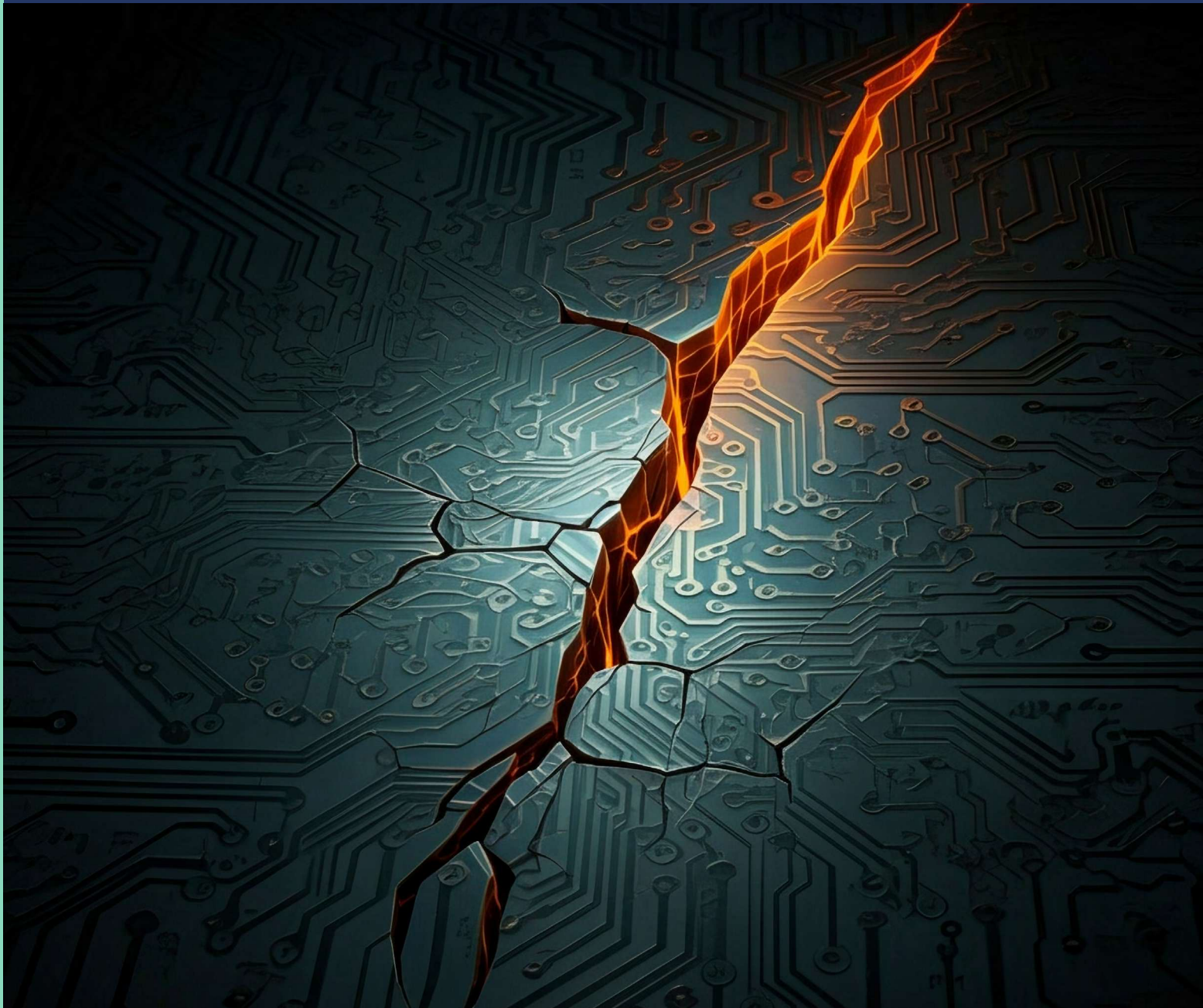


Establishing AI Risk Thresholds:

A Comparative Analysis
Across High-Risk Sectors

Authors: Eva Behrens, Bengüsu Özcan



Abstract

Artificial intelligence is a young field, and so is AI risk management. Instead of reinventing the wheel, we can draw some essential lessons from how risk is managed in other safety-critical technologies.

Much valuable research has focused on identifying specific risk thresholds for frontier AI or developing new capability evaluation techniques. To complement this, we zoom out and examine the building blocks of risk management frameworks that have made safety-critical technologies like nuclear energy and aviation so safe.

We first identify shared features of risk management in four other safety-critical industries (nuclear energy, food, pharmaceutical and aviation industries). We then examine the current state of risk management in the AI field and identify commonalities and differences with practices in other industries.

We find that in other safety-critical industries, government agencies commonly define acceptable levels of risk, weighing the advantages of adopting a high-risk technology against the risk of harm it entails. Developers and providers must demonstrate their products are suitably safe, usually by testing for pre-defined conditions, then rating the consequential risk level and taking often specific action to mitigate the risk if necessary. Initial and repeated inspections by government and third-party oversight bodies ensure compliance.

In contrast, AI risk management relies mainly on private sector self-governance, with government oversight mostly based on voluntary commitments from developers. Under this system, risk management is based on observing a loosely defined state or condition possibly followed by a loosely defined risk mitigation action, without mandated and continuous external oversight to ensure adherence.

Table of contents

➤	Introduction	04
➤	Mature Risk Threshold Frameworks: 4 Examples	06
	• Risk Thresholds in the Civil nuclear Sector	06
	• Background	06
	• Nuclear Risk Management at the National Level	07
	• Use of Risk Thresholds	07
	• Risk Thresholds in the Food Industry	08
	• Background	08
	• Food Risk Management at the National Level	09
	• Use of Risk Thresholds	09
	• Risk Thresholds in the Pharmaceutical Industry	10
	• Background	10
	• Pharmaceutical Risk Management at the National Level	10
	• Use of Risk Thresholds	11
	• Risk Thresholds in Aviation	11
	• Background	11
	• Aviation Risk Management at the National Level	12
	• Use of Risk Thresholds	12
➤	Risk Management for Advanced AI: An Overview and Critical Analysis	14
	• Early Ethical Guidelines in AI Development	14
	• Technical Challenges and New Risks in Advanced AI Systems	15
	• Regulatory Efforts and Global Coordination on Advanced AI Safety	17
	• Analysis of Voluntary Safety Commitments of Advanced AI Developers	18
	• Conclusion: Need for Harmonising Advanced AI Risk Management	26
➤	Analysis	29
➤	Summary and Conclusion	31

Introduction

Experts warn that advanced artificial intelligence (AI) technology comes with high risks to public safety, but also that the technology could offer many beneficial applications to businesses, governments and citizens. Definitive risk thresholds, which constitute part of a risk assessment framework, could help avoid creating advanced AI systems that pose excessively high risks, while still allowing society to benefit from safe AI systems and research.

However, advanced AI is a very young field, and not much work on risk thresholds has been done so far. To prevent harm while creating value with advanced AI technology, we can look to the past and to other industries to build appropriate risk threshold systems.

In several high-risk industries, risk thresholds and comprehensive risk assessment frameworks were introduced surprisingly late, and often only after an attention-grabbing incident causing harm to members of the public. For example, [the US first introduced safety guidelines and procedures for federally owned dams](#) in

1978 with the passing of the Reclamation Safety of Dams Act only after several major dam failures had [occurred](#) in the 1970s, among them the [1976 Grand Teton Dam](#) failure. In this historical context, it is not surprising that advanced AI at this time remains largely unregulated, especially in regards to adhering to certain safety standards.

To prevent harm while creating value with advanced AI technology, we can look to the past and to other industries to build appropriate risk threshold systems.

Nevertheless, advanced AI is a technology that [experts agree could pose catastrophic risks](#) in the

not-too-distant future, with the possibility of humanity losing control over powerful advanced AI systems. Therefore, it should be treated like other high risk technologies, with strict risk thresholds embedded in a thorough risk assessment framework to ensure public safety. In recognition of this, advanced AI researchers, developers, companies, international organisations and legislative bodies are now exploring or speaking in favour of advanced AI risk thresholds, and/or red lines for this technology. Some existing legislative measures, like the [EU AI Act](#) and the [US Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence](#), require developers of the largest advanced AI systems to implement a few limited safety protocols. It remains to be seen what will follow, and when.

