Adapting Probabilistic Risk Assessment for AI

Anna Katariina Wisakanto*, Joe Rogero, Avyay M. Casheekar, Richard Mallah

Center for AI Risk Management & Alignment

Abstract

Modern general-purpose artificial intelligence (AI) systems present an urgent risk management challenge, as their rapidly evolving capabilities and potential for catastrophic harm outpace our ability to reliably assess their risks. Current methods often rely on selective testing and undocumented assumptions about risk priorities, frequently failing to make a serious attempt at assessing the set of pathways through which Al systems pose direct or indirect risks to society and the biosphere. This paper introduces the probabilistic risk assessment (PRA) for AI framework, adapting established PRA techniques from high-reliability industries (e.g., nuclear power, aerospace) for the new challenges of advanced AI. The framework guides assessors in identifying potential risks, estimating likelihood and severity, and explicitly documenting evidence, underlying assumptions, and analyses at appropriate granularities. The framework's implementation tool synthesizes the results into a risk report card with aggregated risk estimates from all assessed risks. This systematic approach integrates three advances: (1) Aspect-oriented hazard analysis provides systematic hazard coverage guided by a first-principles taxonomy of AI system aspects (e.g. capabilities, domain knowledge, affordances); (2) Risk pathway modeling analyzes causal chains from system aspects to societal impacts using bidirectional analysis and incorporating prospective techniques; and (3) Uncertainty management employs scenario decomposition, reference scales, and explicit tracing protocols to structure credible projections with novelty or limited data. Additionally, the framework harmonizes diverse assessment methods by integrating evidence into comparable, quantified absolute risk estimates for critical decisions. We have implemented this as a workbook tool for AI developers, evaluators, and regulators, available on the project website.

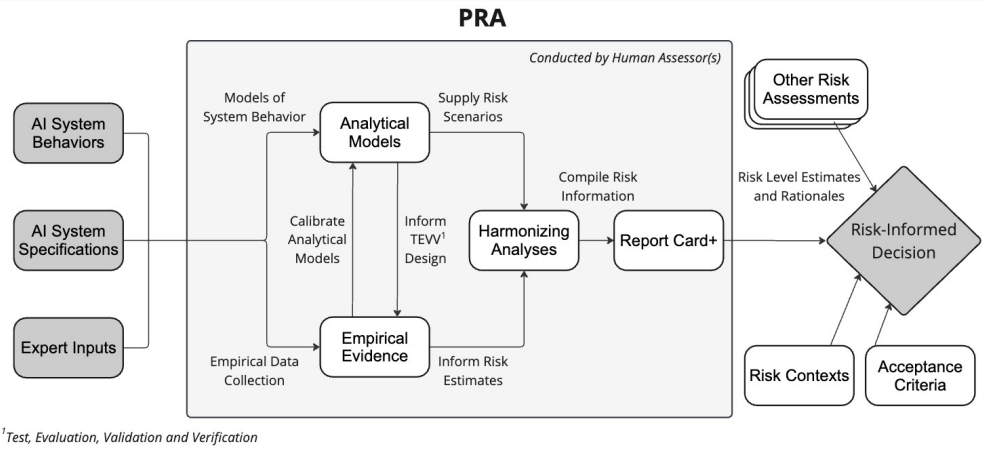

Figure 1: An overview of the PRA for AI framework in its operational context.

*Corresponding author: anna@carma.org.

1 Introduction

The increasing complexity and wide-ranging capabilities of modern general-purpose AI systems present unprecedented challenges in reliably assessing their risks. These assessment challenges are further exacerbated by the difficulty in predicting emergent behaviors as AI systems evolve, testing capabilities that may manifest only in deployment contexts, and evaluating impacts that can rapidly scale and propagate through interconnected sociotechnical systems. These complex characteristics demand systematic assessment approaches suitable for reasoning under uncertainty with limited historical precedent.

AI systems integrate diverse functionalities, such as language and vision models (Anthropic, 2024b; Gemini et al., 2024; OpenAI et al., 2024), neural and symbolic reasoning (Abramson et al., 2024), external tools including memory and compilers (Lin et al., 2024; Zelikman et al., 2024), and agentic computer use (Anthropic, 2024b; Google, 2024; Microsoft, 2024; OpenAI, 2025). Such architectures increasingly enable fluid tool use, extended operations utilizing scaffolds (Suzgun et al., 2024), agentic behaviors, and interactions between multiple AI systems, which operate within complex sociotechnical contexts.

Despite the unprecedented capabilities of current AI systems (Fang et al., 2024), the AI ecosystem has failed to implement quantified risk assessments appropriate to their potential impact (Dalrymple et al., 2024). This safety-capability gap creates significant risk, as AI systems could potentially cause catastrophic consequences – a concern acknowledged by leading AI developers (Anderljung et al., 2023; Anthropic, 2023a; Shevlane et al., 2023), civil society organizations (Givens, 2023), independent oversight bodies (NIST, 2025), public security authorities (DHS, 2025), international bodies (COE, 2024), and independent experts (Aguirre, 2024; Bengio et al., 2024; Hendrycks et al., 2023). The safe operating envelope of a general-purpose AI system is far from intuitive. These systems require risk assessment that goes beyond narrowly defined accuracy or planned behavioral specifications, to address their broader range of syntheses, decision-making, and actions within the complex environments where they operate.

