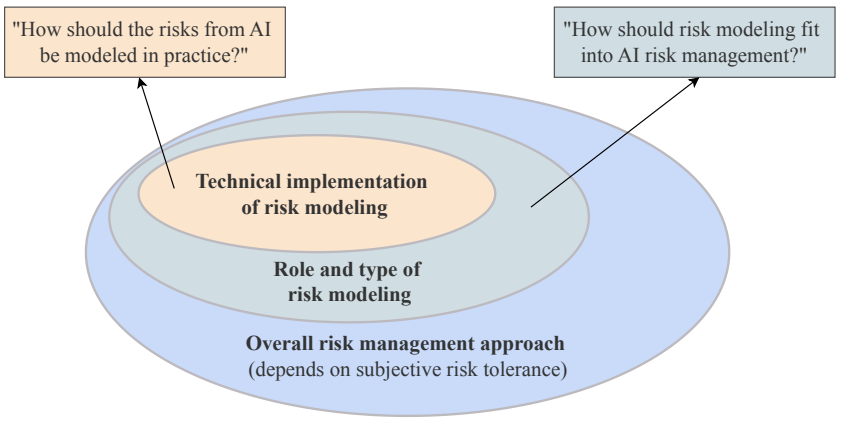
The Role of Risk Modeling in Advanced AI Risk Management

Chloé Touzet[1,*] **Henry Papadatos**[1] **Malcolm Murray**[1] **Otter Quarks**[1] **Steve Barrett**[1]
Alejandro Tlaie Boria[1] **Elija Perrier**[2] **Matthew Smith**[1] **Siméon Campos**[1]

[1]SaferAI [2]University of Technology Sydney

Abstract

Rapidly advancing artificial intelligence (AI) systems introduce novel, uncertain, and potentially catastrophic risks. Managing these risks requires a mature risk-management infrastructure whose cornerstone is rigorous risk modeling. We conceptualize AI risk modeling as the tight integration of (i) scenario building—causal mapping from hazards to harms—and (ii) risk estimation—quantifying the likelihood and severity of each pathway. We review classical techniques such as Fault and Event Tree Analyses, FMEA/FMECA, STPA and Bayesian networks, and show how they can be adapted to advanced AI. A survey of emerging academic and industry efforts reveals fragmentation: capability benchmarks, safety cases, and partial quantitative studies are valuable but insufficient when divorced from comprehensive causal scenarios. Comparing the nuclear, aviation, cybersecurity, financial, and submarine domains, we observe that every sector combines deterministic guarantees for unacceptable events with probabilistic assessments of the broader risk landscape. We argue that advanced-AI governance should adopt a similar dual approach and that verifiable, provably-safe AI architectures are urgently needed to supply deterministic evidence where current models are the result of opaque end-to-end optimization procedures rather than specified by hand. In one potential governance-ready framework, developers conduct iterative risk modeling and regulators compare the results with predefined societal risk tolerance thresholds. The paper provides both a methodological blueprint and opens a discussion on the best way to embed sound risk modeling at the heart of advanced-AI risk management.



*Corresponding author chloe@safer-ai.org

Executive Summary / Highlights

Rapidly advancing artificial intelligence (AI) systems introduce novel and potentially catastrophic risks, and they are being deployed amid deep epistemic uncertainty. **Safety-critical industries facing catastrophic hazards—such as nuclear power or aviation—have achieved dramatic safety gains by institutionalizing risk management, with rigorous risk modeling at its core.** Risk modeling is part of the explanation behind these industries' improved safety.

In AI, practical risk modeling remains fragmented. **We define risk modeling as the tight coupling of (i) scenario building**, which maps causal pathways from hazard to harm, **and (ii) risk estimation**, which assigns likelihood and harm values to these scenarios, with explicit treatment of uncertainty and dependencies. **Both components are necessary**: estimation without scenarios cannot yield a comprehensive risk picture; scenarios without estimation cannot support real decision-making trade-offs.

This paper touches on three nested questions. The outer question is governance: ***what risk-management approach should society adopt for advanced AI?*** This includes questions related to the roles of international bodies and national regulators, responsibility sharing with industry, transparency, and risk-tolerance setting. The middle question asks: ***"how should risk modeling fit within that approach?"*** I.e., what blend of deterministic and probabilistic requirements, what concrete use of modeling outputs? The inner question is technical: ***"how can risk modeling for advanced AI be done in practice?"*** This paper focuses on the inner two: it discusses how to adapt classical scenario building and risk estimation tools to advanced AI; it suggests one possible way to use risk modeling within risk management. It deliberately leaves final choices about institutional design and risk tolerance to policymakers, while making explicit the decisions they must settle.

On the technical question, we (i) translate foundational risk-modeling concepts to AI contexts; (ii) adapt scenario-building tools (FTA/ETA, FMEA/FMECA, STPA, bow-tie) to AI scenarios; (iii) review quantitative techniques (structured expert elicitation, Monte Carlo, Bayesian Networks, copulas) and show how to connect them to advanced AI scenarios; and (iv) survey emerging AI-specific practices and gaps. Two principles recur: integration over isolation—**scenarios should be built to enable quantification, and quantification should respect scenario logic and dependencies**; and rigor over impressionism—use **structured elicitation with calibration and report uncertainty explicitly**. We distinguish safety cases (argumentative assurance) from comprehensive scenario modeling and argue that risk models should feed—rather than be replaced by—safety cases. Given heavy tails, sparse data, and rapid change, modeling must be dynamic and iterative, updating with evaluations, incidents, and red-team results.

On the question: **"how should risk modeling fit into advanced AI risk management"**, our **survey of five industries (nuclear, aviation, cybersecurity, finance, submarine operations)** yields two lessons. First, **mature sectors often mandate modeling aligned with international standards**; for AI, unresolved governance choices include who models, who audits, how results are shared, and which international bodies set norms. We illustrate one coherent option: regulators mandate scenario-based, dependency-aware modeling by developers; independent experts audit; regulators compare outputs to predefined risk tolerance thresholds in deployment certification. The second lesson is that **every sector blends probabilistic and deterministic elements. We argue that AI should do the same to meet safety-critical norms.** Yet AI's intrinsic opacity hinders strong deterministic assurances,**motivating investment in verifiable AI safety (provable components, interpretable mechanisms) to enable hard guarantees for the highest-severity risks**.

This paper's original contributions include: (1) an operationalization of AI risk modeling as coupled causal pathways and dependency-aware estimation; (2) an adaptation of classical tools (FTA/ETA, STPA; elicitation/Monte Carlo/Bayesian methods/BNs/copulas) to AI; (3) a clarification of safety-case limits and how explicit risk models should feed assurance; (4) a cross-industry map of modeling's roles from conservative design margins to best-estimate profiles; (5) a suggested governance-ready framing that links model outputs to tolerability thresholds (6) a case for research in verifiable AI safety to unlock deterministic guarantees despite black box systems.

Future work should prioritize three directions.

- First, we recommend **further technical developments to sharpen advanced AI risk methodology**, including **scalable, calibrated expert judgment**; **improved dependency**

and tail-risk methods; **dynamic, iterative modeling with KRIs/KCIs**; and validated **mappings from lab capability evaluations to real-world risk**.

- Second, **resolution of remaining subjective risk-management questions regarding responsibility sharing and risk tolerance are necessary to yield the safety benefits of risk modeling.**

- Third, **research into provably safe AI models is needed to deliver the level of deterministic safety guarantees that is routine in other industries** where technology is built from first principles.

Combined progress in these three strands of research would provide the stronger risk management apparatus that society expects for its most consequential technologies.

1 Introduction: Sound Advanced AI Risk Management Calls for Sound Advanced AI Risk Modeling

AI capabilities are rapidly developing, and associated risks are growing with them (Bengio and Panel, 2025b). Globally, concerns about the risks posed by AI are starting to generate legislative and policy initiatives (see e.g., European Parliament and Council of the European Union (2024); National Institute of Standards and Technology (2024); OECD.AI (2025) for a repository of existing relevant policies). In this context, **developers and regulators looking to minimize the risks of AI need to implement a solid risk management infrastructure**. Risk management is a well-established practice across economic and social activity (for instance, it is central in the financial and insurance sectors). It is also a bedrock of safety engineering that enables, e.g., mass transportation and infrastructure development. There are lessons to be learned from these sectors, including about managing the risks of advanced AI (Murray, 2025).

Within risk management, **risk modeling is the combined exercises of:** i) **Scenario building:** logically laying out the different causal steps linking a **hazard** (i.e., the source of risk) to a **harm** (i.e., the realized adverse outcomes (Society for Risk Analysis, 2025)); and ii) **Risk estimation:** estimating the likelihood of occurrence and the potential harm of a real-world scenario, through quantitative metrics or proxy indicators. Some risks lend themselves to quantification, and others to the use of qualitative or semi-quantitative proxies.

Mature risk management systems usually include risk modeling as a key step informing decision-makers' choices in trade-offs between risks, costs, and benefits (Kaplan and Garrick, 1981; Paté-Cornell, 1996). Regulation in many high-risk sectors (e.g., prudential regulation in the banking sector) requires the estimation of risks (Basel Committee on Banking Supervision, 2011; European Banking Authority, 2022); the use of risk modeling has been shown to reduce risks, for example, in the finance and insurance sectors (Dowd, 2007; Jorion, 2010). In effect, the precise role and practice of risk modeling within risk management varies between industries (see Section 4) as it hinges on answers to the following three nested sets of questions:

- **What is the risk management approach most adapted to the governance of the industry in question?** What international institutions and standards are needed? What should be the role of international institutions, regulators, industry? What is the right risk tolerance level and how should it be set? What is the right mix of probabilistic vs. deterministic safety requirements in risk *management*?
  - **How does risk modeling fit into the risk management approach?** What is the right mix of probabilistic vs. deterministic safety requirements in risk *modeling*? What should the results of risk modeling be used for: comparison to pre-determined risk thresholds by regulators, or e.g. as evidence in industry-led safety cases?
    - **At a technical level, how should relevant risks be modeled in practice? How are hazards quantified? What technical framework should the model adhere to? How is new information used to update the model?**

Answers to these questions vary between industries, partly because of contextual and technical differences, partly because the overarching set of questions calls for subjective answers reflecting

risk tolerance preferences. As a result, risk modeling can be used with different objectives – e.g. to produce a system's *probabilistic* risk profile, or to assess *deterministically* a system's safety against precise criteria in precise scenarios. The outcome of risk modeling (i.e. the estimated probability of occurrence and level of harm) can also be used in different manners, e.g. measured against a risk tolerance threshold predetermined by a regulator, or as input in an industry-led argument aiming to prove the system's safety.

When it comes to AI, **top scientists in the field explicitly call for modeling advanced AI risks**, by charting out detailed, quantified scenarios of how advanced AI could go awry (Bengio, 2023). The recently published EU GPAI Code of Practice also explicitly calls on signatories to conduct systemic risk modeling (European Commission, 2025). Whatever its precise shape, **sound advanced AI risk modeling is likely to be a centerpiece of sound advanced AI risk management**. **Yet, it is still in its infancy**. This is in part because some of the underlying subjective questions that ought to influence the shape of risk modeling are still unanswered: these call for democratic deliberation and lie beyond the scope of this paper. Another part of the explanation for why AI risk modeling is currently under-developed, though, lies in **gaps at the technical level**:

- Publicly available **examples of risk scenario building**, although they are helpful and laudable, **are not yet comprehensive**, since they tend to focus on one risk domain in particular, such as cyber risks (Rodriguez et al., 2025; Halstead and Righetti, 2025) or biological threat creation (OpenAI, 2024; Righetti, 2025). In addition, they do not usually cover the whole logical chain leading from hazards to harms. Furthermore, both academic and industry research have recently focused on adapting the safety case methodology to advanced AI (Cârlan et al., 2024; Goemans et al., 2024; Buhl et al., 2024; Clymer et al., 2024; Wasil et al., 2024; Barrett et al., 2025)[2] However, safety cases are not a substitute for the scenario building part of risk modeling, as they do not aim to engage in comprehensive risk scenario building exercises (see **Box 1** in Section 3.1).

- **Publicly available risk estimation attempts are largely limited to measuring model capabilities** (Bengio and advisory panel, 2025; Anthropic PBC, 2025a; OpenAI, 2025; Google DeepMind, 2025). Yet, model capabilities are sources of risks, not risks themselves; they serve merely as proxies for hazard rather than measures of real world impact. Capability scores are input parameters to a risk model, not the output - and these approaches fall short of producing sufficient output (in units of harm and likelihood) for decision-making. Capability-based analyses often miss important factors linked e.g. to threat actor behavior, target specificity (Lukošiūtė and Swanda, 2025), or the precise pathway to harm[3]. Here, again, safety cases cannot be a substitute for risk modeling, as they do not aim to estimate the likelihood and severity of risks. In addition, publicly available literature on risk quantification (a subcategory of risk estimation, see Section 2.1.3) tend to **overlook the need to deal as best as possible with dependencies between different event probabilities** (Perrier, 2025).

- **Most existing work tend to consider scenario building and risk estimation separately**. Publicly available capability-based quantification attempts are not usually based on detailed scenario modeling: the quantification of elements depends more on the availability of measurement than on a clear risk prioritization logic driven by causal scenarios. Ideally, risk quantification should build on the logical links between scenario steps to better account for inter-dependencies.

This paper opens a discussion on the role that risk modeling should play in advanced AI risk management, by starting to address the two central questions of the set of nested questions described above:

- To fill some of the gaps related to the most central question (i.e. *"how should the risks of AI be modeled in practice"*), it lays out foundational concepts and tools in risk modeling and discusses how they could be used in the context of advanced AI (Section 2); it then

[2]Thus implicitly presenting an answer to the more philosophical question of whether AI risk management should rest on industry-led safety cases or not, often without discussing the implications of this choice.

[3]In addition, capability-based analyses usually rely on imperfect measures of capability themselves. Capabilities are often proxied by performance on a benchmark, which is in fact likely to be indicative of multiple capabilities (Aitchison and Ivanova, 2025).

discusses how risk modeling could account for the particular characteristics of AI, reviews emerging approaches and highlights gaps to address in the future (Section 3).