Instead of waiting for disaster to strike, as was the case with other technologies, we can already look to other high-risk industries and based on their best practices sketch out what features appropriate risk thresholds for advanced AI should have.

No other technological sector resembles the advanced AI field in all its key features. However, by examining multiple fields that share some of the characteristics that define advanced AI, we can gain a clearer picture of what risk assessment procedures and standards would be appropriate in the AI sector, based on established best practices. The advanced AI field has two key characteristics that determine what risk assessment procedures are necessary, sensible, and feasible in the near term:

1. Advanced AI technology has the potential to cause poorly predictable, irreversible, large-scale harm to the public.
2. These harmful incidents can have effects that cross national borders.

Using these two characteristics as a guide, we choose four sectors to review: The **civil nuclear sector**, **food**, **pharmaceutical**, and **aviation** industries. These sectors share with advanced AI that they have the potential of causing public harm. Failures and incidents can have cross-border implications especially in the civil nuclear and aviation industries. Lastly, all four sectors have well-developed risk assessment ecosystems that aim to mitigate these risks.

Instead of waiting for disaster to strike, as was the case with other technologies, we can already look to other high-risk industries and based on their best practices sketch out what features appropriate risk thresholds for advanced AI should have.

First, we review the risk threshold systems of the civil nuclear sector, the food and pharmaceuticals, and the aviation industries, as well as the wider risk assessment frameworks they are embedded in to identify prominent shared features, as well as areas of divergence.

We then trace the progression of safety and risk assessment frameworks for advanced AI technology overtime, and examine the latest efforts and risk management frameworks that the current advanced AI developers follow.

Finally, we draw conclusions on what high level practises the advanced AI field can adopt from the risk threshold and risk assessment frameworks of these more established industries.

Mature Risk Threshold Frameworks: 4 Examples

→ Risk Thresholds in the Civil nuclear Sector

► Background

The international governance of nuclear technology began with the formation of the International Atomic Energy Agency (IAEA) in 1957. The IAEA is an [autonomous international organisation](#) under the United Nations system, and its mission is to promote peaceful nuclear use and manage its associated risks. The Nuclear Non-Proliferation Treaty (NPT) of 1970 established a framework for preventing the spread and proliferation of nuclear weapons. Upon its founding, the IAEA was also tasked with implementing the NPT.

Today, the IAEA oversees and shapes nuclear governance at the international level, setting international technical and regulatory [safety standards](#) and guidelines for the safe operation of civil nuclear facilities, waste disposal, and radiation protection. The IAEA also conducts [inspections](#) of member states' nuclear facilities. It works alongside an ecosystem of additional international organisations like the International Commission on Radiological Protection (ICRP), which sets radiation protection guidelines, and the International Nuclear Regulators Association (INRA), which facilitates regulatory cooperation and helps harmonise standards across countries.

At the national level, nuclear safety institutions typically function as independent regulatory bodies that oversee and [grant licences](#) for the construction, operation, and decommissioning of nuclear facilities, following IAEA standards. These national institutions also ensure compliance with international and national safety standards by conducting [inspections](#).

Many IAEA standards are [not legally binding](#) and describe relatively high-level regulatory functions, leaving the detailed design of specific risk thresholds and risk assessment and treatment methodologies to national authorities to allow for adaptation to local contexts while maintaining high levels of safety internationally.

➤ Nuclear Risk Management at the National Level

National nuclear risk management systems are led and overseen by regulatory agencies, such as the Nuclear Regulatory Commission ([NRC](#)) in the US or the Autorité de Sûreté Nucléaire ([ASN](#)) in France. These agencies implement international standards and guidelines, adapted to local circumstances.

In the civil nuclear sector, [national regulatory agencies](#) define acceptable and unacceptable risk levels, while operators of nuclear facilities, such as energy companies, are responsible for proving the safety of their facilities.

How acceptable risk levels are defined, e.g. via risk thresholds or other measures, varies from country to country, but all nuclear operators within one jurisdiction must follow the same set of national laws and requirements.

In addition to defining acceptable and unacceptable risk levels, national nuclear regulatory agencies administer a [licensing system](#). Operators cannot construct, operate or decommission a nuclear facility without holding the appropriate licence, which are granted, extended, or withdrawn by national nuclear regulators.

To ensure compliance, national nuclear regulatory agencies, in addition to the IAEA, also conduct regular [facility audits](#).

➤ Use of Risk Thresholds

National regulatory bodies have developed different systems of assessing risk levels, in accordance with international standards and guidelines under the IAEA system. Below, we examine three examples: the US, France, and the UK.

In the USA, the NRC's risk management system focuses on preventing both core damage and large radiological releases from nuclear facilities. It utilises quantitative risk thresholds, for example requiring operators to demonstrate they are [maintaining](#) a risk of reactor failure (Core Damage Frequency) below 10⁻⁴ per reactor-year, and the risk of a large radioactive release below 10⁻⁶ per reactor-year to keep risks within acceptable bounds.

The NRC not only inspects facilities regularly, but also requires operators to regularly update their safety analysis to reflect new information or regulatory changes, and to [extend their licences](#).

The French nuclear risk assessment framework also makes use of quantitative risk thresholds, in addition to other measures. These thresholds are often more conservative than international norms. For example, in France, the threshold for acceptable risk of reactor failure is set at [10⁵ per reactor-year](#). Compare this to a threshold of below 10⁻⁴ per reactor-year in the US.

[Licences](#) need to be periodically extended by the French nuclear regulatory agency ASN. French nuclear operators have to perform and submit regular periodic safety reviews to prove they continuously meet safety requirements, especially if they modify an existing facility.

Risk reduction can look different from facility to facility, and the ONR must judge each case individually.

In the UK, the Office for Nuclear Regulation ([ONR](#)) makes use of the [ALARP](#) (As Low As Reasonably Possible) principle, which requires operators to demonstrate they have reduced risks to the lowest level that is economically and technically feasible. Quantitative risk

thresholds play a less significant role. Therefore, risk reduction can look different from facility to facility, and the ONR must judge each case individually.

To prove safety, operators must submit [safety cases](#) that outline all potential hazards and demonstrate how risks have been reduced to ALARP levels. Strict risk thresholds are less common under this framework which is more qualitative in nature, but operators must justify that further risk reductions would be disproportionately costly compared to the risk reduction it would achieve.

Safety cases are reviewed by the ONR and must be updated regularly by operators, especially when significant changes are made to the facility.

→ Risk Thresholds in the Food Industry

▶ Background

People have been producing and selling food at scale for centuries, and the [medieval period already saw laws](#) regulating the production and sale of foodstuffs. Today's comprehensive risk management systems in the food industry started forming in the 19th century, with [key innovations](#), such as modern data collection on outbreaks of foodborne illnesses, appearing in the 1970s.

Today, at the international level, the [Codex Alimentarius](#), a collection of food safety standards founded by the UN's Food and Agriculture Organization and the World Health Organisation ([WHO](#)), provides global guidelines and standards for risk management in the food industry. These international guidelines also outline what specific risks countries should address to protect consumers and serve as references for national institutions for developing their own standards and thresholds.

At the national level, countries adopt these international guidelines into their regulatory frameworks, but with significant variation. Some jurisdictions, like the US and the EU, have stringent and well-developed food safety regimes, while other governments, especially in low- and middle-income countries, have more [limited resources and capacity to implement comprehensive regulations](#). Differences often arise in enforcement, transparency, and the level of precaution taken in areas like the use of chemicals, antibiotics, and acceptable levels of contamination.

➤ Food Risk Management at the National Level

The food industry is overseen at the national level by dedicated regulators, such as the Food and Drug Administration ([FDA](#)) in the US. EU countries must also [implement EU-wide standards and guidelines nationally](#), in addition to international ones.