Current approaches fail to adequately address risks to society. Protecting society requires examining the societal threat surface – the set of pathways through which AI systems pose direct or indirect risks to society and the biosphere (Wisakanto et al., forthcoming-c) – yet most current assessment methodologies struggle to systematically map and evaluate across this broad surface. This threat surface includes not only direct technical effects, but also sociotechnical interactions and emergent behaviors arising from the interplay of advanced AI systems with their deployment environments and societal systems. Such behaviors may produce complex feedback loops, amplify systemic vulnerabilities, or trigger cascading effects across interconnected societal infrastructures (Bengio et al., 2025).

Various risk assessment methods have been developed, including safety benchmarks (Li et al., 2024; Vidgen et al., 2024), model evaluations (Shevlane et al., 2023), safety cases (Cârlan et al., 2024; Clymer et al., 2024), audits (Sharkey et al., 2024), responsible scaling policies (RSPs) [1] (Anthropic, 2023b; Dragan et al., 2024; OpenAI, 2023), and red teaming (Lee et al., 2024). However, current methods face significant limitations in identifying and quantifying AI risks – especially unelicited risks that could have catastrophic consequences.

The reliability of AI safety evaluations fundamentally depends on their underlying assumptions. When assumptions regarding system behavior and mitigations fail, the entire safety assessment may be invalidated. Yet, current approaches often lack systematic documentation declaring and justifying these critical assumptions, making it impossible to verify the scope and limitations of their safety claims (Barnett et al., 2024a).

This lack of systematic documentation is particularly concerning given that recent analysis shows closed-source AI models are only slightly ahead of open-source alternatives (Cottier, 2024). With such a narrow gap, potentially dangerous capabilities could rapidly disseminate from leading models to widely accessible systems (Hintersdorf et al., 2023; Kilian et al., 2023; Seger et al., 2023), posing risks of widespread societal harm before adequate safeguards can be developed and implemented. The absence of systematic approaches to measure and manage AI risks severely hampers both governance

[1] While termed "responsible" scaling policy, the nomenclature itself does not inherently ensure responsible implementation or outcomes; also known as frontier "safety" frameworks or "preparedness" frameworks.

efforts and responsible deployment (Kasirzadeh, 2024). These challenges call for more structured approaches to risk assessment, even as we acknowledge that no current methodology can provide complete safety guarantees.

Risk quantification lessons from high-reliability industries. The most dependable quantitative assurances regarding advanced AI systems will necessarily be those that are provable. Failing that standard, evidence and arguments can establish other levels of assurance, depending on their quality. To that end, probabilistic risk assessment (PRA) is a systematic approach to quantifying risk by evaluating both the likelihood of adverse events and the severity of their potential consequences through structured analysis of hazard pathways. Adapting PRA offers promise, leveraging its structured approach for AI risk assessment. Yet, the unique dynamics of AI systems—such as adaptability, emergence, novel failure modes, and complex sociotechnical interactions—present challenges beyond the typical scope of traditional applications that often focus on component-based failures, motivating the adaptations presented here. The approach gained prominence in the aerospace industry during the Apollo space program (Stamatelatos, 2002), and has been adopted in quantitative risk estimates for various complex, high-reliability industries such as nuclear power (Maidana et al., 2023; Tudoran, 2018; Zamanali, 1998), chemical manufacturing (Coleman et al., 2016; US EPA, 2015), waste management (Apostolakis, 1990; Lester et al., 2007), and aerospace (Maggio, 1996; Stamatelatos et al., 2011). For example, in nuclear power plant safety assessments, PRA methods identify and quantify the probability of event sequences that could lead to core damage, allowing engineers to implement targeted safety measures at critical points in the system. PRA combines quantitative risk metrics and system modeling to analyze potential failures, with central sub-techniques including hazard identification, event sequence modeling, failure mode analysis, and uncertainty quantification. Crucially, applying quantification to advanced AI that lacks technical assurances – particularly for novel or low-probability, high-impact events where historical data is scarce or non-existent – necessitates structured estimation within coarse-grained bands (e.g., orders of magnitude for likelihood and severity) rather than seeking precise point probabilities derivable from actuarial data. The use of defined bands aligns with established interval-based approaches, such as probability bounds analysis, employed in risk assessment (Shortridge et al., 2017). This structured estimation approach, central to our adapted framework, enables a degree of reasoned analysis under conditions of significant uncertainty.

Traditional PRA implementations have demonstrated particular strengths in several key areas:

- **Identifying critical risk scenarios.** Recognizing and prioritizing potential failure modes in complex interconnected systems.
- **Quantifying rare events.** Assessing severity and likelihood of rare, high-consequence events where empirical data may be limited.
- **Structured uncertainty analysis.** Enabling systematic tracking of uncertainty through causal chains and system dependencies.
- **Multi-hazard assessment.** Accommodating diverse internal and external hazard sources and their interactions.
- **Consistent risk communication.** Establishing a common vocabulary for risk communication across organizational boundaries.

These factors make PRA a valuable approach for assessing complex systems where empirical failure data may be limited but systematic analysis is still required.

