshua Krook. *When Autonomy Breaks: The Hidden Existential Risk of AI*. 2025. arXiv: 2503.22151 [cs.CY]. URL: <https://arxiv.org/abs/2503.22151>.
- [128] Jan Kulveit et al. *Gradual Disempowerment: Systemic Existential Risks from Incremental AI Development*. 2025. arXiv: 2501.16946 [cs.CY]. URL: <https://arxiv.org/abs/2501.16946>.
- [129] SEC & CFTC Staffs. *Findings Regarding the Market Events of May 6, 2010*. Joint Staff Report. [Accessed: 23 July 2025]. Washington, D.C.: U.S. Securities, Exchange Commission, and Commodity Futures Trading Commission, Sept. 2010. URL: <https://www.sec.gov/files/marketevents-report.pdf>.
- [130] Andrei A. Kirilenko et al. “The Flash Crash: High-Frequency Trading in an Electronic Market”. In: *Journal of Finance (Forthcoming)* (2017). [Posted to SSRN 27 May 2011; last revised 18 Apr 2017; accessed 23 July 2025]. DOI: 10.2139/ssrn.1686004. URL: <https://ssrn.com/abstract=1686004>.
- [131] Maurice Chiodo and Dennis Müller. *The Problem of Algorithmic Collisions: Mitigating Unforeseen Risks in a Connected World*. 2025. arXiv: 2505.20181 [cs.CY]. URL: <https://arxiv.org/abs/2505.20181>.
- [132] Kai Ebert et al. *AI, Climate, and Regulation: From Data Centers to the AI Act*. 2025. arXiv: 2410.06681 [cs.CY]. URL: <https://arxiv.org/abs/2410.06681>.
- [133] Intergovernmental Panel on Climate Change (IPCC). *Climate Change 2023: Synthesis Report. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Sixth Assessment Report, Synthesis Report. Core Writing Team: H. Lee & J. Romero (eds.); accessed 2025-07-06. Geneva, Switzerland: IPCC, Mar. 2023, p. 184. DOI: 10.59327/IPCC/AR6-9789291691647. URL: <https://www.ipcc.ch/report/ar6/syr/>.
- [134] Mario J. Molina and F. S. Rowland. “Stratospheric sink for chlorofluoromethanes: Chlorine atom-catalysed destruction of ozone”. In: *Nature* 249.5460 (1974), pp. 810–812. DOI: 10.1038/249810a0.
- [135] M. Norval et al. “The human health effects of ozone depletion and interactions with climate change”. In: *Photochemical & Photobiological Sciences* 10.2 (Feb. 2011), pp. 199–225. DOI: 10.1039/c0pp90044c.
- [136] United Nations Environment Programme. *Montreal Protocol on Substances That Deplete the Ozone Layer*. <https://ozone.unep.org/treaties/montreal-protocol>. n.d. (no date). [Accessed: 24 July 2025].
- [137] International Telecommunication Union and World Benchmarking Alliance. *Greening Digital Companies 2025: Monitoring Emissions and Climate Commitments*. Technical Report. [Accessed: 24 July 2025]. International Telecommunication Union & World Benchmarking Alliance, July 2025. URL: <https://www.itu.int/en/ITU-D/Environment/Documents/Publications/2025/Greening%20Digital%20Companies%202025%20Final.pdf>.
- [138] Shamma Al Qutbah. *AI Rivalries: Redefining Global Power Dynamics*. Trends Research & Advisory Insight. Accessed 2025-07-06. Feb. 2025. URL: <https://trendsresearch.org/insight/ai-rivalries-redefining-global-power-dynamics/>.
- [139] Peter Barnett and Aaron Scher. *AI Governance to Avoid Extinction: The Strategic Landscape and Actionable Research Questions*. 2025. arXiv: 2505.04592 [cs.CY]. URL: <https://arxiv.org/abs/2505.04592>.
- [140] Karna Venkatraj. *Rising Conflicts: An Analysis of Cold War Proxy Wars and Their Modern Application*. Plan II Honors Thesis, University of Texas at Austin. [Accessed: 24 July 2025]. May 2019. URL: <https://repositories.lib.utexas.edu/items/72e35277-79de-48da-81d0-cde8251b27b8>.
- [141] The White House, Office of Science and Technology Policy. *White House Unveils America’s AI Action Plan*. <https://www.whitehouse.gov/articles/2025/07/white-house-unveils-americas-ai-action-plan/>. [Accessed: 24 July 2025]. July 2025.
- [142] Mark Esposito. *AI geopolitics and data centres in the age of technological rivalry*. World Economic Forum, Centre for the Fourth Industrial Revolution. [Accessed: 29 July 2025]. July 2025. URL: <https://www.weforum.org/stories/2025/07/ai-geopolitics-data-centres-technological-rivalry/>.

- [143] Josephine Campbell. *Artificial Intelligence Cold War*. EBSCO Research Starters. [Accessed: 24 July 2025]. 2022. URL: <https://www.ebsco.com/research-starters/diplomacy-and-international-relations/artificial-intelligence-cold-war>.
- [144] Toryn Q. Klassen, Parand Alizadeh Alamdari, and Sheila A. McIlraith. “Epistemic Side Effects: An AI Safety Problem”. In: *Proceedings of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), Blue Sky Ideas Track*. [5 pages; accessed: 29 July 2025]. London, United Kingdom, May 2023. URL: <https://www.southampton.ac.uk/~eg/AAMAS2023/pdfs/p1797.pdf>.
- [145] Craig S Wright. *Cognitive Castes: Artificial Intelligence, Epistemic Stratification, and the Dissolution of Democratic Discourse*. 2025. arXiv: 2507.14218 [cs.CY]. URL: <https://arxiv.org/abs/2507.14218>.
- [146] Michael Lawrence et al. *Polycrisis Research and Action Roadmap 2024: Gaps, Opportunities, and Priorities for Polycrisis Research and Action*. Technical Paper. [Accessed: 29 July 2025]. Victoria, BC, Canada: Cascade Institute, Aug. 2024. URL: <https://cascadeinstitute.org/technical-paper/polycrisisroadmap/>.
- [147] Mustafa Suleyman and Michael Bhaskar. *The Coming Wave: Technology, Power, and the Twenty-First Century’s Greatest Dilemma*. Published Sep 5 2023 (hardcover); accessed 2025-07-19. Bodley Head / Penguin Random House, Sept. 2023, p. 352. ISBN: 9780593593974. URL: <https://www.penguinrandomhouse.com/books/722674/the-coming-wave-by-mustafa-suleyman-with-michael-bhaskar/>.
- [148] Lucas Potter, Orlando Ayala, and Xavier-Lewis Palmer. *Biocybersecurity – A Converging Threat as an Auxiliary to War*. 2020. arXiv: 2010.00624 [cs.CY]. URL: <https://arxiv.org/abs/2010.00624>.
- [149] Jia Yi Goh et al. *Measuring What Matters: A Framework for Evaluating Safety Risks in Real-World LLM Applications*. 2025. arXiv: 2507.09820 [cs.SE]. URL: <https://arxiv.org/abs/2507.09820>.
- [150] Nestor Maslej et al. *Artificial Intelligence Index Report 2025*. Tech. rep. [Accessed: 29 July 2025]. Stanford Institute for Human-Centered Artificial Intelligence, Apr. 2025. URL: <https://hai.stanford.edu/ai-index/2025-ai-index-report>.
- [151] Jamie Bernardi et al. *Societal Adaptation to Advanced AI*. 2025. arXiv: 2405.10295 [cs.CY]. URL: <https://arxiv.org/abs/2405.10295>.