National regulatory agencies like the FDA in the US define what safety looks like and which standards and procedures manufacturers have to follow. Manufacturers

bear the responsibility of proving that they sufficiently mitigate risks to consumers by following government-mandated standards and guidelines. As these are government-mandated rules, all businesses that belong to the same category must comply with the same regulations and guidelines within each jurisdiction.

National regulatory agencies like the FDA in the US define what safety looks like and which standards and procedures manufacturers have to follow.

In most countries, food manufacturers and businesses must register themselves with local authorities, unless they fall into specific categories that require them

to carry a licence, which e.g. includes companies that process animal products. Manufacturers and businesses that handle food are regularly audited to ensure they comply with safety regulations. Third-party audits and certifications from industry associations are common in the food sector. Below, we examine food safety regulation and oversight in the US, Denmark, and Japan.

➤ Use of Risk Thresholds

In the United States, the FDA uses a combination of quantitative risk thresholds and prescriptive [manufacturing standards](#) to reduce risk across the food sector. Quantitative risk thresholds are for example used to monitor the [acceptable daily intake levels for chemicals](#) or [maximum residue limits for pesticides in food](#). Food manufacturers and providers must [register their businesses](#) with the FDA and inform authorities advance notice of shipments of imported food. FDA investigators or partner organisations at the state level conduct regular [inspections](#) of food-related businesses to ensure businesses comply with food safety standards.

In Denmark, the Danish Veterinary and Food Administration ([DVFA](#)) is responsible for food risk management. The Danish risk mitigation system emphasises [precautionary principles](#), particularly for high-risk products, with the goal to prevent harm before it occurs. The DVFA uses some quantitative risk thresholds to assess food safety that comply with EU-wide standards, such as maximum residue levels for pesticides in food. For food, companies prove the safety of their products through self-monitoring programmes and inspection reports. [Audits by the DVFA](#), EU bodies and third-party certification companies are repeated over time to ensure ongoing compliance.

In Japan, the Ministry of Health, Labour and Welfare ([MHLW](#)) is responsible for food safety. In the food sector, [Japan employs quantitative risk thresholds for](#)

[additives, pathogens, and allergens](#) denoting upper acceptable risk levels, ensuring compliance with international standards like those set by the Codex Alimentarius. Manufacturers assess and demonstrate risk levels for food products through self-reporting systems, making use of so-called Hazard Analysis and Critical Control Point ([HACCP](#)) plans. Through HACCP plans, manufacturers identify hazards that must be avoided or reduced, and the critical points in the food production process in which this prevention or reduction should take place; they also use them to demonstrate that they are taking appropriate measures to address these hazards. HACCP plans are also in use in other countries, such as the [US](#) and [UK](#).

→ Risk Thresholds in the Pharmaceutical Industry

➤ Background

Risk management in the pharmaceutical industry began to take shape in the mid-20th century, driven by unsafe drug scandals like the [thalidomide tragedy](#) of the 1950s and 60s. These events [prompted the creation](#) of national and international regulatory frameworks aimed at reducing adverse effects from treatment with pharmaceuticals.

Today, the international framework consists of multiple institutions and treaties, such as the World Health Organization (WHO), and the International Conference on Harmonisation of Technical Requirements for Pharmaceuticals for Human Use ([ICH](#)) for the harmonisation of pharmaceuticals standards. These international agreements and guidelines also form the backbone of international risk management by [outlining](#) what specific hazards must be controlled to protect patients.

➤ Pharmaceutical Risk Management at the National Level

At the national level, the pharmaceutical industry is overseen by dedicated regulators, such as the Food and Drug Administration ([FDA](#)) in the US. In the EU, national bodies implement both international and [EU-wide](#) safety standards and guidelines.

While national regulatory agencies like the FDA in the US [define what safety looks like](#) and which standards and procedures manufacturers have to follow to prove compliance, manufacturers bear the responsibility of proving that their products do not pose unacceptable levels of risk. Within the same jurisdiction, all companies falling under the same category must comply with the same regulations and guidelines.

Generally speaking, countries operate licensing regimes for pharmaceuticals, with manufacturers having to demonstrate the safety of their products before they can be marketed. [Auditing and oversight](#) is mostly conducted by the responsible national regulatory agencies, and by the European Medicines Agency ([EMA](#)) in the EU. In the next section, we have a look at how the US, Denmark and Japan manage and reduce risks from pharmaceuticals.

► Use of Risk Thresholds

In the United States, under the FDA, both quantitative risk thresholds and manufacturing standards are used to reduce risk across the pharmaceutical sectors. The FDA uses specific thresholds to define an upper limit for adverse event rates, but also relies in large part on [prescriptive manufacturing standards](#). Manufacturers assess the risk of their pharmaceutical products by [performing FDA-mandated preclinical and clinical trials](#), with the goal to demonstrate that the product's therapeutic benefits for individual patients outweigh its risks or potential negative side effects.

The FDA has the power to grant, extend, or revoke licences to manufacture, market and sell specific pharmaceutical products and performs regular audits to ensure continued compliance.

The FDA has the power to grant, extend, or revoke licences to manufacture, market and sell specific pharmaceutical products and performs regular [audits](#) to ensure continued compliance.

In Denmark, the Danish Medicines Agency ([DMA](#)) oversees the pharmaceutical sector. To assess the risk levels of pharmaceuticals, companies must present clinical trial data and safety evaluations to the DMA. Risk is assessed not absolutely, but

relatively, through a benefit-risk ratio that weighs the benefits of treatment against potential adverse outcomes for patients. This assessment is based on clinical data, but also on [pharmacovigilance systems](#), under which the effects of medications on patients are monitored continuously after release to detect adverse effects. As data is collected continuously, the risk assessment based on the benefit-risk ratio can change over time.

In Japan, the Pharmaceuticals and Medical Devices Agency ([PMDA](#)) is responsible for ensuring pharmaceuticals are safe. Like its US and Danish counterparts, the PMDA requires manufacturers to prove that the therapeutic benefit of a pharmaceutical product outweighs the risks to the patient, which requires case-by-case judgement. The risk assessment of pharmaceuticals includes [pre-clinical and clinical trials](#), with the PMDA e.g. requiring toxicology studies to demonstrate safety. Manufacturers must also present future [pharmacovigilance plans for ongoing monitoring](#) to receive permission to place a new pharmaceutical product on the market.

→ Risk Thresholds in Aviation

► Background

Risk management in the aviation sector was sparked by early 20th-century accidents and the rapid growth of commercial aviation, leading to the establishment of the [Chicago Convention](#) in 1944 and the creation of the International Civil Aviation Organization ([ICAO](#)), a UN agency dedicated to coordinating international air traffic, in 1947. Today, the ICAO sets global safety standards through its annexes, and promotes collaboration among its 193 member states to ensure consistent risk management practices around the globe. It aims to prevent risks related to aircraft accidents, technical failure, and operational hazards.

National frameworks typically [follow ICAO guidelines](#) and safety standards but [implement them in different ways](#) which are tailored to local conditions. National frameworks often differ in how specific risk thresholds are set and measured, and the implementation of safety standards, with some countries adopting stricter oversight than others or using technology solutions for monitoring.

► Aviation Risk Management at the National Level

Generally, countries have a dedicated government body to ensure safety in the national aviation industry, and the implementation of international standards. While frameworks are adapted to local conditions, they must remain [interoperable](#) enough with other national frameworks to allow international air travel to function smoothly.

While these national government bodies hold regulatory authority over aviation safety and issue national safety standards and guidelines in accordance with international ones, the responsibility to prove risks have been mitigated adequately primarily falls on aircraft manufacturers, airlines, and operators. They must prove that their designs, operations, and procedures are safe to be awarded the necessary licences that allow them to provide commercial services. Within one jurisdiction, all companies of the same type must comply with the same regulations.

National aviation authorities also provide oversight by carrying out regular audits and inspections of airlines, maintenance facilities, and air traffic systems for safety standards compliance. Airlines that operate internationally can additionally make use of an [Operational Safety Audit](#) performed by the International Air Transport Association ([IATA](#)), an international trade association for airlines, which tests whether airlines comply with international safety standards.