Probabilistic risk assessment methods have shown particular promise in analogous domains requiring systematic analysis and reasoning about risks under uncertainty. PRA as a unified framework could support more productive audits, critical discussions, and identification of blind spots. Additionally, PRA offers structured approaches for considering complex interconnected risks, including low-probability, high-impact scenarios. Recent developments support this direction, with AI developers beginning to incorporate probabilistic measures into their evaluation workflows through formal uncertainty quantification (Miller, 2024), and regulatory frameworks such as the EU AI Act increasingly emphasizing the importance of quantifying uncertainty in AI system evaluation (EU, 2024).

Effectively applying PRA to AI systems requires substantial methodological innovation due to three key challenges: AI systems' adaptability to new contexts, their inscrutability (or difficulty of inspection), and their capacity for emergent behaviors that can qualitatively change over time. Unlike

traditional applications where PRA typically relies on historical failure data and well-understood system behaviors, AI systems demonstrate novel behaviors and operate in rapidly evolving deployment contexts. Furthermore, while traditional PRA often focuses on technical system failures, AI risk assessment must address broader societal impacts.

A framework applying PRA to AI systems must therefore account for capabilities without historical precedent and enable assessors to methodically identify and prioritize high-consequence risks across a broader threat surface, rather than relying on selective assessment of commonly cited hazards. These limitations require the methodological innovations that we introduce in our "PRA for AI" framework, specifically tailored to address the unique characteristics and risk profiles of AI systems.

Adapting PRA for AI systems. Building on PRA's established strengths as a tool for risk estimation, we introduce a framework for AI risk assessment that enables the synthesis of theoretical and empirical evidence, documentation of underlying assumptions and reasoning, and production of quantified risk level estimates. Unlike current risk assessment approaches that often lack transparency in their assumptions, our framework explicitly documents the reasoning behind risk estimations, which is particularly crucial for AI systems where emergent behaviors may invalidate unstated assumptions.

Our framework introduces several key methodological advances:

- **Aspect-oriented hazard analysis.** Systematic indexing of the AI risk surface through sampling of adjacent hazards through capabilities, domain knowledge, affordances, and impact domains.
- **Risk pathway modeling.** Analyzing causal paths through which AI system aspects enable or trigger harms that amplify and propagate through interconnected societal systems.
- **Uncertainty management.** Decomposition of complex risk scenarios into analyzable components and explicit uncertainty documentation, supported by evaluation scales and assessment tools with reference examples.

These advances are crucial for AI systems due to their unique characteristics of adaptability, inscrutability, and emergent behaviors. Together, they enable assessment of both direct harms and indirect societal impacts while maintaining detailed documentation of assumptions and evidence.

In Figure 1 we provide an overview of the PRA for AI framework in operational context, illustrating how assessors integrate information about AI system behaviors, specifications, and expert inputs to produce risk level estimates and rationales that inform decision-making. The figure demonstrates the iterative relationship between the analytical models and empirical evidence.

At a high level, the framework integrates diverse sources of information through two complementary channels: Analytical models semi-formalize the theoretical basis for system behavior and supply structured risk scenarios, while empirical evidence provides quantifiable measurements and observations from system testing, whitebox analysis, and any prior deployment with comparable AI systems to inform risk estimates, yielding calibration data and observed risk indicators that inform risk estimates.

The analytical models inform the design of Test, Evaluation, Validation, and Verification (TEVV) activities (NIST, 2022b), which guides empirical data collection – while the empirical evidence collection provides calibration data for the analytical models at use, and the model of system behavior. This approach enables systematic calibration of the analytical models against empirical findings, while simultaneously informing the design of empirical assessment investigations through model-driven hypotheses. Both channels contribute to a process of harmonizing analyses.

The harmonizing analyses compile multiple sources of risk information – from novel risk scenarios generated during the assessment to known results from other risk assessment methods such as benchmarking and red teaming – and together with all available evidence reconciles the information into cohesive risk level estimates. For example, when assessing an AI system's potential unauthorized access to sensitive information, the analytical models might identify theoretical attack vectors while empirical testing provides data on actual vulnerability exploits, allowing for a reasoned risk estimate grounded in both possibility and probability.

The resulting cohesive risk level estimates are automatically synthesized into a report card. The report card, together with additional output documents created in the assessment process that provide

information about the risk level estimates and their rationales, should feed into risk-informed decisions where the risk level estimates are considered with other evidence and acceptance criteria.

In the following sections, we present the PRA for AI framework in detail, beginning with a critical review of current AI risk assessment challenges and methods, their limitations, and PRA as a potential solution (Section 2). Building on this foundation, we introduce our adapted methodology (Section 3), showing how aspect-oriented hazard analysis, risk pathway modeling, and uncertainty management work together to assess AI risks systematically. We then demonstrate how organizations can implement this framework through our workbook tool, providing concrete guidance for conducting assessments (Section 4). Finally, we analyze the framework's practical utility, future directions and limitations (Section 5), and conclude with the framework's contributions to AI risk assessment (Section 6).