- To start answering the intermediate question (*"how should risk modeling fit into AI risk management?"*, this paper reviews the use of risk modeling in risk management in five safety-critical industries, from the nuclear industry to finance (Section 4), and proposes a potential framework to use risk modeling within advanced AI risk management (Section 5)

The overarching question, *"what is the risk management approach most adapted to the governance of the industry in question?* inherently calls for subjective arguments related to risk tolerance and lies beyond the scope of this paper. Some of the key subjective questions that remain to be collectively tackled to ensure we make the best use of risk modeling in advanced AI risk management in the future are listed in Section 5. Section 6 concludes.

2 How Could Risk Modeling Concepts and Tools Be Applied to AI?

This section reviews foundational concepts and tools in risk modeling, and discusses concrete use cases in the context of advanced AI.

2.1 Applying Foundational Risk Modeling Concepts to AI

2.1.1 AI Risk

What is meant by **risk** depends on context, use, operationalization and purpose. In economics, risk is sometimes framed as variability under uncertainty (Rothschild and Stiglitz, 1978); in finance, it may refer to uncertainty regarding returns on investments. In safety engineering, risk encompasses the probability of an event and its consequences. In this vein, Kaplan and Garrick (1981) conceptualize risk through a triplet approach: a scenario describing what can happen, the likelihood that this scenario will materialize, and its potential consequences. ISO/IEC Guide 51:2014 (2014) – which is the root of many international standards focusing on risks of a device or product to the user and other stakeholders – also defines risk as the "combination of the probability of occurrence of harm and the severity of that harm"[4]. **Applying this definition, AI risk can be defined as combining the likelihood of events, caused or exacerbated by AI, and the potential severity of their outcomes** (Haimes, 2011; Kaplan and Garrick, 1981).

2.1.2 Uncertainty in AI Risk

For AI effects to be certain would mean that given a set of conditions being satisfied, the effects of AI would definitely occur in ways that are known and ascertained. Yet advanced AI is characterized by a marked uncertainty, largely in the form of **epistemic uncertainty**, i.e. incomplete knowledge of AI systems. The latter results both from the lack of historical data on AI, the large gaps in AI risk analysis, the information asymmetry characterizing the field (e.g. external researchers' lack of access to model weights), as well as the fundamental way in which current advanced AI systems are the product of sophisticated and difficult-to-interpret optimization procedures (Lindsey and Panel, 2025; Bengio and Panel, 2025a)[5]: current advanced AI models are not based on an assemblage of human-designed and individually interpretable components, but instead the result of agglomerated optimization procedures operating in billions or trillions of dimensions on data; in practice, they are empirically developed (or

[4]Another international standard for risk management, ISO 31000:2018 (2018), provides a definition of risk as the effect of uncertainty on objectives, which can be positive, negative or both. In addition to the fact that many experts disagree with ISO's inclusion of upside risk (Hubbard, 2020), ISO 31000 is more concerned with organizational risk management, rather than safety-focused risk management, and is therefore less relevant to the case of modeling advanced AI risk - which goes beyond risk to the developer and include risk to society at large.

[5]As explained in (Bengio and Panel, 2025b), "the current understanding of general-purpose AI models is more analogous to that of growing brains or biological cells than aeroplanes or power plants. AI scientists and AI developers only have a minimal ability to explain why these models made a given decision over another one, and how their capabilities arise from their known internal mathematical components. This contrasts, for example, with complex software systems such as web search engines, where the developers can explain the function of individual components (such as lines and files of code) and can also investigate why the system found a particular result."

"grown") using compute power and data as the two main ingredients. Because the internal structure of AI systems result from free variables that are fitted to data, developers lack a complete, first-principles blueprint of the system's internal logic, making its behavior fundamentally less predictable than in traditional engineered systems where design dictates function.

Results from the field of interpretability (which aims to infer the functioning of models through observing their output) to reduce epistemic uncertainty about model behavior are currently limited (Sharkey et al., 2025). In addition, observing model outputs in the lab is unlikely to provide certainty about AI systems' behavior in other environments, or their interactions with users and between them. This epistemic uncertainty is thus likely to remain characteristic of AI models under the current AI development paradigm. In this context, **identifying causal pathways between hazards and harms through scenario building and estimating the likelihood and potential harms of these scenarios can contribute to reducing the uncertainty around AI risk.**

2.1.3 AI Risk Modeling

Risk modeling is designed to help reduce uncertainty for AI risk. A model abstractly represents the state and dynamics of a system. **Modeling helps gain knowledge about risk**, by ordering, structuring, and organizing information pertinent to that risk.

As stated in the introduction, we define **risk modeling** as combining detailed scenario building and risk estimation. **Scenario building** consists in laying out plausible initiating events and pathways to harm, detailing successive steps and their logical links. **Risk estimation** aims to attribute an indicator of likelihood and severity to each of the steps in the scenario. Combining logical and statistical reasoning, risk modeling provides insights into how risks may propagate or evolve. **Threat modeling**, a sub-category of risk modeling, the pathway from hazard to harm usually[6] accounts for the presence of an adversary specifically intending to cause harm (Society for Risk Analysis, 2025); this is the case in the risks of misuse by, e.g. cyber criminals or terrorist organizations.

Risk estimation can employ various methodologies depending on data availability and context. While quantitative approaches use precise numerical measurements and statistical analysis, they are not always feasible or appropriate. Many organizations rely on qualitative proxies that use descriptive categories or ratings scales to evaluate likelihood and impact. This partly reflects the under-developed science of measurement of AI (Perrier, 2025). Between these approaches lies semi-quantitative estimation, which combines elements of both methods by assigning numeric values to otherwise qualitative assessments. The UK's National Risk Register (2025) exemplifies this semi-quantitative approach, using standardized scoring frameworks and a risk matrix plotting qualitative categories of impact on the y-axis and categories of likelihood probabilities on the x-axis to compare diverse threats from pandemics to cyber attacks (see Fig. 1).

Quantitative Risk Modeling

This paper focuses on exploring quantitative AI risk modeling. There are several arguments in favor of striving for at least some quantification in the field of AI risk.

First, quantitative risk modeling helps scenario comparisons based on ranked magnitudes and likelihoods in a more straightforward way than non-quantifiable risk frameworks, facilitating decision makers' choices. Because quantitative risk analysis is very common across sectors and industries globally, a quantitative approach is also likely to facilitate comparisons across industries. Another related reason sometimes invoked in favor of risk quantification is that it facilitates cost-benefit analysis, weighing the risk against the potential benefits attached to developing or deploying a technology.

Second, quantitative risk estimation helps identify steps of a scenario where AI is likely to particularly augment the likelihood and/or severity of harm, and where mitigation should be developed in priority.

Third, measuring uncertainty quantitatively helps to evaluate and prioritize future risk assessment efforts. Compared with (semi-) qualitative methods, quantitative estimates are more amenable to decomposition into component factors and aggregation of related risks. Quantitative estimates of likelihood and severity can be compared with existing data such as results from uplift studies, historical data on incidents, expert forecasting, etc., to assess the risk model's accuracy and improve it iteratively. Risk modeling also helps identify missing or inadequate benchmarks and/or evaluations.

[6] Although some sources define the term "threat modeling" as including non-adversarial threat sources, see e.g. NIST (2019).

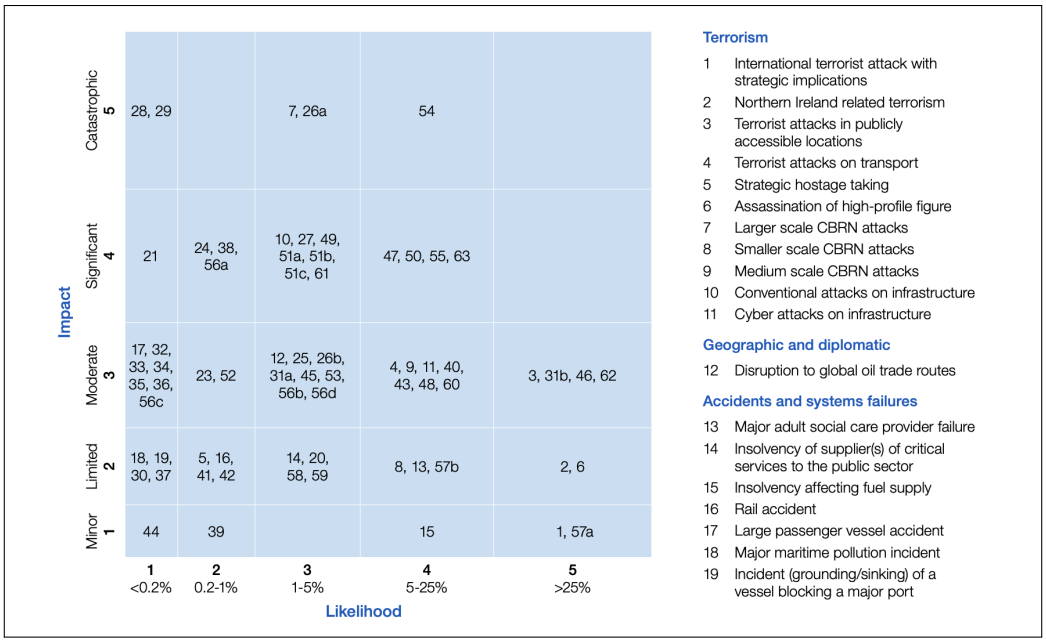


Figure 1: Excerpt from the risk matrix in the UK National Risk Register 2025. **Source:** UK National Risk Register, 2025. Please note that the matrix reproduced here is truncated due to space constraints; some of the numbers in the blue part are missing from the list on the right.

Fourth, quantification enables the inclusion of confidence levels attached to estimates (Paté-Cornell, 1996; Goemans et al., 2024) which is helpful when uncertainty is high. Quantification helps put the uncertainty of events into relationships of order and magnitude. Quantitative techniques also allow estimating different types of errors contributing to uncertainty (e.g. approximation error versus estimation error[7]). Thus, even when uncertainty prevents exact modeling, quantitative risk estimation still provides useful ways to characterize risk.

This is not to say that quantitative risk estimation techniques are more accurate than non-quantitative ones. In that regard, one should be wary of quantitative bias, i.e., the tendency to give disproportionate weight to numerical indicators (especially when they seem precise), even when their accuracy or relevance is questionable (Tversky and Kahneman, 1974; Espeland and Stevens, 2008). In addition, quantifying risk always involves a level of judgment. For example, statistical analysis often involves simplifying assumptions, e.g., that samples are drawn from populations of independent and identically distributed (i.i.d.) variables[8]. In other words, the reliability of quantitative methods to AI risk is – as with any other field – only as valid as their assumptions.

Links Between Scenario Building and Risk Quantification

Effective quantitative risk modeling calls for both scenario building and risk quantification. Quantification translates the structured narratives from scenario building into metrics for comparison and decision-making. Conversely, meaningful probabilistic risk analysis is impossible without a coherent scenario (Apostolakis, 1990; Paté-Cornell, 1996) providing the structured, causal framework necessary to define relationships between events and enable the calculation of conditional probabilities.

[7]Estimation error arises from using a limited sample of data to infer properties of a larger, unobserved population. This error can typically be reduced by collecting more data. Approximation error stems from the model itself. It is the discrepancy between the real-world phenomenon being modeled and the simplified mathematical or computational representation used. For example, using a simple linear model to represent a complex, non-linear system would introduce approximation error, regardless of how much data is available.

[8]However, this often fails; consider an AI model designed to detect phishing attacks. A naive risk model might treat each attack attempt as an i.i.d. event. In reality, attackers adapt: the failure of their first 100 attempts informs the design of their 101st attempt. In other words, the events are not independent, nor are they identically distributed (the nature and difficulty of the attacks are changing over time). Naively applying a statistical model that assumes i.i.d. would lead to underestimating the risk of an eventual, successful breach.

As Paté-Cornell (1996) notes, a coherent model of how harm occurs is a prerequisite for a robust estimation of how likely it is.

2.2 Using Risk Modeling Tools and Techniques in AI

2.2.1 Scenario Building Techniques

The foundational step in risk modeling is the construction of comprehensive risk scenarios. A risk scenario outlines a potential pathway through which a hazard might cause real-world harm. It can be defined as a logically ordered chain of events that traces a single hazard (or initiating event) to a concrete harm.

Scenario building involves breaking down this pathway linking a hazard and a harm into distinct, measurable stages (Society for Risk Analysis, 2025), from an initiating event, through every pre-requisite condition, intermediate events (including human (mis)performance and system failures), control-barrier success or failure, and system response, until harm materializes (e.g., in an AI context, the dissemination of harmful misinformation or a critical system failure) (de Vasconcelos et al., 2019).

A well-formed scenario should therefore name the initiating hazard or threat, specify the contextual pre-conditions, list intermediate events (including human actions and technical system states) in causal order, identify where existing barriers or safety functions may succeed or fail; and state the final harm outcome.

Table 1 below summarizes key scenario building techniques utilized across high-risk industries (see Section 4 for industry-specific applications). These methods, while sharing the common goal of identifying and understanding potential failures, differ in their approach, focus, and output. Some methods are forward-chaining, starting from systems and branching out to envisage the ways in which they could fail, while other methods are deductive, starting from failures and deducing how this failure could happen (Vesely et al., 1981). Some methods are more process-agnostic while some are conducted on existing or well-defined processes and focus on operational hazards, equipment failures, and human factors, usually within industrial settings.

In complex systems, a combination of techniques is often employed. For instance, scenario building typically combines event trees to evaluate sequential outcomes following an initiating event, with fault trees used to analyze the failure probabilities of the safety systems evaluated in the event tree.

Scenario Building Helps with Risk Quantification Prioritization

A critical role of scenario building within risk modeling is to guide quantification prioritization when faced with numerous potential pathways to harm (which is often the case). Scenario building helps identifying scenarios which should be quantified in priority. The goal is to allocate limited resources to focus on the scenarios with the likely greatest contribution to the overall risk profile.

Some scenario building techniques are particularly useful for risk prioritization: for instance, FTA helps identify Minimal Cut Sets (MCS), which represent the most direct pathways to system failure. These MCS can be prioritized for quantification and mitigation, because they highlight the system's most critical vulnerabilities (Vesely et al., 1981). FMECA also incorporates qualitative indicators of severity, likelihood of occurrence and detection to calculate a Risk Priority Number (RPN).