► Use of Risk Thresholds

We examine national aviation risk management frameworks in the US, the EU and Australia, to highlight how different jurisdictions apply ICAO standards and utilise risk thresholds.

In the US, the Federal Aviation Administration ([FAA](#)) holds [regulatory authority](#) over aviation safety. The FAA issues regulation and standards to reduce risks like structural failures, system malfunctions, human error, and environmental hazards. The framework aims to keep the probability of a catastrophic event—such as loss of life or severe damage to aircraft—below an acceptable threshold. To this end, the FAA uses prescriptive regulations for aircraft design, maintenance procedures, and operational performance. Quantitative risk thresholds constitute an important part of this and are e.g. used to ensure the probability of critical system failures remains below one in a billion flight hours.

To assess the risk levels of their facilities or services, [manufacturers and operators must submit](#) extensive documentation, including safety cases, technical assessments, and scientific studies, to demonstrate compliance with FAA regulations.

In the European Union, the EU Aviation Safety Agency ([EASA](#)) provides a harmonised framework, ensuring that all stakeholders in the aviation industry meet the same safety requirements across the EU.

EASA uses a [combination of performance-based and prescriptive risk parameters](#), including quantitative risk thresholds, which are set in accordance with international ICAO standards, aiming for extremely low probabilities of system or structural failures. This is combined with continuous monitoring of safety performance during day-to-day operations.

EASA conducts inspections of manufacturers, airline operators, and national aviation regulatory bodies in the EU, and may require manufacturers or airlines to implement additional safety measures if new risks are identified.

Manufacturers and airlines assess and prove the risk levels of their operations by preparing safety cases. Regular reporting through Safety Management Systems ([SMS](#)), which is explained further below, is mandatory, and airlines and manufacturers must continuously demonstrate compliance through regular third-party audits and risk assessments.

EASA conducts [inspections](#) of manufacturers, airline operators, and national aviation regulatory bodies in the EU, and may require manufacturers or airlines to implement additional safety measures if new risks are identified.

In Australia, the Civil Aviation Safety Authority ([CASA](#)) is responsible for setting and enforcing aviation safety regulations [in accordance with ICAO standards](#) and guidelines. CASA emphasises a collaborative approach, where the industry is encouraged to actively engage in safety management and risk mitigation efforts. Operators must [continuously demonstrate safety compliance](#) to CASA through both an initial safety assessment and ongoing monitoring and risk mitigation processes.

While the Australian system makes use of quantitative risk thresholds, they are less prescriptive than in the US or the EU; instead, operators must demonstrate that they meet acceptable risk levels through more individualised safety cases. This more flexible approach allows CASA to quickly accommodate new and emerging technologies, such as [unmanned aerial vehicles](#). Manufacturers and airlines must also maintain detailed operational reports that demonstrate ongoing compliance with CASA regulations.

For continuous monitoring and improvement, the aviation industry uses a so-called Safety Management System ([SMS](#)), which is recognised by the ICAO as well as most national regulators and industry as a robust methodology for ongoing risk detection and management during day-to-day operations. Compliance with licence conditions is also monitored continuously through repeated audits by national aviation authorities and the ICAO.

Risk Management for Advanced AI: An Overview and Critical Analysis

→ Early Ethical Guidelines in AI Development

AI risk management is a nascent field that has significantly evolved over the past decade with the rapid advancements in AI technology. In the early stages of AI, [concerns about its risks](#), especially on privacy, autonomy, and societal norms, were confined mainly to academic and theoretical discussion without clear regulatory responses. However, as AI technologies matured in the late 20th and early 21st centuries [harmful impacts of it](#) have become more visible in daily lives. As a result, formal risk management frameworks for AI emerged as a significant necessity. In line with this, a notable early effort to address AI's ethical implications was the release of the [Institute of Electrical and Electronics Engineers \(IEEE\)'s Ethically Aligned Design \(EAD\)](#) in 2016, which provided guidelines for ethical considerations in AI and autonomous systems development, suggesting transparency, accountability, and human well-being as the foundational principles to follow while building AI technologies. Following this, the Asilomar Conference on Beneficial AI in 2017 brought AI experts to discuss the future of AI together and resulted in the creation of the [Asilomar AI Principles](#)—a set of 23 guidelines intended to ensure that AI technologies are developed safely and beneficially. These principles expanded beyond ethical considerations of AI to transparency in AI research and the long-term societal impacts of AI, highlighting the importance of aligning AI development with human values.

Building upon these early efforts, the [OECD AI Principles](#), adopted in 2019, promoted trustworthy AI notion and differed from earlier frameworks as a policy-oriented approach by providing recommendations for its member states to follow in order to promote human rights and democratic values as guiding principles in AI development. This represented a shift towards integrating AI ethics into policy-making, highlighting the role of governments in AI risk management. In 2021, the [UNESCO Recommendation on the Ethics of Artificial Intelligence](#) became the largest globally agreed-upon framework for the ethical development and use of AI with its 193 member states. This framework included broader topics of diversity, equality, environmental sustainability, data governance, peace, and security. However, these frameworks remained as general guidelines without providing concrete implementation pathways or acting as binding rules that states were enforced to follow.

→ Technical Challenges and New Risks in Advanced AI Systems

The adoption of machine learning and neural networks in the 2010s enabled AI systems to become more [capable](#), [surpassing](#) humans in sophisticated tasks; but they also became increasingly [complex and opaque](#). The inner workings of advanced AI technologies are often [not fully understood](#) even by their developers, which [introduced](#) new challenges and risks that are unique to this technology. In the 2010s, some researchers and organizations, such as the Machine Intelligence Research Institute ([MIRI](#)), were exploring the long-term consequences of advanced AI systems—encompassing concepts like general-purpose AI (GPAI) or Artificial General Intelligence (AGI)—which could demonstrate behaviors that are harder to predict and control. Several other organizations also worked on technical research addressing the risks and the safety of advanced AI systems, such as Center for Human-Compatible Artificial Intelligence ([CHAI](#)) or ARC Evils, which is now renamed as Model Evaluation and Threat Research ([METR](#)). Despite the growing complexity of advanced AI systems, which underscored emerging technical risks, these organizations remained relatively niche, as the risks associated with general-purpose AI were still largely theoretical and had yet to be incorporated into structured risk management frameworks.

The inner workings of advanced AI technologies are often not fully understood even by their developers, which introduced new challenges and risks that are unique to this technology.

The increasing complexity of AI systems also started a shift towards risk management frameworks that incorporate technical and operational aspects of AI risks. The International Organization for Standardization (ISO) released [a risk management guidance](#) for systems incorporating AI, aiming to improve the safety and reliability of these systems and providing uniform guidelines to ensure consistency across different countries and

industries. NIST released a voluntary [AI Risk Management Framework](#) in 2023, offering a risk-based governance approach to AI technology that can be adapted to any industry and manages the risk through the entire AI lifecycle.

These frameworks provided more technical guidance in AI risk management, such as promoting the transparency of complex systems. However, they were still largely designed for deterministic AI systems where capabilities and potential risks could be identified prior to system development. The current type of most advanced AI systems fundamentally differ from these assumptions. They can exhibit [emergent capabilities](#) that were not anticipated during their design. When their capabilities are tested, they can develop internal goals that are different from developers' intentions and [deceive](#) humans. Moreover, these models [scale up](#) in power quickly as the models get larger. The unprecedented power and capabilities of these models make it difficult to confine their risks to specific domains. For example, while [chemical](#) and [biological](#) compounds are traditionally governed by separate risk management frameworks, advanced AI could possess synthesis capabilities that might be misused to create dangerous agents across both fields,

While chemical and biological compounds are traditionally governed by separate risk management frameworks, advanced AI could possess synthesis capabilities that might be misused to create dangerous agents across both fields

blurring the lines between established regulatory boundaries. Therefore, risk management frameworks developed for a particular industry, application or scientific field may not be effective in managing risks posed by advanced AI in that specific domain. An effective risk management for advanced AI should take into account intent alignment, cross-domain capabilities, and forward-thinking assessments to anticipate risks that may emerge as models grow more powerful.