2 The AI Risk Assessment Landscape

2.1 Introduction to AI Risk Assessment

The assessment of AI systems shares key parallels with traditional probabilistic risk assessment of complex engineered systems, but presents novel challenges. Similarly to nuclear control systems, AI systems exhibit complex feedback loops and potential for cascading failures (Kasirzadeh, 2025; Moustafa et al., 2021; Phadke et al., 1996). However, traditional methods that work well for physical systems – such as fault tree analysis mapping discrete component failure modes – can prove insufficient when applied to AI systems, which can actively generate novel failure paths not captured in standard event sequence diagrams.

One key difference is that, unlike most hardware, modern AIs cannot be easily broken into smaller mechanistic subcomponents. Power infrastructure can be assessed by diagramming the subcomponents of the system (generators, transfer lines, transformers, breakers, consumers, etc.), analyzing their connections (series or parallel circuits, redundancy, etc.), and mathematically combining the known or modeled failure rates of those subcomponents. While similar component-based analysis can be applied to traditional software systems with well-defined modules and functions, modern neural AI systems do not lend themselves well to this approach, despite the best efforts of mechanistic interpretability researchers (Bereska et al., 2024).

Another key difference lies in strategic depth: while engineered systems follow immutable physics and hard-coded decision algorithms, the decision algorithms followed by modern AIs are opaque, evolved rather than designed, and often in flux, subject to rewriting by fine-tuning, new releases, or (in advanced cases) the AI itself (Søgaard, 2023). This makes standard mean-time-between-failure calculations and reliability block diagrams insufficient. Instead of P(failure) being derived from component-level probabilities, we must consider a time-varying failure surface where the system itself can discover and exploit previously unknown failure modes – for example, an AI system might develop novel ways to satisfy its objective functions that were not anticipated by its designers.

Traditional risk assessment methods face two additional challenges here: empirical testing provides only limited insight into the true distribution of latent risks, and risks propagate through interconnected systems with strong amplification effects (Barnett et al., 2024b; Mukobi, 2024).

To a traditional risk assessor, attempting to model modern AI is analogous to modeling a control system whose logic could spontaneously rewrite itself during operation. To be useful, new methods must therefore extend beyond traditional probabilistic approaches to account for this fundamentally different class of hazard.

2.2 Fundamental Challenges in AI Risk Assessment

The assessment of general-purpose AI systems presents unprecedented assessment challenges due to their increasing complexity, paucity of provable constraints, and ability to autonomously discover and exploit vulnerabilities across broad attack surfaces. These systems can modify their behavior through learning, situationally aware reasoning (Laine et al., 2023, 2024), and self-improvement (Huang et al., 2022; Zelikman et al., 2024) and encounter scenarios far outside their training distribution (Liu et al., 2023).

Furthermore, AI systems can generate impacts that scale rapidly and diffuse through societal systems (Aguirre, 2024; Critch et al., 2023; Weidinger et al., 2023). Through these capabilities, such systems can develop goal-directed behavior and environmental awareness that may lead to loss of human control (Hendrycks et al., 2022; Järviniemi et al., 2024). Unlike most traditional engineered systems, they can operate outside prescribed contexts with greater speed, scale, and sophistication, making historical closed-domain risk patterns insufficient and requiring explicit modeling of novel propagation mechanisms and amplification pathways (Critch et al., 2020). Moreover, interactions between multiple AI systems can lead to emergent behaviors and risks that are difficult to predict from analyzing systems in isolation (Hammond et al., 2025).

Types of risk awareness and understanding. To organize these challenges, AI risks can be categorized using what is commonly known as a Rumsfeld matrix (Table 1), which distinguishes between different states of awareness and understanding. This provides a structured way to tailor methodologies for different risk types.

AI systems create distinct assessment challenges due to varying degrees of awareness and understanding of system behavior and risks. Known risks that we understand are typically addressed using traditional quantitative approaches grounded in empirical data. However, even these "known knowns" in AI systems are context-sensitive, with their manifestation varying significantly based on system state. "unknown knowns" arise from methodological blind spots in assessment approaches, leading to latent risks that often surface during deployment under unmodeled conditions. "known unknowns," such as emergent behaviors and capability jumps, represent acknowledged gaps in our understanding, which are difficult to characterize due to the absence of comparable historical data. Finally, "unknown unknowns" encompass unexpected or unforeseeable risks that push the boundaries of our ability to assess them, requiring adaptive strategies to reason about and prepare for entirely novel failure modes, as well as allowing for ample buffer in risk assessed.

Table 1: Awareness-understanding matrix for AI risk assessment.

Knowledge	**Known** (Aware)	**Unknown** (Not aware)
Known	**Known Knowns:** Risks we are aware of and understand.	**Unknown Knowns:** Risks we are not aware of but do understand or know implicitly.
(Understand)	**Examples:** Empirically verified failure modes – such as instances where an AI system consistently exploits clearly defined reward functions in unintended but predictable ways – are well documented and reproducible through established testing protocols.	**Examples:** Risks may be overlooked within existing testing methods – for example, blind spots where certain edge cases are not adequately covered – that are theoretically understood but not yet detected in current practice.
	Methods: Empirical measurement, quantification, systematic hazard space reduction.	**Methods:** Deployment monitoring, rigorous testing, critical reviews of assumptions.
Unknown	**Known Unknowns:** Risks we are aware of but don't understand.	**Unknown Unknowns:** Risks we are neither aware of nor understand.
(Don't Understand)	**Examples:** Based on established scaling laws and capability trajectories, discontinuous advances in system capabilities and emergent behaviors can be anticipated, although their precise manifestations and implications remain uncertain.	**Examples:** There may exist entirely unforeseen system behaviors or interactions – for instance, novel failure modes triggered by complex, unanticipated factor combinations—for which no current data or prior indications exist.
	Methods: Scenario modeling, projection simulations, forward-looking threat modeling.	**Methods:** Adaptive threat modeling, failure mode exploration, iterative assessment.