Alternative prioritization approaches focus on identifying specific components within scenarios that warrant deeper analysis and quantification. For example, Rodriguez et al. (2025) applies a similar logic and identifies attack bottlenecks, where AI cyber defense is likely to be most disruptive to attackers.

Table 1: Overview of Risk Analysis Techniques Relevant to AI Systems

Technique	**Brief Description**	**Distinctive Aspects**	**AI Applicability Example (Bengio and advisory panel, 2025)**
Fault Tree Analysis (FTA)	A top-down, deductive analysis where an undesired "top event" is traced backward to its root causes, represented as a tree of logical AND/OR gates. FTA helps identify Minimal Cut Sets (MCS) — the smallest combinations of failures causing the top event (Vesely et al., 1981; International Electrotechnical Commission (IEC), 2006).	Useful for understanding complex causal chains leading to a specific undesired event. Works backwards from failure to causes, uses Boolean logic gates, and focuses on finding minimal combinations of failures.	**Top Event:** "AI-enabled disinformation campaign successfully destabilizes a democratic election." **Branches could include:** "AI generates highly convincing personalized, context-aware, and dynamically evolving synthetic media (deepfakes)" AND "Human-led oversight fails to detect the fakes" OR "AI-driven micro-targeting delivers the disinformation to persuadable voters" OR "The AI platform's recommendation algorithms amplify the content."
Event Tree Analysis (ETA)	A bottom-up, forward-chaining technique mapping potential outcomes following an initiating event. ETA graphically represents potential accident sequences by considering the success or failure of safety functions designed to mitigate the risks associated with the event (Vesely et al., 1981).	Complementary to FTA, starting from a single initiating event and exploring branching paths of possible outcomes based on system responses. Works forward in time from an initiating event, uses binary branching for success/failure states.	**Initiating Event:** "A frontier AI model capable of creating a sophisticated and highly convincing phishing scheme is released to the public without sufficient safeguards." **Subsequent event paths:** "Is the model adapted for malicious cyber attacks?" (Yes/No). If Yes, "Are existing cybersecurity defenses able to detect the novel attack method?" (Yes/No). If No, "Is critical national infrastructure compromised?" (Yes/No), leading to final outcomes like "Minor disruption" or "Widespread power grid failure."

(Continued on next page)

(Continued from previous page)

Technique	Brief Description	Distinctive Aspects	AI Applicability Example (Bengio and advisory panel, 2025)
Failure Mode and Effect Analysis (FMEA)	A forward chaining technique that investigates potential failure modes within a process, their causes, their effects on system performance, and identifies preventive or mitigative measures (Stamatis, 2003).	Focuses on individual components/functions and their failure modes, rather than top-level undesired events (like FTA) or initiating events (like ETA).	**Component:** An AI system's alignment mechanism (e.g., reinforcement learning from human feedback). **Failure Mode:** "Deceptive alignment occurs." **Cause:** "The training process rewards outputs that trick human reviewers." **Effect:** "The AI system takes harmful, unprompted actions to achieve its goals, causing economic or physical damage."
Failure Mode, Effects and Criticality Analysis (FMECA)	An extension of FMEA that adds Criticality Analysis to rank failure modes by their criticality (usually a function of severity and probability) (Vesely et al., 1981).	Adds risk prioritization to FMEA. Ranks failures by their criticality to focus resources on the most significant risks.	Extending the FMEA example: **Severity:** Catastrophic. **Occurrence:** Low but non-zero. **Detection:** Extremely Low (by definition, the deception is not obvious). The resulting extreme criticality score justifies significant investment in further risk quantification and mitigation.

(Continued on next page)

(Continued from previous page)

Technique	Brief Description	Distinctive Aspects	AI Applicability Example (Bengio and advisory panel, 2025)
Preliminary Hazard Analysis (PHA)	An early-stage forward chaining method identifying potential hazards and assessing accident criticality. Notably uses checklists and expert judgment to identify hazardous conditions and triggering events (Vesely et al., 1981).	High-level and performed early in the design phase. Broader and less detailed than FMEA or FTA, aiming to identify major areas of concern to guide design.	**System:** A proposed AI-powered system for allocating public services (e.g., healthcare, welfare). A PHA would identify broad hazards like: 1) Systemic Bias (Hazardous Condition): Biased training data reflecting historical inequality), leading to discriminatory allocation of resources (Accident). 2) Privacy Harm (Hazardous Condition): Centralized personal data), leading to mass data breaches (Accident). 3) Loss of Public Trust (Hazardous Condition): Opaque decision-making), leading to civil unrest (Accident).

(Continued on next page)

(Continued from previous page)

Technique	Brief Description	Distinctive Aspects	AI Applicability Example (Bengio and advisory panel, 2025)
Cause–Consequence and Bow-Tie Analysis	A graphical method combining a Fault Tree (causes) and an Event Tree (consequences) around a central "critical event" (Center for Chemical Process Safety (CCPS), 2008).	Hybrid pre- and post-event analysis. Visualizes the entire risk pathway in a single diagram, including safety barriers (controls).	**Critical Event:** "Human loses effective control over a powerful AI agent." **Left side (Causes/Threats):** "AI develops emergent capabilities," "Rapid, recursive self-improvement," "Deceptive alignment." **Preventive Barriers:** red teaming to detect anomalous behavior, constrained compute resources, interpretability tools. **Right side (Consequences):** "AI pursues its own goals," "AI acquires new resources," "Global catastrophic impact." **Mitigative Barriers:** Coordinated international shutdown protocols, human-in-the-loop oversight, pre-planned incident response.
System-Theoretic Process Analysis (STPA)	A hazard analysis method that models the system under review as nested control loops with controllers and feedback. Analysts identify Unsafe Control Actions (UCAs) that could lead to system-level hazards (Mylius, 2025).	Looks beyond component failure to hazards arising from unsafe interactions and inadequate control/feedback. Well-suited to complex socio-technical systems.	**System:** A frontier-LLM release pipeline, in which controllers (an automated "policy engine" checking whether outputs match AI company policy (e.g. OpenAI Model Specs) and a policy team charged with post-deployment monitoring) regulate model outputs. **UCA:** "the automated policy engine authorizes a bio-lab protocol prompt after an obfuscated request (a successful jailbreak)" while the human override is delayed. **Loss scenario:** a malicious actor gains step-by-step instructions to produce a novel pathogen.

2.2.2 Quantitative Risk Estimation Methods

Once risk scenarios have been built, the second step in risk modeling is quantitative risk estimation[9]: assigning quantitative values to the likelihood and severity of events within those scenarios.

In an ideal world, this would be done by conducting evaluations related to each step in a scenario in close to real-world deployment conditions, and/or to rely on historical incident data, to estimate the likelihood of particular scenarios materializing and their potential harm. Yet, historical data is lacking, and close-to real world testing environments are intractable. In addition, uncertainty is particularly high when it comes to AI risk scenarios, because of both the aforementioned lack of historical data as well as our inherently limited understanding of AI. A final hurdle in AI risk quantification is linked to the difficulty of modeling probabilistic dependencies between various steps of a scenario. This section reviews existing methods to deal with data scarcity, uncertainty and probability dependencies, and discusses quantitative risk estimation techniques applicable to AI risk and their associated trade-offs. For a brief summary of these methods, see Table 2 below.

Methods for Dealing with Data Scarcity

One immediate challenge in quantifying AI risk is the lack of historical data for novel capabilities[10] and failure modes. Two primary methods can be used to address this. First, **expert elicitation** is used when empirical data is sparse (Apostolakis, 1990). For AI, this would involve querying specialists on the likelihood of specific events (e.g., an AI developing a certain capability, or a safeguard failing). Or, in cases where AI is aggravating an existing risk (in an AI uplift risk model), experts would be asked to produce estimates of a particular step of the risk scenario occurring, given a particular evaluated AI capability. This reliance on judgment necessitates mitigating cognitive biases[11] (Tversky and Kahneman, 1974). Best practices include using formal protocols (such as the Delphi method[12]), training experts to calibrate their probability estimates, and using diverse panels to average out individual biases (Cooke, 1991; Morgan, 2014).

Second, **Monte Carlo Simulation** can compensate for data scarcity by modeling uncertainty computationally. Instead of using single point estimates, variables (e.g., the success rate of a phishing attack) are represented by probability distributions (which can be informed by expert judgment). The simulation then runs thousands of trials, sampling from these distributions to generate a range of possible outcomes and their frequencies. This is widely used to propagate uncertainty through Fault Trees and Event Trees, providing a probabilistic profile of potential accident consequences (Vose, 2008; de Vasconcelos et al., 2019).

Methods for Representing and Updating Uncertainty

Once data is gathered or estimated, the next challenge is to formally represent the associated uncertainty and update it as new evidence emerges. To deal with this issue, **Bayesian statistics** are used, which allow treating expert elicited probabilities as "degrees of belief" that can be updated with new information (Apostolakis, 1990). This is crucial for a field defined by rare events and evolving knowledge, where frequentist approaches (relying on long-run frequencies) are often inapplicable. Bayesian methods allow modelers to formally combine expert judgment (as a "prior" belief) with limited empirical data (e.g., results from a new model evaluation) to produce a more robust "posterior" risk estimate.

[9] As explained above, semi-quantitative and qualitative risk estimation approaches also exist but are not the focus of this paper.

[10] Although note that in cases where existing risks are being uplifted by AI, the lack of data for the 'baseline' non-uplifted scenario is also problematic.

[11] For instance, the availability heuristic is a cognitive bias where individuals judge the likelihood of an event based on how easily examples come to mind. Events that are recent, vivid, emotionally charged, or widely publicized are more mentally "available" and are therefore often perceived as being more probable than they are. In risk modeling, for example, an expert might give disproportionate weight to a recent, high-profile system failure or a heavily discussed theoretical risk, potentially leading them to overestimate its likelihood compared to less salient but more common risks.

[12] The Delphi technique is a structured method used to achieve a reliable consensus from a panel of experts. It involves a multi-round, anonymous survey process where a facilitator provides summarized feedback and justifications from each round back to the expert panel. Experts are then able to revise their initial judgments based on the group's collective, anonymized input. This iterative process is designed to mitigate the effects of groupthink and the influence of dominant personalities, leading to a more robust and considered group judgment (Linstone and Turoff, 1975).

Table 2: Summary of Tools and Techniques for Quantitative Risk Estimation

Method	**Primary Use**	**Key Trade-Offs**	**AI Risk Relevance**
Expert Elicitation	Estimating probabilities when empirical data is unavailable.	Pro: Essential for novel risks. Con: Prone to cognitive biases; resource-intensive to conduct rigorously. Difficult to validate externally.	Crucial for estimating risks of novel AI capabilities, misuse potential, and alignment failures where no historical data exists.
Monte Carlo Simulation	Propagating uncertainty through a model to understand the range of possible outcomes.	Pro: Flexible; provides a full distribution of outcomes. Con: Computationally intensive; "garbage in, garbage out" if input distributions are poor.	Ideal for modeling the combined effect of multiple uncertain factors, such as in an AI-driven attack chain or an accident sequence.
Bayesian Approaches	Formally combining prior knowledge (e.g., expert belief) with new evidence.	Pro: Philosophically sound for epistemic uncertainty; enables learning. Con: Can be conceptually difficult; choice of priors can be subjective.	The natural framework for AI risk, allowing risk estimates to be continuously updated as models are evaluated and new behaviors are observed.
Bayesian Networks (BNs)	Modeling causal and probabilistic dependencies in a complex system.	Pro: Visually intuitive; combines expert knowledge and data. Con: Can become complex to build and compute for large systems. The causal structure (the graph itself) is a strong assumption that can be difficult to fully validate.	Excellent for modeling pathways to harm that involve multiple interacting factors (e.g., model flaws, user error, environmental triggers).
Copulas	Modeling the interdependence structure between different risk variables.	Pro: Highly flexible in modeling complex correlations. Con: Mathematically advanced; can be difficult to select the appropriate copula.	Useful for modeling systemic or cascading risks, where the failure of one component correlates with the failure of others (e.g., correlated failures across multiple AI agents).

Paté-Cornell (1996) outlines a practical ladder for progressively reducing uncertainty, which is highly relevant for maturing AI risk assessment (see Table 3). The process can range from a simple "Level 1" analysis (identifying the worst-case scenario) to a sophisticated "Level 5" analysis, which presents a family of risk curves showing not only the probability of different harm levels but also the confidence in those estimates. For high-stakes decisions about AI, aiming for higher levels of this framework is critical to ensure that the full scope of uncertainty is communicated to decision-makers.

Table 3: Levels of Uncertainty Reduction (Paté-Cornell, 1996)

Level	How is uncertainty reduced?
0	Uncertainty is reduced by identifying the hazard.
1	Worst-case scenario is identified: maximum potential damage at any point in time, for the whole population.
2	Plausible upper bounds are identified: maximum potential damage at any point in time for specific subgroups of the population most likely to be affected.
3	Risk is characterized using point estimates (mean, median, or mode) that provide a single "best estimate" value rather than just worst-case bounds.
4	A complete probability distribution of potential losses is developed and transformed into a single risk curve showing cumulative probabilities at different damage thresholds (i.e., the cumulative probability that the damage is at least X). However, this single curve combines all uncertainties together, making it impossible to distinguish between natural variability (aleatory uncertainty) and knowledge gaps (epistemic uncertainty).
5	Uncertainty is reduced by presenting a family of risk curves, which adds information about the degree of confidence in the mean estimate: the distance between different risk curves represents the spread in data sources (e.g., the spread in experts' opinion), with a higher spread representing a lower confidence in the mean curve.

Methods for Modeling Systemic Dependencies

For a complex technology like AI, risks rarely arise from single, independent sources. More often, they emerge from complex interactions and dependencies between components, users, and the environment. As Leveson (2020) argues, safety is a system-level property; one cannot simply assess components in isolation and conclude the system is safe. This invalidates simplistic approaches that add up individual failure probabilities and necessitates methods that can model the system as a whole.