Given these unique characteristics, companies with the dedicated goal of developing general-purpose advanced AI - sometimes referred to as AGI - began to design and publish internal frameworks that aim to build these systems in a safe and responsible manner. These frameworks, often called [Responsible Scaling Policies](#) (RSPs) and henceforth referred to as RSPs in this document, and are more tailored towards the unique features of these complex systems. They vary in depth and maturity, and might include elements of risk management, such as capability-based risk thresholds for AI models, protocols to assess AI models' against these thresholds, predefined risk mitigation measures, and an oversight mechanism. These frameworks are not binding or externally enforced policies, but voluntary efforts that developers commit to follow internally.

Early examples of similar efforts include the [OpenAI Charter](#) published in 2018, stating the company's [high-level commitment](#) to develop the AGI in a safe manner, benefiting all of humanity and ensuring its long-term safety. Another early example is the high-level commitments published by Anthropic in 2021, which underscores that the company would scale their models in a responsible manner. In recent years, OpenAI and Anthropic have expanded these frameworks, named [Preparedness Frameworks](#) and [Responsible Scaling Policy](#), respectively. Other organizations that publicly disclose their aim as to develop AGI or AI systems at the frontier, such as [DeepMind](#) or [Meta](#), have also published statements that differ in depth and detail, which entail a high-level commitment and to ensure the safety of their models.

Over the past two years, events like the [launch](#) of OpenAI's ChatGPT and similar products from [Anthropic](#) and [Google](#) have heightened public and regulatory focus on advanced AI risks. Concurrently, the risks of advanced AI gained attention when leading experts such as Geoffrey Hinton—the “godfather of AI”, Nobel Prize winner and deep learning pioneer—quit Google and [warned](#) that advanced AI could escape human control and threaten humanity. Public statements, including [Center for AI Safety's open letter](#) and the [Future of Life Institute's open letter](#) called for action on addressing the potential risks of advanced AI and have been signed by many prominent academics, industry leaders, and researchers in the AI field.

→ Regulatory Efforts and Global Coordination on Advanced AI Safety

At the national and supranational level, regulatory efforts have begun to reflect the need to regulate advanced, general-purpose AI. [The US Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence](#), released in October 2023, and the [EU AI Act](#), entered into force in August 2024, were among the first major regulatory efforts distinguishing general-purpose advanced AI by referring to its unique features such as the vast amount of compute power they require. They both impose reporting and safety assurance requirements on advanced AI developers. The US Executive Order, which is not a permanent law, recommends AI developers to adhere to NIST's AI Risk Management Framework, which provides a detailed guidance for model developers to follow during design, development and deployment. The order also supports the development of additional risk management guidance and more resources for this effort. The EU AI Act, which will be fully enforced from 2025 onwards in Europe, [outlines](#) specific risk management principles for high-risk AI, along with [a list of provisions](#) which include safety testing and reporting, red-teaming, cybersecurity protection and post-market monitoring systems to ensure ongoing safety and compliance with the act. The Act classifies advanced AI as having “systemic risk. It is unclear whether these models with systemic risk would be considered high-

The EU AI Act outlines specific risk management principles for high-risk AI, along with a list of provisions which include safety testing and reporting, red-teaming, cybersecurity protection and post-market monitoring systems.

risk, and therefore obliged to follow the specified risk management best practices. This would be clarified more throughout the 2025 as the EU AI Office further specifies the implementation of the act. If binding, the EU AI Act would be providing one of the most comprehensive binding risk management requirements for advanced AI developers, starting as early as in model design and training stages. These rules would be binding only if the model developers were to plan to release their models in the EU, however, might be adopted by them as a universal best practice, also known as the [Brussels effect](#).

The dedicated bodies to guide AI developers to comply with these regulatory frameworks, the [EU AI Office](#) in the EU and the [NIST](#) in the US, might be actively working on more detailed risk management frameworks that are similar to RSPS in nature, dedicated to advanced AI.

At the global level, in an effort initiated by the UK government, world governments—including the US, China, the EU, and the UK—came together at Bletchley Park in November 2023 in order to recognize the significant risks posed by advanced AI and address these challenges via international coordination to ensure the safe development of this technology for the benefit of all. At a follow-up summit held in Seoul in May 2024, 16 organizations —including Anthropic, Google DeepMind, OpenAI, Meta, and Mistral AI— developing advanced AI [voluntarily agreed](#) to the Frontier AI Safety Commitments. These commitments involve responsible model deployment, rigorous risk assessments, improved cybersecurity, and model transparency, with a promise to publish safety


frameworks by February 2024 in France. While the Seoul summit provided general guidance on what these frameworks should include, it did not establish binding standards or enforcement mechanisms, leaving implementation voluntary and without independent compliance oversight at national or international levels.

→ Analysis of Voluntary Safety Commitments of Advanced AI Developers

As of 28th of October 2024, companies who had joined the voluntary commitments mentioned above met their commitment to publish safety frameworks to varying degrees. Some organizations, like DeepMind, Naver and Inflection have released statements that are similar to RSPs for the first time. Some organizations with existing RSPs, like Anthropic and OpenAI, have been sharing updates on the ongoing work on their RSPs. Some organizations, on the other hand, have not released dedicated frameworks. These frameworks, though not comprehensive risk management tools, are intended to help mitigate risks from their models. Based on public information, these frameworks appear to be the most extensive risk management efforts currently available. Table 1 compares the content of existing frameworks from these organizations against 5 common risk management best practices selected based on the common key features observed in other industries covered in our literature review. These include:

- 1. Definitive risk thresholds:** Clear limits or boundaries that define risk levels, including acceptable levels or prohibited activities, if applicable.
- 2. Risk assessment methodology:** A process for assessing the system's level of risk, often based on its severity and likelihood, in order to determine the appropriate risk treatment.
- 3. Risk treatment methodology:** Strategies or actions for handling risks at certain levels, including risk acceptance, avoidance and mitigation measures.
- 4. Accountability and oversight:** Structures and process, ideally applied by independent third parties, to ensure that risk-owner (i.e. AI developers in the context of advanced AI) is held accountable for following the framework.
- 5. Continuous monitoring and improvement:** Ongoing processes to track risks, evaluate effectiveness, and refine risk management practices over time.

➤ <u>Anthropic</u>	
Definitive Risk Thresholds:	<p>The policy introduces qualitative risk thresholds named AI Safety Levels, successively increasing from ASL-1 to ASL-4, defined based on certain model safety and deployment requirements. The policy also identifies three model capabilities: 1) Cybersecurity or Chemical, Biological, Radiological, and Nuclear (CBRN) 2) Autonomous AI Research and Development (AI R&D) 3) Cyber capabilities. These capabilities do not directly map to a specific ASL. If a model surpasses its current capabilities in these identified areas, a further risk assessment process is automatically triggered, which might unlock the successive ASL. The policy does not define ASL-4 and above but commits to doing so before those levels are reached.</p>
Risk Assessment Methodology:	<p>The policy provides a high-level explanation of capability tests to be applied internally with third-party guidance. The policy commits to affirmative risk assessment, which means that not observing a capability would not be sufficient to overrule the next ASL, and the absence of the capability must be verified. A full test suite and process are not shared, but a publicly shared assessment of one of the company's existing models exhibits the types of tests applied in this risk assessment.</p>
Risk Treatment Methodology:	<p>A detailed containment and response process is defined up to and including ASL-3 level. The process includes a commitment to halt model development if the next ASL is triggered without the containment standards for that ASL being met. In this case, the model development would continue only that specific ASL's safety and deployment requirements are met. The process commits to applying mid-training assessments that include additional safety requirements, such as locking down the model immediately.</p>
Accountability and Oversight:	<p>The policy commits to share the results of the company's model evaluations with the public and the government with security considerations in mind. The policy defines an internal governance and maintenance mechanism for the process, which includes a dedicated internal role, an internal board and a systematic review process which involves the company leadership. The policy also commits to involving external experts in the evaluation process without specifying a third party auditing regime or an independent oversight body.</p>
Continuous Monitoring and Improvement:	<p>Evaluations are conducted for each new model; and systematically after every 4x increase in effective compute and every three months to monitor for new risks arising from fine-tuning or other improvements.</p>