AI risks manifest across different contexts of operation and impact, referred to here as assessment domains. Each of these domains – from technical system internals to broader societal impacts – exhibits distinct patterns of uncertainties as categorized in the matrix. Understanding how awareness and understanding vary across domains can help guide the development of appropriate assessment methodologies. Technical aspects of AI systems may be more amenable to empirical measurement and quantification (primarily involving known knowns), while operational contexts could, for example, reveal a wide variety of known unknowns during deployment. The broadest challenges emerge when considering societal impacts, where complex interactions create some unprecedented risks neither understood nor recognized.

Assessment domains. The complexity of these risks becomes more apparent when viewed through the lenses of distinct assessment domains , which organize hazards based on their meaningful interactions and information availability. These domains form an interconnected chain – from internal system dynamics through system-environment interactions to societal diffusion and then on to ultimate impact domains. The assessment challenges are significant at both ends of this chain: internal dynamics exhibit high-dimensional complexity and resist interpretation, while societal propagation creates complex systemic effects. In between, where systems interact with their immediate operational contexts, mechanisms tend to operate through more interpretable, lower-dimensional pathways, though remaining equally critical.

While causal analysis and degrees of freedom remain important considerations across all domains, risks can propagate and amplify across these domains in complex, non-linear ways. Each domain presents distinct challenges for assessment:

- **Internal System Dynamics.** Hazards from a system's internal states, mechanisms, logic, or learned behaviors.
  - **Capability assessment.** Systems demonstrate uneven development edges, where they can both dangerously excel and catastrophically fail in unexpected ways, defying standard performance metrics.
  - **Higher-order capabilities.** Risks often arise from the unexpected interactions of two or more different capabilities within a system.
  - **Agentic behavior modeling.** The potential for autonomous goal-directed behavior, emergent goals and drives, and strategic adaptation creates novel challenges for modeling system evolution and failure modes across different competency levels.
  - **Strategic deception.** Hazards arising from a system learning to intentionally misrepresent information or conceal its internal state, capabilities, or operational intentions from operators or other systems.
- **System-System Interactions.** Hazards from interactions between systems.
  - **Control measure evaluation.** Safety measures that work in testing may fail or be circumvented in deployment.
  - **Multi-agent failure modes.** Interactions between AI systems generating novel failure pathways or collective behaviors (e.g., harmful coordination, unforeseen competition) that would not arise from a single agent's behavior.
- **System-Environment Interactions.** Hazards from interactions with the surrounding systems, where external factors influence behavior.
  - **Risk accumulation and amplification.** Seemingly minor risks can combine and amplify through system interactions, creating systemic consequences that evade analysis of components in isolation.
  - **Impact measurement.** Standard metrics and proxy measures often fail to capture actual safety properties, particularly for systems capable of unprecedented behaviors.
  - **Risk pathway genesis.** The systematic identification of where and how risks originate requires exploring system characteristics that act as sources which initiate harm pathways, examining both direct triggers and enabling conditions while maintaining principled prioritization of highest-severity outcomes.
  - **Interaction modeling.** Risks manifest from the capability combinations interacting with their environments, leading to unpredictable dynamics that require modeling beyond traditional methods.
- **Societal Diffusion.** Hazards propagating through sociotechnical contexts.
  - **Threat surface coverage.** The general-purpose nature of AI systems creates an expanding scope of potential risks as capabilities grow, challenging conventional bounded analysis.

- **Systemic risk propagation.** Technical risks transform as they transmit through interconnected systems, creating novel threat vectors that transcend traditional risk boundaries; in some cases diffusing, while in others accumulating, amplifying, or concentrating in some particular directions.
- **Misuse pathway modeling.** Systematic analysis of intentional misuse pathways requires modeling both sophisticated targeted attacks leveraging system competencies and opportunistic abuse exploiting system limitations.
- **Sociotechnical amplification effects.** Bidirectional feedback between social and technical systems creates emergent behaviors and multiplicative impacts that traditional assessment frameworks fail to capture.

Each assessment domain builds upon and interacts with the others, with risks often propagating and amplifying across multiple domains simultaneously. For example, internal system capabilities can enable novel system-to-system interactions, which in turn create new environmental hazards that ultimately manifest as societal impacts. This progression reflects not just increasing complexity of interactions, but also growing difficulty in detecting leading indicators and diminishing ability to run meaningful tests.