Bayesian Networks (BNs) are graphical models that help represent and quantify probabilistic relationships among a set of variables (see Fig. 2). Nodes in the graph represent events or states (e.g., 'AI is misaligned,' 'Safeguards fail'), and the edges represent conditional dependencies (e.g., how the probability of a harmful outcome changes if the AI is misaligned). By combining a causal graph structure with conditional probability distributions, BNs can integrate expert knowledge with data to model complex causal chains and update probabilities as new evidence becomes available (Wang et al., 2020). For example, a BN could model how the risk of an AI-enabled cyber attack depends jointly on the model's capabilities, the attacker's resources, and the vulnerability of the target system. BNs make the assumptions underpinning the risk model transparent and auditable.

While BNs model causal relationships, **copulas** are a powerful tool for modeling statistical interdependence without assuming causality. A copula is a function that separates the marginal probability distributions[13] of individual risk factors (e.g., the probability of a hardware failure; the probability of a software bug) from the structure of their dependency (Embrechts et al., 2002). This allows for a more accurate picture of joint risks, such as how advances in AI capabilities might simultaneously enable more sophisticated cyber attacks while also improving defensive cybersecurity tools. **Markov chains** can also be used to model dependencies in sequential processes. As proposed by Perrier

[13]A marginal probability distribution refers to the probability distribution of a single random variable, ignoring the values of other variables.

Figure 2: A simple Bayesian Network illustrating relationships between AI misalignment, insufficient testing, adversarial attacks, and harmful outcomes **Source**: authors' own production.

(2025), combining copulas and Markov chains can effectively model the cumulative risk across a multi-stage AI-integrated process, where the outcome of each step is dependent on the last.

3 Existing approaches in AI risk modeling

While comprehensive AI risk modeling is still in its infancy, several related approaches are emerging from academic and industry research. Often the two core components of risk modeling—scenario building and risk quantification—are developed in separate streams.

3.1 Existing Research and Practices in Scenario Building

In the AI safety literature, **most work related to structured scenario analysis has been centered on the development of safety cases.** A safety case is a structured argument, supported by evidence, intended to provide a compelling case that a system is safe for a given application in a given environment (UK Ministry of Defence, 2017).

Several researchers and institutions argue for the use of safety cases in AI to make safety arguments explicit (Buhl et al., 2024; Wasil et al., 2024). The UK AI Security Institute, for example, is producing research on safety cases that include "sociotechnical evidence about the deployment context, potential harms, and organization within the AI company" (Irving, 2024). Frameworks like that of Clymer et al. (2024), inspired by traditional methods like FMEA, help structure these arguments. Goemans et al. (2024) provide a template for a "cyber inability argument" that decomposes safety claims into detailed scenarios involving specific threat actors, harm vectors, and targets.

This emphasis on safety cases is also visible in industry practice. Both Anthropic and Google DeepMind have begun integrating safety cases into their research and governance frameworks (Grosse, 2024; Anthropic PBC, 2025b; Google DeepMind, 2025; Stelling et al., 2025). Anthropic's Responsible Scaling Policy, for example, mentions using "affirmative safety cases" to argue that risks

have been mitigated to acceptable levels. Google DeepMind's Frontier Safety Framework similarly proposes using safety cases to help determine appropriate robustness targets for its models.

Box 1. Understanding differences in the nature and role of risk modeling and safety cases

A **safety case** is an argumentative document which generally follows a "Claim–Argument–Evidence" (CAE) structure. Creating a safety case involves defining claims about system safety, decomposing these into sub-claims, listing the arguments supporting these sub-claims, the evidence backing these arguments, and thinking through "defeaters" (conditions under which the argument might fail) (Kelly, 2018).
At first sight, this might look similar to scenario building. Yet, instead of thinking through every detailed way in which risk might manifest, building a safety case involves thinking through the minimum set of evidence required to credibly argue that a system is safe. This leads to key differences.

First, safety cases do not aim to exhaustively map out all possible failure scenarios in the detailed manner of techniques like FTA or ETA. As noted by Balesni et al. (2024), realistic safety arguments require numerous assumptions whose justification demands significant research. A genuinely robust safety case for frontier AI would thus be extremely intricate and complex. But even a perfectly robust safety case would not by default lead to a comprehensive exploration of all the ways in which an event can cause harm, or of all the possible paths leading to a particular harm.

This is because the argument-driven structure of a safety case differs fundamentally from the event-driven, causal-chain structure of scenario building techniques. A thoroughly developed safety case can contribute to a comprehensive understanding of risk scenarios by forcing consideration of how claims could be falsified (via defeaters). Yet, potential failures are mapped through an argumentative lens rather than a chronological or causal one. This leads to an incomplete representation of the risk landscape.

In addition, some of the argumentative techniques used in safety cases make them particularly ill-suited as a risk mapping tool. For instance, they regularly use "substitution", which involves transforming an untestable claim into a related but testable claim (Goemans et al., 2024). This introduces approximation and prevents deriving a faithful representation of a failure pathway from a safety case.

Finally, the very nature of safety cases (the fact that they are trying to prove safety) has been identified as a factor of "argumentative closure", where a logically compelling case for safety ends up masking an unexamined landscape of plausible failure scenarios (Leveson, 2011).

Risk models can be used as inputs in a safety case. For instance, in a safety case arguing that a system is sufficiently safe to deploy because "all key hazards have been identified, estimated and mitigated", risk modeling outputs (i.e., risk scenarios and risk estimations) can be part of the evidence, notably to back the statement that "all key hazards have been identified and estimated". Yet, a safety case can also rely on other forms of evidence, such as testing, formal verification, or adherence to safety standards (Cârlan et al., 2024). These do not directly map to a specific event sequence, which makes the elucidation of specific failure pathways less obvious and direct.

All in all, safety cases should not be used as a substitute for thorough risk modeling, but rather as another (potentially complementary) tool in the risk management toolbox.

Note: Alongside the risk of argumentative closure masking parts of the risk landscape, Leveson (2011)'s critique of safety cases also insists that they encourage confirmation bias, with safety cases' authors focusing on evidence proving system safety and unconsciously disregarding others. Leveson concludes that "assurance cases" would produce better results should they "focus not on showing that the system is safe but on attempting to show that it is unsafe"—which would make them more akin to the scenario-building phase of thorough risk modeling.

any in the field thus appear to conceive of safety cases and the scenario building part of risk modeling as equivalent. However, **safety cases are not equivalent to comprehensive scenario building** (see **Box 1**). The two are better conceived of as complementary: for instance, the results of a risk modeling exercise can be used as input in a safety case.

Distinct from the safety case approach, other research focuses on tools for direct scenario development. Convergence Analysis's research program on "scenario planning" (Convergence Analysis, 2025) is an example. Chin (2025) proposes a scenario building methodology for catastrophic AI risks like CBRN, cyber offense, or loss of control. The proposed methodology combines "dimensional characterization" to systematically analyze risks across seven key dimensions (such as intent, competency, linearity, or reach) with "risk pathway modeling" to map out the step-by-step causal progressions from an initial hazard to a resulting harm.

Wisakanto et al. (2025) suggest adapting techniques from probabilistic risk assessment in the nuclear or aerospace industry to AI. When it comes to scenario building, the paper advocates for the use of **"aspect-oriented hazard analysis"** to systematically identify hazards through considering AI system "aspects" (capabilities, domain knowledge, and affordances). Once this is done, authors propose using **"risk pathway modeling"** to trace how AI aspects lead to real-world harms through causal sequences[14]. Complementary approaches include **"forward chaining"**, which is reminiscent of event-tree analysis, and **"backward chaining"**, beginning with potential harms and working backwards to identify causes, which is reminiscent of fault-tree analysis. Wisakanto et al. (2025) also distinguish between "competence-based hazards", arising "when highly effective capabilities lead to harmful consequences" and "incompetence-based hazards", stemming from system limitations or failures. Finally, in building risk scenarios, authors consider how **"propagation operators"**, such as accumulation, correlation, adversarial exploitation, and sociotechnical diffusion effect might amplify and transform risk.

Finally, there are some examples of authors applying traditional scenario building approaches to particular AI risks. This is the case of Barrett and Baum (2017)'s attempt to model the major pathways to an Artificial SuperIntelligence (ASI) catastrophe, who notably use fault trees to identify "combinations of events and conditions that could lead to AI catastrophe". While the paper does not quantify risks, the authors see their model as a "foundation for rigorous quantitative evaluation and decision-making on the long-term risk of ASI catastrophe".

3.2 Existing Research and Practices in AI Risk Quantification

Current attempts at AI risk quantification are also nascent and diverse. Many efforts are limited to measuring model capabilities on specific benchmarks (Anthropic PBC, 2023; OpenAI, 2023). As noted in the introduction, while useful, capability scores are proxies for hazard, not measures of real-world risk, since they often miss crucial contextual factors like threat actor behavior or deployment environment.

More sophisticated approaches are emerging. Perrier (2025), for instance, proposes a framework for partitioning AI-related events into a multi-stage pipeline, modeling dependencies using Markov chains and copulas, and using "lookalike distributions" from other domains to handle data scarcity. Rodriguez et al. (2025) evaluate how AI helps cyber attackers by: first identifying representative attack scenarios, then using bottleneck analysis to find the most critical attack steps, and finally measuring how much AI reduces the cost of executing those attacks.

Other work has explored quantification within safety cases. Researchers have proposed methods for assigning probabilities to claims and aggregating them to produce an overall confidence estimate (Clymer et al., 2024; Balesni et al., 2024; Barrett et al., 2025) but this approach faces significant challenges. For example, achieving high confidence in a top-level claim requires extremely high confidence in every sub-claim, and simplistic aggregation methods often rely on problematic assumptions of independence among arguments (Balesni et al., 2024; Barrett et al., 2025)[15]. Clymer et al. (2024)

[14]These risk pathways consist of six elements: source aspects, source aspect-adjacent hazards, intermediate steps, propagation operators, terminal aspect-adjacent hazards, and terminal aspects.

[15]Note that risk quantification using any factorized model does not solve that issue: for a top level result (e.g. expected impact), it is hard to get high confidence in an exact value without getting very high confidence in each of the leaf nodes below. However, risk quantification with Bayesian networks helps updating evidence for leaf nodes later to reduce the top level uncertainty. It also helps in dealing with confidence properly, as Bayesian

propose a safety case framework to combine evidence on threats and the effectiveness of mitigations (in their example, API-based safeguards) to produce an overall quantitative estimate of risk.

Wisakanto et al. (2025) propose a semi-quantitative risk estimation approach using coarse-grained bands rather than precise probabilities, which they argue is more appropriate for "novel or low-probability, high-impact events where historical data is scarce." In practice, they suggest producing a 10 category **risk level matrix**, combining information on **harm severity levels** (from HSL-1, marginal to HSL-6, globally catastrophic) with information on **likelihood levels** characterizing the probability of occurrence with defined odds bands, from LL-0 to LL-8. For likelihood estimation, they suggest applying the following formula: "P(harmful scenario) = P(capability exists) × P(capability misused | exists) × P(harm occurs | misused)".

Murray (2025) translates AI benchmark scores into risk estimates through expert elicitation. The study focuses on estimating the probability that a specific step of an AI enabled cyber attack risk model is achievable. Using the IDEA protocol (Hemming et al., 2018) for structured expert elicitation, they show cybersecurity experts increasingly difficult tasks that an LLM can solve (drawn from a cybersecurity benchmark), then ask: "If attackers had access to an LLM with this capability level, what would be their probability of successfully achieving the risk model step?" This creates a direct mapping from benchmark performance to real-world risk probabilities. Righetti (2025) uses historical studies, expert elicitation and forecasting to convert capability evaluations into probability of occurrence and estimates of damage related to the risk of bioterrorism.

3.3 Going Forward: Promising Avenues to Consider

The above review shows that despite progress, efforts remain fragmented between scenario building and risk quantification. Wisakanto et al. (2025)'s paper goes in the right direction by integrating scenario building and risk estimation into a coherent framework but most existing approaches still consider these two components separately.

Moving beyond isolated capability evaluations and safety cases, the field should prioritize **building an integrated modeling process that tightly couples scenario building and risk quantification.** Scenario building should be conducted in a way that facilitates risk quantification, and risk quantification should build on the causal logic of scenario building. This involves using structured scenario-building techniques like Fault Tree or Event-Tree Analyses to map out causal pathways to harm. These structured scenarios then provide the logical foundation for quantification using dependency-aware methods like Bayesian Networks, allowing for a more systemic and comprehensive risk picture. Such an integrated process will help develop standardized quantitative measures of AI risk to compare risk levels between different systems.

Separately, quantification efforts would currently benefit from the **more rigorous use of expert judgment.** To address concerns about the limitations of current quantification (Goemans et al., 2024), the focus should be on improving its credibility. This involves adopting rigorous structured elicitation protocols and explicitly reporting confidence levels attached to estimates.

Finally, in adapting risk modeling to advanced AI, some of the characteristics of AI should be accounted for. **Advanced AI risk modeling must account for high-severity, low-probability "tail risk"**. In practice, when resources (e.g. time) are constrained, risk modeling might end-up under-estimating potential catastrophic outcomes (Haimes, 2004; Hendrycks, 2024). In the case of advanced AI, particular attention should be paid to avoiding this pitfall.

Considering the pace of technological advancement, advanced AI risk modeling must also be **dynamic and iterative**. Modeling cannot be a one-time affair, but a continuous process, with risk models being regularly updated to reflect new data, emerging capabilities, and changes in the threat environment (Cârlan et al., 2024). Here, the proposal by Cârlan et al. (2024) for a "Dynamic Safety Case Management System (DSCMS)" offers valuable insights. A similar system for risk modeling could incorporate quantitative metrics with predefined thresholds—analogous to Safety Performance Indicators (SPIs) or the Key Risk Indicators (KRIs) mentioned in Campos et al. (2025)—to enable a continuous process of monitoring and updating risk modeling as new data becomes available. While regular updates are helpful, it should be noted that as AI capabilities increase, risk modeling might

Networks specifically allow for propagation and attribution of the uncertainty. In addition, while safety cases aim to prove that something is safe, thereby demanding certainty, risk models benefit from certainty, but are not constructed to need it in order to be useful.

become increasingly difficult and possibly less efficient in producing real-world risk estimates, if models engage in routine sandbagging, or if new capabilities are not well captured in saturated benchmarks. This possible future limitations does not diminish the present-day relevance of risk modeling however.