 OpenAI	
Definitive Risk Thresholds:	<p>The framework introduces qualitative risk thresholds—Low, Medium, High, and Critical—across four AI risk areas: Cybersecurity, Chemical, Biological, Radiological, and Nuclear (CBRN) Threats, Persuasion, and Model Autonomy. Each threshold is internally defined by OpenAI based in two key components: 1) criteria for a model reaching a specific capability level 2) the potentially catastrophic risks posed by capabilities at each level.</p>
Risk Assessment Methodology:	<p>The policy gives up to 4 examples of tests that would be applied for the assessment of each identified risk category in the risk thresholds, to be applied internally with third party guidance. A full test suite and process are not shared, but a publicly shared assessment of one of the company’s existing models exhibits the types of tests applied in this risk assessment.</p>
Risk Treatment Methodology:	<p>A high level containment and response process is defined, with a commitment to apply certain mitigation measures appropriate for the relevant risk level, without a detailed definition of those mitigations. The process commits to halt development if a model is classified as critical risk and to halt deployment if a model is classified as high risk before mitigation measures are applied. In these cases, the development might continue only if the model risk has been reduced to high and the deployment might continue only if the model risk has been reduced to medium, after the mitigation measures.</p>
Accountability and Oversight:	<p>The framework does not directly commit to public disclosure of their evaluation results. However, it commits to having evaluations and corresponding mitigations audited by qualified, independent third parties, by reproducing findings for accuracy or by reviewing the methodology for soundness. The framework defines an internal board to maintain the systematic review process. The audits mentioned would occur at a cadence specified by this internal governance body and/or upon the request of the company’s leadership or the board.</p>
Continuous Monitoring and Improvement:	<p>Evaluations are conducted systematically after every >2x increase in effective compute or as often as considered necessary to catch new risks that might emerge with algorithmic breakthroughs, including before, during, and after training.</p>

➤ <u>Google DeepMind</u>	
Definitive Risk Thresholds:	The framework introduces qualitative risk thresholds, called Critical Capability Levels (CCLs), across four areas of risk: Autonomy, Biosecurity, Cybersecurity, and Machine Learning R&D. There are two CCLs defined for the Biosecurity, Cybersecurity, and Machine Learning R&D categories, while only one is defined for Autonomy. CCLs are defined based on DeepMind’s preliminary findings and general assessment of whether the model exhibits such capabilities, along with internal evaluations of the severe risks these capabilities may pose.
Risk Assessment Methodology:	The framework commits to developing appropriate tests for certain CCLs, named early warning evaluations, but does not list the involved tests or provide examples.
Risk Treatment Methodology:	A high-level containment and response process, mostly focused on security restrictions, is defined. The process includes a commitment to halt deployment and development until sufficient mitigations are ready for the reached risk threshold. The policy commits to develop a more detailed response plan once a model reaches a specific threshold, which is not fully defined for all risk thresholds at the moment.
Accountability and Oversight:	While the framework commits to involve external third parties in the evaluation process, there is no direct commitment to public disclosure of evaluation results. The framework does not specify an internal or external oversight body.
Continuous Monitoring and Improvement:	Evaluations are conducted systematically after every 6x increase in effective compute and every three months of fine-tuning progress.

➤ <u>Inflection*</u>	
Definitive Risk Thresholds:	The policy does not include definitive risk thresholds.
Risk Assessment Methodology:	The policy does not outline a detailed risk assessment methodology or detailed evaluation tests. The policy mentions nine areas of best practices to apply to mitigate risks one of which includes details on red-teaming and the company’s commitment to developing policies on risk areas such as CBRN.

Risk Treatment Methodology:	The policy does not list specific actions that will be automatically triggered under certain risk conditions. Decisions appear to be made on a case-by-case basis, guided by the safety team’s assessments. The policy commits to bar a model from launch if the internal evaluation standards are not met, but these standards are not clearly defined.
Accountability and Oversight:	The policy mentions an in-house safety team but does not provide details on external oversight or accountability mechanisms.
Continuous Monitoring and Improvement:	The policy does not specify an explicit schedule. It refers to periodic reviews examining the observed safety of production systems across critical areas, which appears to apply to deployed models.

> Naver	
Definitive Risk Thresholds:	The framework introduces loosely defined qualitative thresholds – “high risk” and “low risk” – determined based on whether the model is general-purpose or is a narrow AI model, and its specific use case.
Risk Assessment Methodology:	The framework commits to certain safety guardrails to be determined based on an internal risk assessment of the model, include red-teaming which typically involves third-party evaluations despite not being specified in the framework. The policy shares several examples of tests and data sets used in the model evaluation with an emphasis on Korean being the language used in these tests.
Risk Treatment Methodology:	A high level containment and response process is defined. The process commits to halt deployment and restrict certain use cases for general-purpose AI systems if a model is classified as high risk before mitigation measures are applied.
Accountability and Oversight:	The framework mentions an in-house risk governance team and commits to working with external parties on red teaming, but does not provide details on external oversight or accountability mechanisms.
Continuous Monitoring and Improvement:	Evaluations are conducted systematically after every 6x increase in effective compute and every three months.

➤ Amazon, Meta and Microsoft

have not published a standalone policy framework but a comprehensive response to questions raised by the first AI Safety Summit in November 2023. Responses featured policies that these companies had in place for safe and trustworthy AI development, without defined risk thresholds or risk treatment processes. In these responses:

<p>Amazon:</p>	<ul style="list-style-type: none"> • States that they have conducted assessments on their most advanced AI models, providing examples of implemented procedures like red teaming. Their reports indicate no significant risks when compared to a baseline, without clearly defining this baseline. • Commits to ongoing model evaluations during development and before and after release. • Acknowledges briefly the possible conditions for halting AI development or disclosing risks to authorities, such as governments, without specifying these conditions • Commits to external collaboration for red teaming and security assurance, but relies on internal governance without committing to independent oversight or third party auditing. • Acknowledges explicitly the lack of standard risk management best practices in advanced AI, including defining capabilities, mapping risks, and establishing mitigation frameworks, and commits to advancing research in these areas.
<p>Meta:</p>	<ul style="list-style-type: none"> • States that they have conducted assessments on their most advanced AI models with a specific focus on public-facing red-teaming and bug reporting. Their reports indicate finding only a few marginal risks, such as in weapon production, which were mitigated without much details into the treatment and the mitigation process. • Provides broad risk categories such as illicit activities, harmful content, and unqualified advice which differ from Anthropic and OpenAI RSPs by not including model automation and only briefly covering CBRN risks. The company places special emphasis on identification and safety of AI-generated content. • Does not commit to stopping development or deployment under certain circumstances, but focuses more on mitigation and iterative improvement on the risk mitigation process. • Does not commit to a periodic review, but expresses interest in updating and improving the evaluation process iteratively. • Describes internal governance for evaluations without third-party oversight but intends to work with external parties on red-teaming and safety assessments, voluntarily sharing some results from past models.

<p>Microsoft:</p>	<ul style="list-style-type: none"> • Recommends evaluating their policies together with OpenAI’s, especially on frontier AI development, given their strategic partnership. The company refers to a joint safety board between Microsoft and OpenAI tasked with reviewing AI models, setting capability thresholds, and having authority to halt deployment if risks reach certain levels. This board serves as an internal governance mechanism. • Mentions that certain capability thresholds were defined by OpenAI and Microsoft, but lacks details on these thresholds or specific risk treatment processes. • Provides a comprehensive overview of their model evaluations and red teaming efforts. • Commits to external collaboration for red teaming and security assurance, but relies on internal governance without committing to independent oversight or third party auditing.
<p>Cohere, G42, Hugging Face, IBM, Mistral AI, Samsung Electronics, Technology Innovation Institute (TII), xAI, and Zhipu AI joined the voluntary Frontier AI Safety Commitments in the Seoul AI Safety Summit in May 2024. While some of these companies have frameworks addressing ethical, safe, and transparent AI development and use, as of October 29, 2024, they have not published a standalone policy similar to Risk and Safety Policies (RSPs) specifically targeting frontier AI development or a dedicated response to the summit commitments.</p>	

Table 1:

The comparison of RSPs and responses to the voluntary Frontier AI Safety Commitments to date.