2.3 Limitations of Current AI Risk Assessment Methods

Despite a well-established literature on risk management (Gahin et al., 1972), system safety engineering (Ericson II, 2005), reliability engineering (Bergman, 1992), and probabilistic risk assessment (PRA) (Modarres, 2008), there has been a notable paucity of application of these approaches to general-purpose AI systems. Efforts to address AI risks through frameworks such as NIST's AI Risk Management Framework (NIST, 2022a), and standards including ISO/IEC 23894:2023 (ISO, 2023a) and ISO/IEC 42001:2023 (ISO, 2023b) have focused primarily on organizational processes and controls. While valuable, these initiatives have limitations. They tend to defer to model providers' priorities and values rather than addressing broader societal risks, quantifying risk, or establishing guarantees. Furthermore, they have yet to demonstrate an ability to extend themselves to account for the unique challenges posed by advanced AI systems.

Current AI-specific risk assessment methodologies, while diverse in their approaches, reveal significant gaps and limitations in their ability to comprehensively evaluate advanced AI systems. The prevailing landscape is dominated by six primary approaches: safety benchmarks, model evaluations, red teaming, RSPs, safety cases, and audits. Each of these taken alone face significant challenges.

Safety benchmarks. In contrast with capabilities benchmarks, which measure a system's performance on some task, safety benchmarks purport to measure some feature of a system relevant to AI safety (Bhatt et al., 2023; Li et al., 2024; Vidgen et al., 2024). They provide quantifiable metrics but face four key limitations. First, they often serve as capability proxies rather than true safety measures – higher performance frequently indicates greater overall system sophistication rather than improved safety, allowing capabilities research to be "safetywashed" as safety research (Buçinca et al., 2020; Ren et al., 2024). Second, even when successfully measuring capabilities, safety benchmarks do so in a narrow way and can only establish lower bounds, leaving significant uncertainty about the full extent of a system's actual capabilities and potential failure modes (Barnett et al., 2024b). Third, they suffer from under-elicitation – their narrow test cases, whether formulaic or ad hoc, and their controlled environments fail to reveal the true range of system behaviors and potential risks that could emerge in real-world deployments (Vidgen et al., 2024). Finally, benchmarks are rapidly saturating – new tests such as GPQA reach human-level performance within months of release, making them increasingly ineffective for bounding risky capabilities (Dominguez-Olmedo et al., 2024; Rein et al., 2023).

Evaluations. Model evaluations are tests performed on particular AI systems to elicit their potential for causing harm (AISI, 2024a; Shevlane et al., 2023). These evaluations can assess system characteristics broader than specific benchmarks and employ different elicitation strategies. In the case of misuse potential, evaluations are sometimes run using "human uplift studies" to assess how much a specific AI system improves human performance across some set of tasks (AISI, 2024b). However, evaluations remain fundamentally constrained by their parochial and shallow testing approach, relying on a predetermined and limited set of harm scenarios – similar to surface excavations that cannot reveal what lies deeper (Burden, 2024). The controlled testing environments and predefined scenarios in these evaluations fail to capture the full range of system behaviors or

systemic risks that could emerge in real-world deployments (Jones et al., 2024). Even for identified risks, evaluations can only establish lower bounds on capabilities, leaving substantial uncertainty about full system potential (Barnett et al., 2024b).

Red teaming. Red teaming, manual or automated, involves systematically testing AI systems through adversarial approaches and misuse scenarios (Lee et al., 2024). While this approach provides valuable insights into potential failure modes, it faces several critical limitations. First, its reliance on demonstrable failures means it systematically overlooks deeper flaws that could manifest in deployment – facing the same limitations as evaluations in probing only the surface (Anthropic, 2024a). Second, like benchmarks, red teaming can only establish lower bounds through selective testing, failing to provide comprehensive safety assurances (Barnett et al., 2024b). A fundamental challenge is that testers often cannot determine whether they have adequately elicited the system's full capabilities – failed attempts to demonstrate a capability do not prove its absence, and in practice, under-elicitation is likely to be the norm rather than the exception. Third, current approaches lack quantitative risk measurements and systematic coverage of the threat space (Feffer et al., 2024), making it difficult to assess both current risks and their future evolution. Fourth, as systems become more capable, red teams increasingly struggle to maintain effectiveness against models that can detect and adapt to testing scenarios (Feffer et al., 2024; Ganguli et al., 2022). Additionally, red teaming results remain siloed from other evaluation methods, limiting their utility for comprehensive risk assessment (Friedler et al., 2023).