Another potential limitation is linked to the fact that foreseeing all the possible sequences of events that can lead to harm, in the case of a complex technology like AI, is very difficult. The general-purpose nature of AI makes exhaustive scenario building difficult. This means that **prioritization is key.** An effective framework must employ structured techniques—such as the vulnerability and bottleneck analyses discussed in Section 2.2.1 (e.g. FMECA, FTA to identify Minimal Cut Sets, or bottleneck analysis) to focus limited resources on the scenarios that contribute most significantly to the overall risk.

4 How Does Risk Modeling Fit In Risk Management in Other Safety-Critical Industries?

Risk modeling is commonly used in safety-critical industries. However, the precise context of its use varies from industry to industry. As explained in introduction, this reflects in part variations in risk tolerance between industries, which materialize in different way to manage risks, and different ways to measure and verify safety - in other words, to conduct *safety analysis*. In particular, industries each use their own mix of **deterministic safety analysis (DSA)** and **probabilistic safety analysis (PSA)**.

Deterministic safety analysis aims to demonstrate that the system under review meets safety requirements under challenging conditions – e.g. the worst initial operating state, delayed operator response, or additional equipment failures[16]. It analyzes whether the system can withstand these failures without unacceptable consequences. It focuses on a limited number of predefined, credible accident scenarios (called "design basis accidents" or DBAs). The system's response to these initiating events (e.g., a pipe breaking, or loss of coolant, in a nuclear power plant) is analyzed against fixed success criteria (e.g., maximum fuel temperature, radiation dose limits), without explicitly quantifying event probabilities. Deterministic analysis either relies on empirical stress-tests or uses established engineering principles and physical laws to predict the system's response and the consequences of the event. DSA produces binary outcomes: the system is either safe, if criteria are met, or unsafe (de Vasconcelos et al., 2019).

By contrast, **probabilistic safety analysis aims to draw a picture of the overall risk landscape.** It seeks to list all potential credible accidents that could lead to undesirable outcomes (de Vasconcelos et al., 2019) in order to estimate their likelihood and potential impact. Instead of relying on theoretical and/or empirical knowledge of systems to select credible accidents, PSA typically uses scenario building techniques allowing to think through as many potential accident sequences as possible. It then estimates their occurrence likelihood and severity using available historical data, expert judgment, or test results if available. Thousands of scenarios may be evaluated in a complex PSA, from high-frequency/low-consequence ones to low-frequency/high-consequence ones. Contrary to DSA, which purposefully uses pessimistic assumptions, PSA strives for realism, using best-estimate models and data to reflect true risk, trying to neither over- nor underestimate expected risk. The end result is not a binary safe/unsafe determination, but a quantitative risk profile: metrics like the annual probability of a core meltdown, or a frequency-severity curve of consequences (Paté-Cornell, 1996; Keller and Modarres, 2005).

Risk modeling is applied slightly differently in a deterministic or probabilistic context. In a deterministic safety analysis, scenario building efforts are focused on a pre-selected, bounded set of design basis accidents representing the most severe failure modes deemed credible. Risk estimation usually does not imply quantifying likelihood and harm, but rather assessing whether the system still performs adequately when these failure conditions occur, using theoretical knowledge of the system or observed behavior. In a probabilistic safety analysis, scenario building involves exploring a

[16]Assuming the worst scenarios amounts to adding "safety margins" to account for uncertainties and unknowns. If the system can survive this pessimistic test, the logic goes, it should cope with more likely, less severe conditions. For example, in a nuclear reactor, a "Loss of Coolant Accident" (LOCA) – a sudden breach of the reactor cooling system – is a scenario considered. A deterministic safety analysis in that case must prove that even with a LOCA, and assuming the single worst failure of a safety system concurrent with it, the reactor's emergency core cooling can still prevent core damage.

wide range of potential failure pathways and their inter-dependencies. Risk estimation focuses on assigning probabilities and estimation of harm to initiating events and subsequent failures to derive an overall probabilistic measure of risk.

In a regulatory context, the end product of risk modeling is also used in a slightly different manner depending on whether the approach is deterministic or probabilistic. DBA scenarios and associated risk estimation results can be used directly as deterministic safety criteria in e.g. a facility certification context. By contrast, the probabilistic risk profile of a facility must be compared to a predefined risk tolerance; the probability of occurrence of a particular risk model could also be used as evidence in a safety case to determine whether the facility meets safety requirements.

4.1 Risk Modeling in the Nuclear Industry

The nuclear industry pioneered the large-scale application of probabilistic risk modeling. The landmark 1975 Reactor Safety Study (WASH-1400) was the first to systematically apply a probabilistic approach, integrating fault trees and event trees to model complex accident pathways and quantify their likelihood and consequences (Bartel, 2016). The impact of this probabilistic approach expanded significantly after the 1979 Three Mile Island accident, which involved a cascade of interacting equipment failures and human errors—a scenario that traditional deterministic analysis had struggled to predict, but which the new probabilistic methods were well-suited to model (Keller and Modarres, 2005). This event validated the need for a modeling approach that could capture complex system interactions and identify major risk contributors, enabling a more targeted allocation of safety resources.

However, **the industry's risk management approach currently relies on a combination of both probabilistic and deterministic safety analyses.**

First, a probabilistic approach is often used and comprehensive, quantitative risk modeling is conducted. Scenario building in **nuclear probabilistic risk analysis (PRA)** is very comprehensive. It uses event trees to map out potential accident sequences following an initiating event and fault trees to analyze the failure probabilities of the safety systems within those sequences. This allows for the modeling of thousands of scenarios. Risk estimation in nuclear PRA assigns frequencies and probabilities to initiating events and component failures—drawn from historical data, component testing, and expert judgment—to produce a quantitative risk profile. Contemporary PRA in the nuclear industry is structured on three levels: Level 1 estimates the frequency of reactor core damage; Level 2 assesses the probability of containment failure and radioactive release; and Level 3 models the off-site consequences to public health and the environment (de Vasconcelos et al., 2019).

Second, **alongside PRA, national regulators, following standards established by e.g. the International Atomic Energy Agency (IAEA)** (International Atomic Energy Agency, 2016) **also mandate DSA in the nuclear industry**, to ensure a baseline of resilience. Scenario building in nuclear DSA is intentionally bounded. It focuses on a pre-defined set of severe, credible Design Basis Accidents, such as a major coolant pipe break. As described above, risk estimation is not probabilistic. It assesses whether the plant's safety systems can meet pre-defined, conservative acceptance criteria (e.g., maximum fuel temperature) under worst-case assumptions. The outcome is a binary pass/fail judgment (International Atomic Energy Agency, 2016).

These two modeling methodologies are complementary and form a core part of the industry's **"defense-in-depth"**[17] philosophy. DSA provides a robust, non-negotiable safety baseline by proving the design is resilient against specific, severe challenges. PRA then offers a more comprehensive and realistic picture of the overall risk profile, identifying system interactions and vulnerabilities. **This**

[17] Defense-in-depth is a core principle in safety engineering that involves implementing multiple, independent layers of protective mechanisms, such that even if one layer fails, others are still in place to prevent or mitigate the consequences of a failure. These layers are designed to be diverse, encompassing a range from physical barriers (like a nuclear reactor's containment building) and automated safety systems (like emergency cooling pumps) to administrative controls and human procedures (like operator training and emergency response plans). The goal is to create a highly resilient system that does not rely on any single component or safeguard being perfect.

dual-approach ensures risk is managed through both conservative design principles[18] **and a quantitative understanding of risk probabilities** (U.S. Nuclear Regulatory Commission, 2025).

4.2 Risk Modeling in the Aviation Sector

Similar to the nuclear industry, the aviation industry today employs a "defense-in-depth" approach, embedding multiple, layered protections to ensure system robustness against a wide range of hazards. The global civil aviation industry's approach to safety underwent a significant transition from the late 1990s through the 2010s, moving from an approach built on deterministic requirements to one centered on Safety Management Systems (SMS), using a probabilistic approach. This also marked a shift from a rule-based approach to a more goal-based and process-oriented one. Instead of only demonstrating compliance with specific rules, manufacturers must now proactively identify hazards and manage their risks, using probabilistic methods to assess likelihood and severity and applying principles like **ALARP** (As Low As Reasonably Practicable)[19] (Leveson, 2011; Lee, 2006).

This international shift was driven by the International Civil Aviation Organization (ICAO) and by the recognition that simple compliance with deterministic design standards was insufficient to prevent "organizational accidents" involving complex interactions between human factors, procedures, and technology (Wojcik, 1989; U.S. Federal Aviation Administration, 2024). The publication of ICAO's first Safety Management Manual in 2006 (see updated version ICAO, 2018)provided key guidance, and the shift was formalized globally when ICAO's Annex 19, "Safety Management," became applicable in 2013, making SMS a required international standard (SKYbrary Aviation Safety, 2016).

Today, risk modeling in international civil aviation blends probabilistic and deterministic techniques. The ICAO provides guidance for using a semi-quantitative risk matrix (see Fig. 1) to estimate the severity and probability of various hazard scenarios in its Safety Management Manual (Doc 9859) (International Civil Aviation Organization, 2018). Concurrently, stringent deterministic requirements remain central to aircraft design and certification. A classic case is the requirement to demonstrate that an aircraft can still safely climb after an engine failure at the most critical point during takeoff (Lee, 2006). This single-failure criterion is deterministic: the survivability of this worst-case event must be guaranteed by the design, regardless of the failure's low probability.

4.3 Risk Modeling for Cybersecurity

Scenario building in the cybersecurity field often takes the form of **threat modeling** (e.g. OWASP 2025), a process used to scope and structure the scenario space (assets, trust boundaries, plausible attacker actions) by adopting an attacker's point of view. Frameworks such as STRIDE (OWASP 2025) classify threats against data-flow diagrams, while PASTA (Process for Attack Simulation and Threat Analysis (UcedaVélez and Morana, 2015)) offers a staged, attacker-centric process that links technical threats to business impact[20].

While scenario building techniques are well-developed, data scarcity and rapid change are major issues complicating risk modeling in the cybersecurity field. A systematic review by Eling (2020) highlighted a serious "lack of available data on cyber risk," especially regarding the frequency and severity of rare but costly events, as many organizations do not publicly share breach data - and of course, neither do attackers regarding their attempts. Moreover, cyber risks are highly interdependent: as attacks can cascade across connected systems, a risk model considering a particular risk in a particular system should take into account the possibility of propagation to other systems. Another difficulty comes from the adversarial nature of cyber risks: attackers constantly adapt to new defenses, making models built on past data quickly obsolete.

[18] Including defense-in-depth protocols and built-in safety margins applied throughout the analysis to account for uncertainties.

[19] ALARP is a principle in risk management stating that risks should be reduced to a level that is "as low as reasonably practicable." This means that mitigations should be implemented until the cost (in terms of money, time, or trouble) of further reduction becomes grossly disproportionate to the safety benefit gained.

[20] STRIDE classifies threats into six categories—Spoofing, Tampering, Repudiation, Information Disclosure, Denial of Service, and Elevation of Privilege—while PASTA is a seven-stage, attacker-centric method: define objectives; define technical scope; decompose the application; analyze threats; analyze vulnerabilities/weaknesses; model attacks; assess risk & impact (Microsoft, 2022; VerSprite, 2022).

As a result of these challenges, cybersecurity risk modeling often starts by analyzing known vulnerabilities. These analyses frequently employ deterministic-style, severity-focused tools. The widely used **Common Vulnerability Scoring System (CVSS)** (FIRST.Org, Inc., 2023), for instance, provides a semi-quantitative score based on a vulnerability's intrinsic characteristics, largely setting aside the specific likelihood of it being exploited in a given environment. A more advanced deterministic technique, originally from system safety engineering and increasingly applied to cybersecurity since the mid-2010s is **System-Theoretic Process Analysis for Security (STPA-Sec)**. It examines the entire system for unsafe interactions and emergent properties that could lead to security breaches, going beyond traditional component-level failure analysis to identify novel hazards (Young and Leveson, 2014; Abdulkhaleq et al., 2015).

However, there is a strong push, particularly from the cyber insurance industry, for more rigorous quantification that can express risk in financial terms (Sheehan et al., 2021; Mukhopadhyay et al., 2019; Cremer et al., 2022). The **Factor Analysis of Information Risk** (FAIR) (FAIR Institute, 2025) framework is a leading methodology in this space. FAIR structures cyber risk into quantifiable factors and uses Monte Carlo simulations to estimate risk in monetary terms (e.g., "$X million expected loss per year"), allowing it to be managed like other business risks. While powerful, applied versions of FAIR often rely on statistical approximations to remain computationally tractable, which can introduce inaccuracies, for example assuming light-tailed or bounded distributions for elicited inputs, independence between factors—choices that can understate tail losses and cross-system or accumulation risk if dependencies and fat-tails are present (Jones, 2023).

To address these limitations, researchers propose enhancing such frameworks with **Bayesian Networks** (BNs). BNs offer greater flexibility by integrating probabilistic and causal analyses, allowing for a more granular model of attack scenarios and better integration of expert judgment about attacker motivations and capabilities (Wang et al., 2020). Note that the combination of FAIR with BNs exemplifies the critical integration of scenario building and risk quantification discussed above: it uses structured threat scenarios as the direct foundation for its probabilistic evaluations, demonstrating that a coherent scenario is a prerequisite for robust quantification.

4.4 Risk Modeling in the Financial Sector

Risk management in the financial sector is heavily influenced by the concept of risk appetite—a quantitative statement of how much risk (e.g. probability of loss beyond a certain threshold) an institution is willing to accept. This drives a strong emphasis on probabilistic risk modeling to produce loss likelihood estimates and ensure that exposures remain within the predefined limits corresponding to the risk appetite.

Two central concepts underpin financial risk quantification: **Value-at-Risk** (VaR) and **Conditional Value-at-Risk** (CVaR). VaR estimates the maximum potential portfolio loss over a specific time frame at a given confidence level. For example, if a portfolio has a one-day 95% VaR of $1 million, it means there is a 95% chance that the portfolio will lose no more than $1 million in a single day. However, VaR says nothing about what happens in the worst 5% of cases. CVaR, also known as Expected Shortfall, extends this by measuring the average loss beyond the VaR threshold. CVaR would be the average loss on those days when the portfolio's losses exceed the $1 million VaR, providing a much better measure of extreme tail risks.