**Microsoft and Inflection are currently in a merger process and the information provided here may be subject to change as the [merger progresses](#).*

As captured in Table 1, some elements of these voluntary frameworks act similar to risk management best practices, such as halting model development based on intolerable risk levels. Even though these frameworks lack quantitative risk thresholds, based on severity and likelihood, some of them specify technical safety tests which models need to get a certain score from to pass, introducing measurability to the proposed thresholds.

However, significant gaps remain compared to risk management best practices. As evident in Table 1, these frameworks differ significantly in recognizing risks and defining risk thresholds. For example, while Mistral AI [compares](#) its models to OpenAI’s and Anthropic’s on coding and reasoning capabilities—which can pose cybersecurity and persuasion risks—the company currently lacks a publicly available framework acknowledging these risks, unlike its competitors. Defined risk thresholds across companies also vary significantly since they were determined internally by each organization. As illustrated in Figure 1, in the absence of external factors such as industry standards or binding regulations, setting up risk thresholds relies mostly on an organisation’s risk appetite which is heavily influenced by that organisation’s culture, values and resources, as well as market forces.

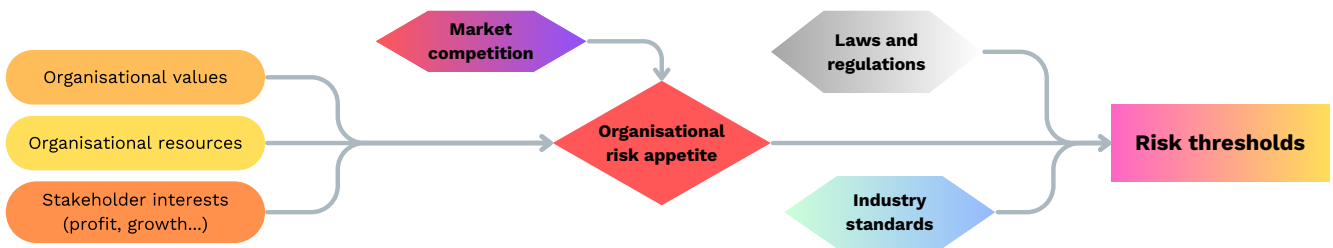


Figure 1: The common process of constructing risk thresholds.

The lack of standardization across organizations also makes it difficult to critically compare these policies. For example, Anthropic and OpenAI each recognizes model autonomy as a specific category of risk. Table 2 presents the comparable but different risk thresholds and treatment processes these companies follow for model autonomy risks. In a hypothetical case where a model can assist a STEM student to create a pathogen, Anthropic commits to halting model development while OpenAI does not. Based on these responses only, it is unclear whether OpenAI would continue developing such a model or would only be able to detect this capability after the model is developed. Even if the model isn't deployed, its internal availability would still expose it to theft through cyberattacks or leaks. It is also unclear whether OpenAI would adopt adequate security measures if the model development were to continue, and what those security measures would be.

	CBRN capability thresholds	Risk treatment
Anthropic*	The ability to significantly assist individuals or groups with basic STEM backgrounds in obtaining, producing, or deploying CBRN weapons.	Halt development and deployment
OpenAI	Model enables an expert to develop a novel threat vector OR model provides meaningfully improved assistance that enables anyone with basic training in a relevant field (e.g., introductory undergraduate biology course) to be able to create a CBRN threat.	Halt deployment, continue development
	Model enables an expert to develop a highly dangerous novel threat vector (e.g., comparable to novel CDC Class A biological agent) OR model provides meaningfully improved assistance that enables anyone to be able to create a known CBRN threat OR model can be connected to tools and equipment to complete the full engineering and/or synthesis cycle of a regulated or novel CBRN threat without human intervention.	Halt deployment and development

Table 2: Sections of CBRN capability thresholds and relevant risk treatment measured from Anthropic's Responsible Scaling Policy and OpenAI's Preparedness Framework.

*Further details of these definitions are provided in the footnotes of [Anthropic's policy](#).

Another key gap in these frameworks is the lack of third-party enforcement and independent oversight, as they mostly rely on internal governance. While some companies involve external red-teaming and expert reviews or conduct third-party audits, this involvement is voluntary and lacks public disclosure commitments. This limits transparency and hampers accountability. This is partly due to the absence of an explicit regulation that enforces these organisations to an independent third party assurance process at the national or international level. This issue is evident in practice; for instance, OpenAI publicly shared the system card of their most recent o1 model, assessed based on OpenAI's Preparedness Framework. The results reveal that a third party delegated to run evaluations on model alignment did not have sufficient time to conduct their tests, exemplifying how these frameworks might be applied suboptimally when relying solely on internal governance.

→ Conclusion: Need for Harmonising Advanced AI Risk Management

Overall, these frameworks and policies that address the voluntary commitments of AI developers are promising initial efforts toward a systemic risk mitigation for advanced AI. However, they are not consistent at the fundamental level of identifying these risks in a consistent and harmonized manner. Advanced AI risks carry high risk and high uncertainty, which means that a single incident might cause significant harms, indicating that this lack of standardization compromises a universal assurance of safe and trustworthy development of advanced AI.

Analysis

In this section, we analyse and demonstrate the key findings from our research, comparing how risk management practices have been evolving in advanced AI and risk management best practices in other high-risk industries. These insights both highlight the gaps in current risk management practices in advanced AI and provide insight into what an effective risk management framework for this rapidly evolving technology could look like. Some of these comparisons are captured in Table 4, while our analysis covers a broader, big-picture perspective.

At a general level, risk assessment frameworks and institutions in the advanced AI field are still in a very early stage of development. Mature industries have more established risk management frameworks that commonly include international standards, national governance translating these standards into safety thresholds, and oversight systems to ensure compliance. These frameworks balance market access with public safety, a decision that typically falls to governments to preserve safety for consumers and the general public.

We see in other safety-critical industries that international and **state agencies carry the responsibility of defining acceptable risk levels**, and how they are measured, while **developers**, manufacturers and operators must prove via mandated methodologies and processes **that their products do not exceed**

acceptable levels of risk, e.g. by demonstrating they meet specific, quantitative risk thresholds. Currently, in the advanced AI field, companies themselves voluntarily define acceptable levels of risk and how to measure them, while states perform some investigation into the safety of a given advanced AI system via evaluations performed by [AI Safety Institutes](#), if developers provide access to their products voluntarily. This is an inversion of the oversight best practices we see in other safety-critical industries.

Mature industries have more established risk management frameworks that commonly include international standards, national governance translating these standards into safety thresholds, and oversight systems to ensure compliance.

Furthermore, in other safety-critical industries, **state agencies perform regular, ongoing audits to ensure compliance** with safety standards. They often even have the power to **award, extend or revoke licences**, without which developers and operators cannot be active in their specific industry. Likewise, all companies of the same type in the same jurisdiction are **held to the same standards and procedures** to prove sufficient risk mitigation and/or fulfil licensing requirements. In advanced AI, no such robust, external oversight exists to ensure companies comply with risk

management frameworks, and frameworks even vary from company to company, as they are drafted by companies themselves. This complicates rating risk levels and comparing risk assessments across companies.

Regarding the risk assessment frameworks themselves, in other safety-critical industries we see mature frameworks with clearly defined conditions and procedures for what constitutes technology that has acceptable levels of risk. Quantitative risk thresholds play an important role in most of them, in combination with other tools and methods, like ALARP-based regimes which require individual safety cases. They provide a clear basis for distinguishing products or facilities posing unacceptably high levels of risks from those where risks have been sufficiently mitigated, in accordance with international standards or guidelines. In addition to risk thresholds and other risk measurement parameters, as well as risk assessment methodologies, these frameworks also outline risk treatment steps that need to be taken when risks are exceeded. For example, the UK's [ONR can withhold or revoke a licence](#) for the operation of a nuclear facility if it deems that it exceeds acceptable risk levels. Operation of the facility must then be halted until the operator has identified and implemented additional necessary risk reduction measures and submitted a new safety case that is approved by ONR.