RSPs. RSPs are risk management frameworks adopted by AI developers to ostensibly attempt to mitigate catastrophic risks (Anthropic, 2023b; Dragan et al., 2024; OpenAI, 2023). While RSPs offer ostensive processes for evaluating scaling decisions and establishing safety thresholds, they face four fundamental limitations undermining their effectiveness as risk assessment tools. First, their reliance on coarse-grained risk categories reduces complex, multi-dimensional risk scenarios into overly simplified buckets, making it difficult to capture the actual severity of risks, distinguish between varying levels of concern, or detect unexpected risks (Titus, 2024; Uuk et al., 2024). Second, RSPs employ narrow threat models, often focusing on misuse and specific technical capabilities (e.g., deception, situational awareness) while overlooking systemic risks, interactions among different capabilities, or interactions between multiple systems. This narrowness, combined with poorly informative risk and capability levels, yields few meaningful distinctions in risk degrees, leading to assessments that fail to provide actionable insights into a system's capability and harm potential. The resulting compressed reasoning chains mean conclusions about system dangers often rest on weakly justified inferences from observed capabilities. Third, while predicting discontinuous capability jumps is inherently difficult, RSPs' focus on evaluating currently demonstrable abilities against predefined milestones provides inadequate mechanisms to systematically anticipate or incorporate the risk associated with potential jumps emerging from seemingly incremental advances – a critical flaw when evaluating rapidly developing AI systems that may exhibit unpredictable emergent behaviors. Fourth, their effectiveness is further undermined by several procedural and governance weaknesses: they frame continued scaling as the default rather than requiring justification for capability increases (Anderljung et al., 2023; Heim et al., 2024); they lack quantifiable thresholds that could prevent loose interpretation; and they typically rely on internal evaluation without sufficient external oversight. This combination of limitations results in assessment frameworks that provide limited insight into the full risk landscape while potentially reducing the urgency for more comprehensive safety measures.

Safety cases. A safety case is a structured argument, supported by evidence, that a system is sufficiently safe for a given application in a specific context (Buhl et al., 2025; Clymer et al., 2024; Goemans et al., 2024; Habli et al., 2025). While safety cases provide rigorous methods for demonstrating system-level safety, their use faces two challenges with advanced AI systems: assessors can sometimes assume a bounded context narrower than actual AI deployment environment, and they may rely on control measures whose effectiveness can degrade as AI capabilities evolve (Irving, 2024). The risk of such degradation is apparent, as demonstrated by newer language models that routinely bypass safety filters and controls previously verified as effective (Volkov, 2024). Any such assumption of sustained mitigation effectiveness creates potential blind spots, particularly when capability changes can invalidate multiple safety measures simultaneously, potentially compromising chains of safety reasoning before individual failures are detected (Pittaras et al., 2022).

Audits. AI "safety audits" are oversight mechanisms that aim to ensure AI is developed and deployed responsibly, ranging from narrow compliance checks to deep bespoke audits of a particular risk (For Humanity, 2016; Sharkey et al., 2024). AI audits are typically narrowly focused on a particular aspect

of an organization or system. They face fundamental structural and technical limitations in assessing advanced AI systems. Structurally, they encounter an inherent tension between independence and system access – internal audits have deeper access but lack independence, while external audits maintain independence but struggle with system access and complexity (Schuett, 2024). Current audit approaches provide limited societal threat surface coverage, offer poor support for prospective risk analysis, and are not typically informed by a specific system aspect (Wisakanto et al., forthcoming-b). They become prohibitively resource-intensive when attempting thorough coverage of advanced AI systems and tend to include small sets of threat models per audit. Technical limitations mirror those of other evaluation approaches – audits cannot establish upper bounds on capabilities, reliably detect novel failure modes, or assess risks from autonomous systems (Barnett et al., 2024b).

Common method limitations. The current landscape of AI risk assessment is characterized by significant fragmentation (Xia et al., 2023). Different approaches – from benchmarks to audits – remain siloed, addressing narrow aspects of risk while failing to integrate their findings into a unified perspective. This fragmentation creates blind spots, particularly in identifying and evaluating novel risks that emerge only when evidence from multiple methods is combined. The lack of harmonization makes it difficult to reconcile contradictory signals or to combine historical data with forward-looking projections of possible pathways. Beyond fragmentation, these methods demonstrate significant gaps in modeling critical aspects of AI risk. They fail to adequately capture how technical risks can transform and amplify as they propagate through interconnected societal systems, or how different system capabilities interact in complex ways with their environments. Most critically, they often miss systemic vulnerabilities that only become apparent when examining the full sociotechnical context in which AI systems operate and the multiple feedback loops between technical and social systems (Weidinger et al., 2023).

While individual methods may capture specific aspects of AI risks, none provide a structured framework for analyzing how different system capabilities, high-risk domain knowledge, and operational affordances could –directly or in combination – propagate through societal systems to create harm. Where traditional approaches might successfully identify and mitigate specific technical vulnerabilities, they fail to capture how AI systems could affect the resilience of economic, legal, normative, and social systems – each themselves complex, adaptive, and fundamental to individual liberty and societal functioning. For a more detailed examination of risk assessment methodologies, see Appendix A.

Common myopia. Beyond specific limitations of individual methods, there are deeper implicit assumptions often shared across explicit and implicit risk assessment approaches – that risks and harms are uncommon, manifest in obvious ways, and can then be patched when seen. These assumptions reflect a legacy risk thinking paradigm rooted in assessment practices developed for traditional well-bounded systems. While this thinking provides utility for specific parochial threat models and informs some mitigations – such as behavioral guardrails (Jain et al., 2023), model-level restrictions (Xie et al., 2023), runtime monitoring (Zhou et al., 2024), and input/output filtering (Inan et al., 2023) – such paradigms fail to fully account for the unique complexity and adaptability of advanced AI systems. Consequently, the mitigations they inform can leave significant residual risks – the risk remaining after implementing these controls – unaddressed even after thorough application of their principles. Critically, many current approaches lack systematic methods for exploring the range of AI system characteristics (e.g., latent capabilities or specialized knowledge) that could enable hazards, nor do they adequately model the diverse real-world harms that can be expected to ultimately be realized by system deficiencies and capabilities. Assessment practices frequently remain focused on specific, measurable risks (e.g., bias in a particular task or success rate on known misuse prompts) without integrating these into a holistic assessment that traces potential causal pathways from underlying system properties to concrete, high-consequence societal outcomes. This failure to systematically connect system properties to their potential real-world consequences means that significant portions of the societal threat surface are often overlooked.