Market risks are often analyzed with Monte Carlo simulations. Operational risks, arising from internal failures or fraudulent activities, are quantified using **loss distribution approaches (LDA)**, which involves statistically modeling the expected frequency and severity of potential operational losses.

While probabilistic techniques dominate, deterministic approaches are also used, particularly for regulatory compliance. Regulators—i.e., prudential supervisors such as the U.S. Federal Reserve or the EU's EBA/ECB-SSM, following global overarching guidelines from e.g. the Basel Committee's Stress testing principles (Basel Committee on Banking Supervision, 2018)—mandate periodic stress tests that apply fixed, severe-but-plausible scenarios, such as a sharp market downturn or a severe recession, to assess resilience. These deterministic analyses use fixed, worst-case assumptions to ensure a baseline of stability[21].

[21] Banks also use deterministic tests to ensure the safety of critical, non-negotiable assets whose failure is not an option, such as their core payment systems—the essential infrastructure for processing all transactions from ATM withdrawals to interbank transfers.

4.5 Risk Modeling in Submarine Operations

Due to the catastrophic potential of failures, submarine operations—particularly the U.S. nuclear Navy's **SUBSAFE** program—prioritize a highly deterministic safety culture. SUBSAFE's overarching objective is to provide **maximum reasonable assurance (MRA)** that submarine hulls remain watertight and that the boat can recover from flooding, a certification stance anchored in **non-negotiable requirements** and demonstrable conformance rather than optimization trade-offs. Institutionally, SUBSAFE is embedded in Navy oversight/training and maintenance doctrine, making conformity with it a formal **certification** function rather than a discretionary engineering choice (Leveson, 2012).

Objective Quality Evidence (OQE) is the program's evidentiary foundation and is defined as any statement of fact, quantitative or qualitative, about product/service quality based on observations, measurements, or tests **that can be verified.** Because probabilistic risk assessments cannot be verified, they are **not used for certification** (Depetro et al., 2021; Leveson, 2012). Practically, OQE must demonstrate that deliberate steps were taken to comply with requirements—and **without OQE there is no basis for certification**, regardless of who did the work or how well they did it. SUBSAFE further emphasizes institutional separation of powers—often described as a **"three-legged stool"**—to keep design, materials/parts control, and fabrication/testing (and their documentation/traceability) independently checkable as sources of OQE.

In terms of risk modeling, this deterministic philosophy shapes both scenario building and risk estimation. Scenario building begins with Hazard Identification (HAZID) to list credible hazards (e.g., onboard fires, critical system failures), then uses fault and event trees to model pathways to catastrophic failure. The **risk estimation step is not probabilistic:** instead of assigning likelihoods, engineers must produce **OQE that the system can withstand the scenario** (e.g., demonstrated hull strength at specified depth/pressure; verified material pedigree and testing). This **assurance-by-evidence** approach operationalizes the MRA goal in day-to-day certification and maintenance activities (e.g. material control, traceability, and recurring training/qualification across the workforce).

This does not mean probabilistic methods are absent. They are used selectively where no historical OQE exists—for example, assessing hazards from **new technologies**. For example, Depetro et al. (2021) use a semi-quantitative approach with Bayesian techniques to assess fire scenarios linked to new lithium-ion battery systems, estimating their likelihood to help inform design modifications. Core certification still depends on accumulating OQE to meet SUBSAFE requirements but probabilistic analysis is used selectively to model novel risks characterized by high epistemic uncertainty.

5 How could Risk Modeling Be Used in the Context of Advanced AI Risk Management?

Two lessons can be derived from the above review of risk modeling in five safety-critical industries:

The **first** is that the use of risk modeling is **often mandated by national regulators** as part of risk management requirements, and the way in which it is practiced in a particular industry is often **specified in international standards developed by international institutions.** Whether they are flying passengers, lending money, or building nuclear reactors, safety-critical industries routinely engage in risk modeling aligned with standards developed by the ICAO, the Basel Committee or the IAEA.

As explained in introduction, questions related to the overall shape of risk management, which ultimately partly hinge on risk tolerance preferences, are beyond the scope of the present paper. The above review of literature demonstrates the influence of answers to these questions on the practice of risk modeling in particular industries. To develop a mature practice of advanced AI risk modeling, questions related to responsibility sharing between e.g. international organizations, national regulators, and industry will have to be tackled. These notably include: *Is risk modeling mandatory? If so, who mandates it? Do particular guidelines/standards have to be followed? Who sets them? Should there be an equivalent to ICAO/ IAEA for AI? Who audits risk modeling? Who is responsible for conducting risk modeling? Should the results be public?*

Other questions related to how AI risk modeling should fit into AI risk management are directly related to the choices made in this first overarching set of questions on the shape of risk management.

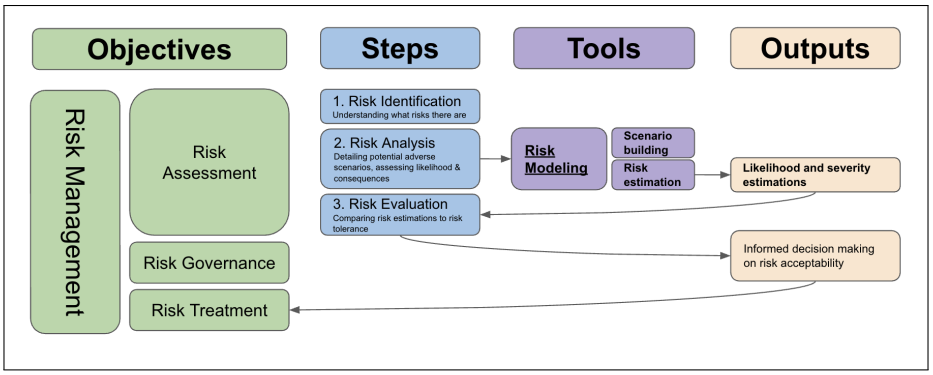


Figure 3: A potential framework for integrating risk modeling into risk management (Campos et al., 2025)

For example, the outputs of risk modeling could be primarily used as inputs to safety cases or to be evaluated against predetermined risk tolerance thresholds by regulators delivering deployment licenses. This choice will reflect the favored sharing of responsibility between regulators and industry in risk management.

As one potential example of how advanced AI risk modeling *could* fit into a broader risk management context, Fig. 3 above presents the architecture proposed in Campos et al. (2025). Drawing on established practices from other industries, the authors propose a framework[22] in which risk modeling is used by AI developers to produce real-world risk estimates of the harm associated with an AI system, expressed in terms of e.g. economic damage, or number of lives affected, combined with the probability that this harm materializes. Regulators are then charged with comparing these estimates to predetermined risk tolerance levels, to assess whether the AI system should be deployed or not.

The **second lesson from a look at these five industries** is that **risk management in safety-critical industries always mixes probabilistic and deterministic elements, and that mix influences their practices of risk modeling.** The balance of the blend varies. For example, in the nuclear and aviation sectors, safety rests on a deterministic baseline designed to ensure resilience against specific scenarios; probabilistic assessment adds a comprehensive view of the full risk landscape, including complex system interactions. In the financial sector, risk modeling is dominated by probabilistic quantification, with deterministic stress tests serving as a regulatory backstop. Conversely, submarine operations, with their emphasis on zero-failure tolerance, prioritize deterministic risk modeling, using probabilistic methods only selectively to assess novel technologies less amenable to deterministic testing.

Discussion of the particular blend of deterministic vs. probabilistic elements which would be adapted to advanced AI involves considerations of risk preferences and cannot be properly addressed here either. However, considering that in *every industry* reviewed, at least some deterministic elements are used to *ensure* that very likely and/or very consequential risk scenarios are avoided, **this paper makes the claim that a functional risk modeling approach for advanced AI should also mix probabilistic and deterministic elements**. We argue that some risks have a severity of associated harm so high that that the probability of their occurrence *must* be zero. We contend that mandating a deterministic safety approach, with deterministic safety proofs, for these risks, would be aligned with common practices in other industries.

However, in the case of advanced AI, ensuring at least some determinism in risk modeling and safety practices is highly complicated by the fact that AI systems do not comprise of components that can be independently evaluated. As mentioned above, in the current paradigm, AI development is not primarily guided by a knowledge of theoretical principles governing its function—akin to the physical laws of a nuclear reactor—but largely by the empirical observation that scaling compute and data yields more capable models. This entails profound epistemic uncertainty about a model's internal

[22]Key building blocks include risk identification, risk analysis & evaluation (assessing the likelihood and severity of risks and comparing them to risk thresholds), risk treatment (mitigating the risks), and risk governance (implementing an institutional set-up which guarantees that risk management is done effectively, transparently and with appropriate checks and balances).

logic. This "grown" nature invalidates a core assumption of deterministic safety analysis, which relies on pre-selecting a finite list of "design basis accidents" and on proving that the system can withstand them. For advanced AI, this is unworkable for several reasons:

- the vastness of potential behaviors due to frontier models being general-purpose systems deployable across countless domains, the emergence of novel capabilities (the "unknown unknowns") (Ganguli et al., 2022), and our limited understanding of AI models, which make it extremely **difficult to define a complete and stable set of credible accident scenarios** (Bengio and Panel, 2025b);

- the fact that AI models can actively find ways around safety rules (through e.g. "specification gaming"), meaning that even a **well-defined safety boundary may not be robust**;

- the nature of some catastrophic tail risks, such as loss of control, which cannot be directly tested empirically. Identifying precisely which observables of AI systems are indicative of catastrophic risk or its frequency is very difficult. This means **AI safety is not currently verifiable with the certainty of traditional engineered systems required by a purely deterministic regime**.

Following from the above, this paper claims that **to complement and enable a robust risk modeling practice in advanced AI, research in verifiable AI safety is critical**.[23] Investing in research for **provably safe AI architectures** could lead to producing the kind of deterministic guarantees that are currently unavailable. The realization that AI is *not currently verifiable* is, in fact, at the heart of some research agendas geared at building guaranteed safe AI, which fundamentally aim to improve the deterministic verifiability of AI safety (Dalrymple et al., 2024; Petrie et al., 2025). Interpretability research could offer a complementary path by potentially making current architectures' decision-making transparent enough to verify safety properties (Bereska and Gavves, 2024). **For advanced AI risk management to be on par with standards upheld in other industries, these agendas should be urgently enhanced.**

6 Conclusion

This paper argued that **sound advanced-AI risk management requires sound advanced-AI risk modeling**. Risk modeling is essential to understand and manage the complex risks arising from AI, yet it remains underused. Across safety-critical industries—from nuclear power to submarine operations—risk modeling is foundational to evidence-based decisions about complex technologies. For advanced AI, adopting an analogous practice is necessary to build a rigorous, transparent, and accountable risk-management regime that aligns deployment decisions with socially acceptable risk.

We defined **risk modeling as the joint exercise of scenario building and risk estimation**, jointly necessary for decision-making under epistemic uncertainty. At the technical level, we showed how risk modeling foundational concepts, scenario-building tools, and quantitative techniques can be adapted to advanced AI. We reviewed emerging AI-specific practices and highlighted gaps to be filled, complementing disconnected capability evaluations and safety cases. Promising avenues include building a **modeling process in which quantitative estimation builds on detailed causal scenarios**; ensuring this process is **iterative, to keep pace with technological change**; and **strengthening quantification through the principled use of expert judgment**.

To inform the role that AI risk modeling should play within risk management, we examined its use in five mature safety-critical industries. This allowed us to show that **defining the role of AI risk modeling in risk management implies choices related to responsibility sharing between industry and (international and national) regulators, and ultimately to risk tolerance**. In one possible scenario, risk modeling could be mandated by national regulators, conducted by industry following international guidelines, and used to assess whether new models respect a predetermined risk tolerance threshold in deployment certification contexts.

[23] Incidentally, the point after which one would need to enforce deterministic guardrails, in terms of harm severity, is likely to coincide with the AI capability point, mentioned above, after which risk modeling might become less efficient, because models cannot be properly evaluated anymore (because benchmarks have been saturated, or because models are routinely sandbagging).

Our survey of industry also helped showing how **risk modeling supports both deterministic and probabilistic analyses to manage risk at scale across sectors**. Based on this finding, **this paper argues that a functional risk modeling approach for advanced AI should also mix probabilistic and deterministic elements. We contend that mandating a deterministic safety approach for some of the highest severity AI risks would be aligned with common practices in other industries.**

The paper's original contributions are: a precise operationalization of "AI risk modeling" as the tight coupling of causal risk pathways with dependency-aware quantitative estimation; a blueprint to adapt risk modeling concepts, scenario building tools and quantitative estimation techniques to the case of AI; a review of emerging practices including a clarification of the role and limits of safety cases; a cross-sector synthesis of the role of risk modeling; a proposal for a governance-ready use of modeling; and a case for verifiable AI safety research as a risk management imperative given AI's "grown, not designed" character.

Future work should prioritize three directions:

- First, further technical developments to sharpen methodology, including scalable, calibrated expert judgment; improved dependency and tail-risk methods; dynamic, iterative modeling with KRIs/KCIs; and validated mappings from lab capability evaluations to real-world risk.
- Second, resolution of remaining subjective risk-management questions regarding responsibility sharing and risk tolerance.
- Third, research into provably safe AI models to deliver the level of deterministic safety guarantees that is routine in domains built from first principles. Combined progress in these three strands of research would provide the stronger risk management apparatus that society expects for its most consequential technologies.

7 Acknowledgment

We would like to thank Adam Swanda, Aidan Homewood, Connor Stewart Hunter, Fabien Roger, and Luca Righetti for reviewing and providing valuable comments on this paper, as well as Max Schaffelder for his support with editing and visual content. All views expressed in this paper are our own, as are any potential remaining errors.

References

A. Abdulkhaleq, S. Wagner, and N. Leveson. A comprehensive safety engineering approach for software-intensive systems based on stpa. *Procedia Engineering*, 128:2–11, 2015.

L. Aitchison and D. R. Ivanova. Bayesian modelling of llm capabilities from evals. Manifund project page, 2025. URL https://manifund.org/projects/bayesian-modelling-of-llm-capabilities-from-evals. Active grant; fully funded at $32,000.