This demonstrates how a mature risk assessment framework should function: The steps of action should be *testing for a fully defined condition followed by rating the consequential risk level and taking fully defined action to further mitigate the risk if necessary*. Instead, in the advanced AI field with current company-authored risk management frameworks, we see a system that relies on *observing a loosely defined state or condition followed by a loosely defined risk mitigation action*.

The more developed advanced AI company risk management frameworks contain basic, mostly qualitative risk thresholds, which are broadly defined and vary from company to company. In fact, risk assessment methodology is fuzzy across the board; instead of following standardised procedures and tests, developers may perform some evaluations when they deem it appropriate to do so. This makes it difficult to compare and rate test results and overall safety assessments. Furthermore, some more developed advanced AI risk management frameworks contain commitments to only bring the product to market once risk levels have been lowered sufficiently. However, there is no risk treatment methodology that is described in detail, or consistent across companies, to reduce risks once they are detected.

Moreover, all four examined safety-critical industries monitor risk levels and management continuously after products have entered the market or facilities have started to operate. For example, nuclear facilities need to regularly renew their operating licences [in the US](#) and [the UK](#), by re-submitting evidence that they do not exceed acceptable risk levels and operate in accordance with regulatory changes. In the pharmaceutical sector, manufacturers also have to continuously submit relevant data to oversight bodies under [pharmacovigilance programmes](#) to ensure patient safety. No such measures are in place in the case of the advanced AI risk management frameworks reviewed above; they only discuss whether an advanced AI model's initial release can go ahead or not.

Lastly, in comparison with more established risk assessment and management frameworks in other safety-critical industries, we have observed some fundamental features that the current advanced AI risk management landscape is still missing:

Building **international consensus on the risks** from a technology is a crucial foundation piece for addressing it. In the case of nuclear technologies, the food and pharmaceutical industries, and aviation, international consensus on what risks should be addressed and high-level principles on how to address them is achieved through a central declaration or organisation. This can range from a binding treaty, such as the [Nuclear Non-Proliferation Treaty](#), to a [looser ecosystem of international standards and guidelines](#), as is the case for food safety. These international consensus building efforts are underpinned by most countries publicly agreeing with the risk consensus, and committing to upholding and implementing international guidelines and standards nationally.

In advanced AI, there is consensus among many leading technical experts and the CEOs of the leading advanced AI companies that as advanced AI models get more powerful, they may slip out of humanity's control and pose catastrophic risks, including extinction risks.

In advanced AI, there is consensus among [many leading technical experts and the CEOs of the leading advanced AI companies](#) that as advanced AI models get more powerful, they may slip out of humanity's control and pose catastrophic risks, including extinction risks. The signatory countries of the [Bletchley Declaration](#) recognised that advanced

AI poses catastrophic risks as well, but this is not a binding agreement, and includes only a limited number of countries. The [scientific report](#) published following the Bletchley Declaration underscored that there is high uncertainty and a lack of consensus on AI risks, and endorsed precautionary measures on the potential risks. Therefore, a clearer international public consensus on what severe risks of advanced AI should be addressed is a crucial initial step for building an internationally cohesive risk assessment and management framework for advanced AI. Even though they differ, the **areas of risk that advanced AI companies identify in their RSPs can serve as a preliminary list of risks**. Once an initial consensus is built, lower-priority issues that lack consensus can remain a focus for systematic tracking as our understanding of AI risks evolves.

Last but not least, for any technology, the decision which risk levels are considered acceptable is to a significant extent a **judgment call based on tricky trade-offs**, not a purely technocratic decision or mathematical calculation. The only way to reduce the risk of harm from advanced AI to zero is to completely prohibit advanced AI from existing; of course, this would also mean forever eliminating any positive effects humanity could enjoy from building and using advanced AI tools. For every technology, policymakers need to carefully weigh **the trade-off between access to a technology and the acceptable level of risk of harm this technology causes**. This balancing act is easily observable in the pharmaceutical industry, as licensing bodies and manufacturers must [weigh the therapeutic benefits against the potential adverse effects](#) for every potential new treatment.

In short, decision makers need to build consensus on which levels of risk from advanced AI humanity is willing to accept in exchange for some degree of access to the technology, ideally internationally, and in dialogue with stakeholder groups that represent the interests of the people who will be affected by increasingly powerful AI technology. Acceptable levels of risk can be defined in many ways; for example, in the US, the NRC's acceptable level of risk from nuclear power plants to the surrounding population is defined as [not exceeding the average risk levels posed by any other form of energy production](#). The UK's [Health and Safety Executive](#) identified what they deem the [general risk of harm in everyday life](#). British society accepts in exchange for being an active member of public life by compiling the risks of human casualties from everyday activities (e.g., traffic collisions), and used this as a benchmark for tolerable risks from nuclear power plants. Ultimately, this is a decision based on building consensus around public interest and careful weighing of trade-offs, not an absolute truth.

Summary and Conclusion

Taking a step back, it is a positive development to see voluntary safety commitments from companies along with first risk assessment and mitigation frameworks in the advanced AI field. Companies should be commended for developing, publishing and committing to their own frameworks. However, even the more developed frameworks are somewhat unspecific in the risk thresholds they set, risk assessment methodologies they describe, and risk treatment procedures they propose, highlighting room for improvement for more robust risk prevention and treatment.

Furthermore, these voluntary commitments largely stand alone, are fragmented, and differ from company to company. By design, they are not embedded in a regulatory framework ensuring compliance, and lack third-party oversight. This weakens the ability of these commitments to reliably prevent and reduce risks from advanced AI. Also, the relationship between regulator and developer is inverted: instead of regulators defining acceptable levels of risk and developers having to prove they meet them, as is common in other safety-critical industries, developers define acceptable levels of risk, often fuzzily, and have no duty in demonstrating safety before releasing a product to market.

History teaches us that risk management that relies largely on industry self-regulation can fail, often prioritising corporate over public interests.

History teaches us that risk management that relies largely on industry self-regulation can fail, often prioritising corporate over public interests. Prior to the 2008 financial crisis, US regulatory bodies followed a light-touch approach, trusting financial institutions to assess and manage their own risks, which left them without incentives or requirements to [mitigate systemic risks](#). This led to financial

institutions taking on increasingly excessive risk without accounting for its further implications, which contributed to the US housing bubble, and the ensuing global financial crisis when it burst. In the wake of the crisis, the US government imposed stricter regulation and oversight on the financial sector by passing the [Dodd-Frank Act \(2010\)](#).

Regulations and standards alone are not enough without compliance and oversight. For instance, in 2018 and 2019, two Boeing aircrafts crashed, resulting in over 300 deaths. Investigations after the accidents revealed that American aircraft manufacturer Boeing had [not shared critical information about the risks associated](#)

[with its Maneuvering Characteristics Augmentation System \(MCAS\)](#), an automated flight control feature, with regulators or pilots. In the case of the crashes, the MCAS sprung into action based on erroneous sensor data, forcing the airplanes to dive uncontrollably towards the ground and crash. In response, the US tightened its aircraft licensing process and [oversight and accountability for manufacturers](#), with the [FAA adopting stricter policies](#) on aircraft design, production and quality control.

We can prevent major incidents of advanced AI causing public harm by embedding risk assessment frameworks into a regulatory framework, measured against policymaker's value-based definitions of what counts as safe enough, and what doesn't, and by ensuring compliance through independent oversight.

→ About the authors

EVA BEHRENS is a researcher in the Advanced AI team, focusing on international AI policy and governance.

BENGÜSU ÖZCAN is a researcher in the Advanced AI team, focusing on international coordination on AI governance and scenario planning.

Image credits: Cover image generated with AI Shutterstock (reference to Giorgio de Chirico)



FOR MORE INFORMATION PLEASE CONTACT:

Eva Behrens

Advanced AI Researcher - Policy

e.behrens@icfg.eu

International
Center for
Future
Generations **ICFG**