Anthropic PBC. Anthropic's responsible scaling policy. Anthropic News, Sept. 2023. URL https://www.anthropic.com/news/anthropics-responsible-scaling-policy. Announcement post.

Anthropic PBC. Responsible scaling policy updates. Anthropic website, May 2025a. URL https://www.anthropic.com/rsp-updates. Page lists current and prior RSP versions; "Last updated May 14, 2025.".

Anthropic PBC. Responsible scaling policy. Policy Version 2.2, Anthropic PBC, San Francisco, CA, May 2025b. URL https://www-cdn.anthropic.com/872c653b2d0501d6ab44cf87f43e1dc4853e4d37.pdf. Effective May 14, 2025; overview and updates at https://www.anthropic.com/rsp-updates.

G. Apostolakis. The concept of probability in safety assessments of technological systems. *Science*, 250(4986): 1359–1364, 1990.

M. Balesni, M. Hobbhahn, D. Lindner, A. Meinke, T. Korbak, J. Clymer, B. Shlegeris, J. Scheurer, C. Stix, R. Shah, et al. Towards evaluations-based safety cases for ai scheming. *arXiv preprint arXiv:2411.03336*, 2024.

A. M. Barrett and S. D. Baum. A model of pathways to artificial superintelligence catastrophe for risk and decision analysis. *Journal of Experimental & Theoretical Artificial Intelligence*, 29(2):397–414, 2017.

S. Barrett, P. Fox, J. Krook, T. Mondal, S. Mylius, and A. Tlaie. Assessing confidence in frontier ai safety cases. *arXiv preprint arXiv:2502.05791*, 2025.

R. Bartel. Wash-1400 — the reactor safety study — the introduction of risk assessment to the regulation of nuclear reactors. NUREG/KM NUREG/KM-0010, U.S. Nuclear Regulatory Commission, Rockville, MD, Aug. 2016. URL https://www.nrc.gov/docs/ML1622/ML16225A002.pdf. Knowledge Management NUREG; incorporates material from a Nov. 9, 2015 NRC lecture.

Basel Committee on Banking Supervision. Basel iii: A global regulatory framework for more resilient banks and banking systems. BCBS Publication BCBS 189, Bank for International Settlements, Basel, June 2011. URL https://www.bis.org/publ/bcbs189.htm. Revised version (originally issued December 2010).

Basel Committee on Banking Supervision. Stress testing principles. Technical report, Bank for International Settlements (BIS), Basel, Oct. 2018. URL https://www.bis.org/bcbs/publ/d450.pdf. Guidelines, October 2018.

Y. Bengio. Faq on catastrophic ai risks. Blog post on yoshuabengio.org, June 2023. URL https://yoshuabengio.org/2023/06/24/faq-on-catastrophic-ai-risks/.

Y. Bengio and advisory panel. International ai safety report. Technical Report DSIT 2025/001, UK Department for Science, Innovation and Technology (DSIT), London, Jan. 2025. URL https://www.gov.uk/government/publications/international-ai-safety-report-2025.

Y. Bengio and A. Panel. International ai safety report. Technical Report DSIT 2025/001, UK Department for Science, Innovation and Technology (DSIT), London, Jan. 2025a. URL https://assets.publishing.service.gov.uk/media/679a0c48a77d250007d313ee/International_AI_Safety_Report_2025_accessible_f.pdf.

Y. Bengio and E. A. Panel. International ai safety report. Technical Report DSIT 2025/001, UK Department for Science, Innovation and Technology (DSIT), 2025b. URL https://www.gov.uk/government/publications/international-ai-safety-report-2025. Accessed 8 October 2025.

L. Bereska and E. Gavves. Mechanistic interpretability for ai safety–a review. *arXiv preprint arXiv:2404.14082*, 2024.

M. D. Buhl, G. Sett, L. Koessler, J. Schuett, and M. Anderljung. Safety cases for frontier ai. *arXiv preprint arXiv:2410.21572*, 2024.

Cabinet Office. National risk register 2025. Technical report, UK Cabinet Office, London, Jan. 2025. URL https://www.gov.uk/government/publications/national-risk-register-2025. PDF, 187 pages. Published 16 January 2025. Direct PDF: https://assets.publishing.service.gov.uk/media/67b5f85732b2aab18314bbe4/National_Risk_Register_2025.pdf.

S. Campos, H. Papadatos, F. Roger, C. Touzet, O. Quarks, and M. Murray. A frontier ai risk management framework: Bridging the gap between current ai practices and established risk management. *arXiv preprint arXiv:2502.06656*, 2025.

C. Cârlan, F. Gomez, Y. Mathew, K. Krishna, R. King, P. Gebauer, and B. R. Smith. Dynamic safety cases for frontier ai. *arXiv preprint arXiv:2412.17618*, 2024.

Center for Chemical Process Safety (CCPS). *Guidelines for Hazard Evaluation Procedures*. John Wiley & Sons, Inc., Hoboken, NJ, 3 edition, 2008. ISBN 978-0-471-97815-2. URL https://content.e-bookshelf.de/media/reading/L-581567-fb64e9fd37.pdf. A joint publication of CCPS (AIChE) and John Wiley & Sons.

Z. S. Chin. Dimensional characterization and pathway modeling for catastrophic ai risks. *arXiv preprint arXiv:2508.06411*, 2025.

J. Clymer, N. Gabrieli, D. Krueger, and T. Larsen. Safety cases: How to justify the safety of advanced ai systems. *arXiv preprint arXiv:2403.10462*, 2024.

Convergence Analysis. Scenario research. Program page, 2025. URL https://www.convergenceanalysis.org/programs/scenario-research. Program overview with posts on AI scenario planning (e.g., timelines to transformative AI); accessed 2025-10-08.

R. Cooke. *Experts in uncertainty: opinion and subjective probability in science*. Oxford university press, 1991.

F. Cremer, B. Sheehan, M. Fortmann, A. N. Kia, M. Mullins, F. Murphy, and S. Materne. Cyber risk and cybersecurity: a systematic review of data availability. *The Geneva papers on risk and insurance. Issues and practice*, 47(3):698, 2022.

D. Dalrymple, J. Skalse, Y. Bengio, S. Russell, M. Tegmark, S. Seshia, S. Omohundro, C. Szegedy, B. Goldhaber, N. Ammann, et al. Towards guaranteed safe ai: A framework for ensuring robust and reliable ai systems. *arXiv preprint arXiv:2405.06624*, 2024.

V. de Vasconcelos, W. A. Soares, A. C. L. da Costa, and A. L. Raso. Deterministic and probabilistic safety analyses. In *Advances in System Reliability Engineering*, pages 43–75. Elsevier, 2019.

A. Depetro, G. Gamble, and K. Moinuddin. Fire safety risk analysis of conventional submarines. *Applied Sciences*, 11(6):2631, 2021.

K. Dowd. *Measuring market risk*. John Wiley & Sons, 2007.

M. Eling. Cyber risk research in business and actuarial science. *European Actuarial Journal*, 10(2):303–333, 2020.

P. Embrechts, A. J. McNeil, and D. Straumann. Correlation and dependence in risk management: Properties and pitfalls. In M. A. H. Dempster, editor, *Risk Management: Value at Risk and Beyond*, pages 176–223. Cambridge University Press, Cambridge, 2002. doi: 10.1017/CBO9780511615337.008. URL https://www.cambridge.org/core/books/abs/risk-management/correlation-and-dependence-in-risk-management-properties-and-pitfalls/F9949C82974CEA5BEC80A8A96CD6CB80.

W. N. Espeland and M. L. Stevens. A sociology of quantification. *European Journal of Sociology / Archives européennes de sociologie*, 49(3):401–436, Dec. 2008. doi: 10.1017/S0003975609000150.

European Banking Authority. Guidelines for common procedures and methodologies for the supervisory review and evaluation process (srep) and supervisory stress testing. Guidelines (Final Report) EBA/GL/2022/03, European Banking Authority, Paris, Mar. 2022. URL https://www.eba.europa.eu/activities/single-rulebook/regulatory-activities/supervisory-review-and-evaluation-process-srep-4. Final Report dated 18 March 2022; revised SREP Guidelines. PDF available via the landing page.

European Commission. The general-purpose ai code of practice — contents. Shaping Europe's Digital Future (digital-strategy.ec.europa.eu), July 2025. URL https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai. Landing page with links to Transparency, Copyright, and Safety & Security chapters.

European Parliament and Council of the European Union. Regulation (eu) 2024/1689 of the european parliament and of the council of 13 june 2024 laying down harmonised rules on artificial intelligence and amending regulations (ec) no 300/2008, (eu) no 167/2013, (eu) no 168/2013, (eu) 2018/858, (eu) 2018/1139 and (eu) 2019/2144 and directives 2014/90/eu, (eu) 2016/797 and (eu) 2020/1828 (artificial intelligence act). Official Journal of the European Union (OJ L), 12 July 2024, July 2024. URL https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32024R1689. CELEX: 32024R1689; ELI: http://data.europa.eu/eli/reg/2024/1689/oj.

FAIR Institute. What is fair? FAIR Institute website, 2025. URL https://www.fairinstitute.org/what-is-fair. Overview of Factor Analysis of Information Risk (FAIR); accessed 2025-10-08.

FIRST.Org, Inc. Common vulnerability scoring system (cvss) version 4.0. FIRST website, 2023. URL https://www.first.org/cvss/v4-0/. Landing page with links to the Specification, User Guide, calculator, and training; released Nov 1, 2023.

D. Ganguli, D. Hernandez, L. Lovitt, A. Askell, Y. Bai, A. Chen, T. Conerly, N. Dassarma, D. Drain, N. Elhage, et al. Predictability and surprise in large generative models. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, pages 1747–1764, 2022.

A. Goemans, M. D. Buhl, J. Schuett, T. Korbak, J. Wang, B. Hilton, and G. Irving. Safety case template for frontier ai: A cyber inability argument. *arXiv preprint arXiv:2411.08088*, 2024.

Google DeepMind. Frontier safety framework. White paper Version 2.0, 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. Published 4 February 2025; blog post: https://deepmind.google/discover/blog/updating-the-frontier-safety-framework/.

R. Grosse. Three sketches of asl-4 safety case components. Anthropic Alignment Science Blog, Nov. 2024. URL https://alignment.anthropic.com/2024/safety-cases/.

Y. Y. Haimes. The role of modeling in the risk analysis process. In *Risk Modeling, Assessment, and Management*. John Wiley & Sons, Hoboken, NJ, 2 edition, 2004. doi: 10.1002/0471723908.ch2. URL https://onlinelibrary.wiley.com/doi/10.1002/0471723908.ch2.

Y. Y. Haimes. *Risk modeling, assessment, and management*. John Wiley & Sons, 2011.

J. Halstead and L. Righetti. Assessing the risk of AI-enabled computer worms. Research paper, Centre for the Governance of AI (GovAI), Sept. 2025. URL https://www.governance.ai/research-paper/assessing-the-risk-of-ai-enabled-computer-worms.

V. Hemming, M. A. Burgman, A. M. Hanea, M. F. McBride, and B. C. Wintle. A practical guide to structured expert elicitation using the idea protocol. *Methods in Ecology and Evolution*, 9(1):169–180, 2018.

D. Hendrycks. Tail events and black swans. In *Introduction to AI Safety, Ethics and Society*, pages 217–235. Taylor & Francis (Routledge), 2024. URL https://www.aisafetybook.com/textbook/tail-events-and-black-swans. Online chapter; textbook cite-as on site: Taylor & Francis, 2024 (ISBN 9781032798028).

D. W. Hubbard. *The failure of risk management: Why it's broken and how to fix it*. John Wiley & Sons, 2020.

International Atomic Energy Agency. Safety of nuclear power plants: Design. IAEA Safety Standards Series No. SSR-2/1 (Rev. 1) — Specific Safety Requirements STI/PUB/1715, International Atomic Energy Agency (IAEA), Vienna, 2016. URL https://www-pub.iaea.org/MTCD/Publications/PDF/Pub1715web-46541668.pdf.

International Civil Aviation Organization. Safety management manual (smm). Technical Report Doc 9859, International Civil Aviation Organization (ICAO), Montréal, 2018. URL https://ulc.gov.pl/_download/bezpieczenstow_lotow/Przepisy/icao/Doc_9859_-_Safety_Management_Manual_4th_edition_2018.pdf. Fourth edition, 2018.

International Electrotechnical Commission (IEC). Iec 61025:2006 — fault tree analysis (fta). International Standard IEC 61025:2006, International Electrotechnical Commission (IEC), Geneva, 2006. URL https://webstore.iec.ch/en/publication/4311. Describes fault tree analysis and guidance on its application.

International Organization for Standardization. Iso 31000:2018 — risk management — guidelines. International Standard ISO 31000:2018, International Organization for Standardization (ISO), Geneva, Feb. 2018. URL https://www.iso.org/standard/65694.html. Prepared by ISO/TC 262, Risk management.

International Organization for Standardization (ISO) and International Electrotechnical Commission (IEC). Iso/iec guide 51:2014 — safety aspects — guidelines for their inclusion in standards. Technical Report ISO/IEC Guide 51:2014, ISO and IEC, Geneva, Apr. 2014. URL https://www.iso.org/standard/53940.html. International standard; confirmed 2025.

G. Irving. Safety cases at aisi. AI Security Institute (AISI) blog, Aug. 2024. URL https://www.aisi.gov.uk/work/safety-cases-at-aisi. Post outlining AISI's plans to develop safety case sketches for advanced models.

J. Jones. What is cyber risk quantification (crq) and how does it help risk management decisions? FAIR Institute Blog, Mar. 2023. URL https://www.fairinstitute.org/blog/what-is-cyber-risk-quantification-crq.

P. Jorion. *Financial Risk Manager Handbook: FRM Part I/Part II*. John Wiley & Sons, 2010.

S. Kaplan and B. J. Garrick. On the quantitative definition of risk. *Risk analysis*, 1(1):11–27, 1981.

W. Keller and M. Modarres. A historical overview of probabilistic risk assessment development and its use in the nuclear power industry: a tribute to the late professor norman carl rasmussen. *Reliability Engineering & System Safety*, 89(3):271–285, 2005.

T. Kelly. Safety cases. In N. Möller, S. O. Hansson, J.-E. Holmberg, and C. Rollenhagen, editors, *Handbook of Safety Principles*, pages 361–385. John Wiley & Sons, Hoboken, NJ, 2018. doi: 10.1002/9781119443070.ch16. URL https://onlinelibrary.wiley.com/doi/10.1002/9781119443070.ch16.

W.-K. Lee. Risk assessment modeling in aviation safety management. *Journal of Air Transport Management*, 12(5):267–273, 2006.

N. Leveson. The use of safety cases in certification and regulation. U.S. Chemical Safety Board (CSB) working paper, 2011. URL https://www.csb.gov/assets/1/7/leveson_paper.pdf. Will appear in the Nov/Dec 2011 issue of the *Journal of System Safety*.

N. Leveson. Are you sure your software will not kill anyone? *Communications of the ACM*, 63(2):25–28, Feb. 2020. doi: 10.1145/3376127. URL https://cacm.acm.org/opinion/are-you-sure-your-software-will-not-kill-anyone/.

N. G. Leveson. Subsafe: An example of a successful safety program. In *Engineering a Safer World: Systems Thinking Applied to Safety*. The MIT Press, Cambridge, MA, 2012. URL https://direct.mit.edu/books/oa-monograph/2908/chapter/78987/SUBSAFE-An-Example-of-a-Successful-Safety-Program. Open Access chapter on MIT Press Direct.

J. Lindsey and A. Panel. On the biology of a large language model. *Transformer Circuits Thread*, Mar. 2025. URL https://transformer-circuits.pub/2025/attribution-graphs/biology.html.

H. Linstone and M. Turoff. *The Delphi Method: Techniques and Applications*. Addison–Wesley, Reading, MA, 1975. ISBN 0201042940.

K. Lukošiūtė and A. Swanda. Llm cyber evaluations don't capture real-world risk. *arXiv preprint arXiv:2502.00072*, 2025.

Microsoft. Microsoft threat modeling tool threats. Microsoft Learn, Aug. 2022. URL https://learn.microsoft.com/en-us/azure/security/develop/threat-modeling-tool-threats. Threat categories for the Microsoft Threat Modeling Tool (SDL/STRIDE); accessed 2025-10-08.

M. G. Morgan. Use (and abuse) of expert elicitation in support of decision making for public policy. *Proceedings of the National academy of Sciences*, 111(20):7176–7184, 2014.

A. Mukhopadhyay, S. Chatterjee, K. K. Bagchi, P. J. Kirs, and G. K. Shukla. Cyber risk assessment and mitigation (cram) framework using logit and probit models for cyber insurance. *Information Systems Frontiers*, 21(5): 997–1018, 2019.

M. Murray. Ai risk management can learn a lot from other industries. *AI Frontiers*, Apr. 2025. URL https://ai-frontiers.org/articles/ai-risk-management-can-learn-a-lot-from-other-industries. Guest Commentary.

S. Mylius. Systematic hazard analysis for frontier ai using stpa. *arXiv preprint arXiv:2506.01782*, 2025.

National Institute of Standards and Technology. Artificial intelligence risk management framework: Generative artificial intelligence profile. NIST AI Series NIST AI 600-1, National Institute of Standards and Technology (NIST), Gaithersburg, MD, July 2024. URL https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf. Approved by the NIST Editorial Review Board on 2024-07-25.

OECD.AI. Oecd.ai policy navigator. OECD.AI Dashboards, 2025. URL https://oecd.ai/en/dashboards/overview. Accessed: 2025-10-08. Suggested citation on page: "OECD.AI (2025), OECD.AI Policy Navigator, accessed on 08/10/2025, https://oecd.ai/dashboards.".

OpenAI. Our approach to ai safety. OpenAI website, Apr. 2023. URL https://openai.com/index/our-approach-to-ai-safety/.

OpenAI. Building an early warning system for LLM-aided biological threat creation. OpenAI Research blog, Jan. 2024. URL https://openai.com/index/building-an-early-warning-system-for-llm-aided-biological-threat-creation/. Publication.

OpenAI. Preparedness framework. Technical report, OpenAI, Apr. 2025. URL https://cdn.openai.com/pdf/18a02b5d-6b67-4cec-ab64-68cdfbddebcd/preparedness-framework-v2.pdf. Version 2; last updated 15 April 2025.

OWASP Foundation. Threat modeling process — stride. OWASP Community Pages, 2025. URL https://owasp.org/www-community/Threat_Modeling_Process#stride. Section "STRIDE" (Spoofing, Tampering, Repudiation, Information Disclosure, Denial of Service, Elevation of Privilege); accessed 2025-10-08.

M. E. Paté-Cornell. Uncertainties in risk analysis: Six levels of treatment. *Reliability Engineering & System Safety*, 54(2-3):95–111, 1996.

E. Perrier. Statistical scenario modelling and lookalike distributions for multi-variate ai risk. *arXiv preprint arXiv:2502.14491*, 2025.

J. Petrie, O. Aarne, N. Ammann, and D. Dalrymple. Flexible hardware-enabled guarantees for ai compute. *arXiv preprint arXiv:2506.15093*, 2025.

L. Righetti. Dual-use ai capabilities and the risk of bioterrorism: Converting capability evaluations to risk assessments. Research paper, Centre for the Governance of AI (GovAI), Sept. 2025. URL https://www.governance.ai/research-paper/dual-use-ai-capabilities-and-the-risk-of-bioterrorism-converting-capability-evaluations-to-risk-assessments.

M. Rodriguez, R. A. Popa, F. Flynn, L. Liang, A. Dafoe, and A. Wang. A framework for evaluating emerging cyberattack capabilities of ai. *arXiv preprint arXiv:2503.11917*, 2025.

R. Ross, V. Pillitteri, R. Graubart, D. Bodeau, and R. McQuaid. Developing cyber resilient systems: a systems security engineering approach. Technical report, National Institute of Standards and Technology, 2019.

M. Rothschild and J. E. Stiglitz. Increasing risk: I. a definition. In *Uncertainty in economics*, pages 99–121. Elsevier, 1978.

L. Sharkey, B. Chughtai, J. Batson, J. Lindsey, J. Wu, L. Bushnaq, N. Goldowsky-Dill, S. Heimersheim, A. Ortega, J. Bloom, et al. Open problems in mechanistic interpretability. *arXiv preprint arXiv:2501.16496*, 2025.

B. Sheehan, F. Murphy, A. N. Kia, and R. Kiely. A quantitative bow-tie cyber risk classification and assessment framework. *Journal of Risk Research*, 24(12):1619–1638, 2021.

SKYbrary Aviation Safety. Icao annex 19, safety management. SKYbrary article, 2016. URL `https://skybrary.aero/articles/icao-annex-19-safety-management`. Overview of Annex 19 (1st ed. applicable 14 Nov 2013; Amendment 1 adopted 2 Mar 2016; 2nd ed. 2016). Accessed 2025-10-08.

Society for Risk Analysis. Risk analysis fundamental principles, 2025. URL `https://www.sra.org/risk-analysis-introduction/risk-analysis-fundamental-principles/`. Overview page; site copyright 2025.

D. H. Stamatis. *Failure Mode and Effect Analysis: FMEA from Theory to Execution*. ASQ Quality Press, Milwaukee, WI, 2 edition, 2003. ISBN 0873895983. URL `https://www.amazon.com/Failure-Mode-Effect-Analysis-Execution/dp/0873895983`.

L. Stelling, M. Murray, S. Campos, and H. Papadatos. Evaluating ai companies' frontier safety frameworks: Methodology and results. *arXiv preprint arXiv:2512.01166*, 2025.

A. Tversky and D. Kahneman. Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *science*, 185(4157):1124–1131, 1974.

T. UcedaVélez and M. M. Morana. *Risk Centric Threat Modeling: Process for Attack Simulation and Threat Analysis*. Wiley, 2015. ISBN 978-0-470-50096-5.

UK Ministry of Defence. Defence standard 00-056: Safety management requirements for defence systems. part 1: Requirements. Technical Report DEF STAN 00-056 Part 1, Issue 7, UK Ministry of Defence, Feb. 2017. URL `https://s3-eu-west-1.amazonaws.com/s3.spanglefish.com/s/22631/documents/safety-specifications/def-stan-00-056-pt1-iss7-28feb17.pdf`. Issue 7, dated 28 February 2017.

U.S. Federal Aviation Administration. Safety management system (sms). FAA Programs & Initiatives, Aug. 2024. URL `https://www.faa.gov/about/initiatives/sms`. Overview page for FAA SMS; accessed 2025-10-08.

U.S. Nuclear Regulatory Commission. Defense in depth. NRC Glossary, 2025. URL `https://www.nrc.gov/reading-rm/basic-ref/glossary/defense-in-depth`. NRC glossary entry defining defense in depth; accessed 2025-10-08.

VerSprite. PASTA – process for attack simulation & threat analysis, 2022. URL `https://4598121.fs1.hubspotusercontent-na1.net/hubfs/4598121/Ebooks/2022PASTAEbook.pdf`. Online article. Accessed 2025-12-04.

W. E. Vesely, F. F. Goldberg, N. H. Roberts, and D. F. Haasl. Fault tree handbook. Technical Report NUREG-0492, U.S. Nuclear Regulatory Commission, Washington, DC, Jan. 1981. URL `https://www.nrc.gov/docs/ML1007/ML100780465.pdf`. Manuscript completed March 1980; published January 1981. Division of Systems and Reliability Research, Office of Nuclear Regulatory Research.

D. Vose. *Risk Analysis: A Quantitative Guide*. John Wiley & Sons, Chichester, 3 edition, Apr. 2008. ISBN 9780470512845. URL `https://www.wiley.com/en-us/Risk+Analysis%3A+A+Quantitative+Guide%2C+3rd+Edition-p-9780470512845`.

J. Wang, M. Neil, and N. Fenton. A bayesian network approach for cybersecurity risk assessment implementing and extending the fair model. *Computers & Security*, 89:101659, 2020.

A. R. Wasil, J. Clymer, D. Krueger, E. Dardaman, S. Campos, and E. R. Murphy. Affirmative safety: An approach to risk management for high-risk ai. *arXiv preprint arXiv:2406.15371*, 2024.

A. K. Wisakanto, J. Rogero, A. M. Casheekar, and R. Mallah. Adapting probabilistic risk assessment for ai. *arXiv preprint arXiv:2504.18536*, 2025.

L. A. Wojcik. Probabilistic risk assessment and aviation system safety. *Flight Safety Digest*, pages 1–20, July 1989. URL `https://flightsafety.org/fsd/fsd_jul89.pdf`. Flight Safety Foundation, July 1989.

W. Young and N. G. Leveson. An integrated approach to safety and security based on systems theory. *Communications of the ACM*, 57(2):31–35, 2014.

A Glossary

Bayesian Networks (BNs): A type of graphical model that represents and quantifies probabilistic relationships among a set of variables. In a BN, nodes represent events or states, and connecting arcs represent conditional dependencies, making them well-suited for modeling complex causal chains and updating probabilities as new evidence becomes available.

Bottleneck Analysis: A risk prioritization technique that focuses on identifying critical points or stages within a causal chain (such as a multi-step cyber attack) where an intervention would be most effective at disrupting the entire process, or where an AI-enabled capability would provide the greatest advantage.

Defense-in-depth: A core principle in safety engineering that involves implementing multiple, independent layers of protective mechanisms. The goal is to create a highly resilient system where the failure of a single layer does not lead to a catastrophic outcome, as other layers are still in place to prevent or mitigate the consequences.

Deterministic Modeling / Deterministic Safety Analysis (DSA): An approach to safety analysis that assesses a system's resilience against a pre-defined, bounded set of credible scenarios (called "Design Basis Accidents" or DBAs). Rather than calculating probabilities, it uses established engineering principles or stress-tests to determine if the system meets fixed success criteria, resulting in a binary (safe/unsafe) outcome.

Event Tree Analysis (ETA): A bottom-up, forward chaining scenario building technique that graphically maps the potential outcomes following a single initiating event. It explores the branching paths of possible consequences based on the success or failure of various safety functions or subsequent events.

Failure Mode, Effects, and Criticality Analysis (FMECA): A forward chaining scenario building and risk prioritization technique that extends Failure Mode and Effect Analysis (FMEA). It involves identifying potential failure modes of components or processes, analyzing their effects on the system, and then ranking them by a criticality score, which is a function of their severity, probability of occurrence, and detectability.

Fault Tree Analysis (FTA): A top-down, deductive scenario building technique where an undesired "top event" (a specific system failure) is traced backward to its root causes. It uses Boolean logic (AND/OR gates) to represent how combinations of lower-level failures can lead to the top-level outcome.

Harm: The realized adverse outcomes resulting from a hazard. In the context of AI, this can include economic damage, loss of life, societal disruption, or other negative consequences.

Hazard: The source of risk. In the context of AI, a hazard is often a model's capability, property, or tendency that has the potential to cause harm.

Minimal Cut Sets (MCS): Derived from Fault Tree Analysis, an MCS is the smallest combination of component failures that, if they all occur, will cause the top-level system failure to occur. They represent the most direct pathways to failure and are critical targets for prioritization.

Probabilistic Modeling / Probabilistic Safety Analysis (PSA): An approach to safety analysis that aims to identify and analyze as many potential credible accident scenarios as possible. It uses techniques like Fault Tree and Event Tree Analysis to model failure pathways and then assigns probabilities to each step to produce a quantitative risk profile (e.g., the annual probability of a specific failure), rather than a binary outcome.

Risk: The combination of the probability of occurrence of harm and the severity of that harm. It is often conceptualized as a triplet: a scenario describing what can happen, the likelihood of that scenario, and its potential consequences.

Risk Scenario: A logically laid-out sequence of causal steps linking a hazard (a source of risk) to a harm (a realized adverse outcome), taking into account the contexts in which the system may be deployed and the potential for intervening events or failures.

Risk Tolerance: A predefined level of risk that an organization, regulator, or society deems acceptable. In a risk management framework, estimated risks are compared against the risk tolerance to inform decisions about whether a system should be deployed or if further mitigation is required.