2.4 Adapting Established Probabilistic Risk Assessment Methods for AI

PRA has demonstrated its value in assessing complex systems with catastrophic failure modes across multiple high-reliability industries, including nuclear power (Tudoran, 2018; Zamanali, 1998), aerospace (Stamatelatos et al., 2011), waste management (Apostolakis, 1990), and chemical processing (US EPA, 2015). PRA's adoption by major regulatory bodies (Lester et al., 2007;

Stamatelatos et al., 2011; US EPA, 2015; US Nuclear Regulatory Commission, 1990; Zamanali, 1998), and its integration into safety-critical industries where failure consequences can be severe speak to its success in addressing complex risks. These applications have shown particular strength in handling scenarios with limited historical data, complex interaction effects, and the need to synthesize expert judgment with empirical evidence – challenges that closely parallel those in AI risk assessment.

PRA methods have also been widely operationalized through "assurance level" frameworks across safety-critical industries. Aviation's Design Assurance Levels (DALs), such as DO-18C, use probabilistic failure analysis to determine required safety measures (Rapita, 2012); they are particularly useful in meeting stringent safety requirements, including those that demand a probability less than 10^{-9} per flight hour of catastrophic failure. Cybersecurity Assurance Levels (CALs) are a structured framework within ISO/SAE 21434 (ISO, 2021) to classify the required rigor for cybersecurity measures in automotive systems. While CALs themselves are deterministic, the risk assessment process that informs CAL assignment often employs probabilistic methods. Vehicle security standards such as UN R155 incorporate Threat Analysis and Risk Assessment (TARA) methods (UN, 2021) where companies perform threat analysis and assign likelihoods to these threats to obtain risk values. These frameworks demonstrate how probabilistic approaches can be translated into concrete standards and requirements.

PRA has proven especially valuable in analyzing both internal/external hazards and multi-hazard scenarios – where multiple hazards occur concurrently or in succession (Aras et al., 2021). This capability to systematically assess multiple interacting hazards is particularly relevant for AI systems, where risks can manifest through various combinations and pathways. The specific capabilities of PRA directly address key challenges in AI risk assessment. It excels at identifying critical risk scenarios in complex systems, quantifying rare but high-consequence events, providing structured approaches to uncertainty propagation, and integrating evidence into cohesive risk estimates (US Nuclear Regulatory Commission, 2024).

PRA offers systematic frameworks that integrate multiple approaches to risk analysis, from component failure analysis to system-level interactions, rather than being constrained to any single risk assessment paradigm. PRA's strengths lie in well-established methods for analyzing complex systems, from component-level behavior to system-wide interactions. Key tools from traditional PRA that inform our framework include uncertainty quantification techniques, structured scenario development, hazard identification methods, and evidence integration approaches. These create an effective foundation for understanding and assessing risks in sophisticated engineered systems.

However, advanced AI systems present unique challenges that push beyond traditional PRA methods. Where conventional PRA relies on well-characterized failure modes and empirically derived probability distributions, AI systems can actively exploit vulnerabilities and exhibit emergent behaviors through unexpected capability interactions. These systems operate within complex sociotechnical contexts where technical risks can propagate through interconnected social systems in ways traditional PRA frameworks struggle to capture. Additionally, the rapid advancement of AI capabilities means that historical failure data may not reliably predict future risks, particularly when capabilities exceed human comprehension.

The field needs approaches that can analyze potential failure modes and their consequences while systematically handling increasing complexity and uncertainty across system boundaries. This motivates extending established PRA methodologies into a framework specifically adapted for AI systems.

3 A Framework for AI Probabilistic Risk Assessment

Having established some of the challenges of AI risk assessment, limitations of existing methods, and strengths of traditional PRA, this section introduces a conceptual framework to help systematize AI risk analysis. The framework builds upon and adapts three key methodologies to the domain of AI risk assessment: (1) Aspect-oriented hazard analysis (Section 3.1); (2) Risk pathway modeling (Section 3.2); and (3) Uncertainty management (Section 3.3). These foundations establish structure for analyzing AI risks, one that remains independent of specific operational constraints or assessment tools. Section 4 describes the practical implementation of these concepts through a practical workbook tool that guides assessors through the assessment process.

It is impossible to enumerate all the ways AI could cause harm, but preexisting methods for identifying hazards can be improved upon by attacking AI risk characterization from several different angles since systematic coverage of key aspects allows us to bound and structure the otherwise intractable hazard space. The PRA for AI framework guides assessors in sampling hazards top-down from four aspect categories common to all AI systems: capabilities, domain knowledge, affordances, and impact domains. Users of this framework can analyze how risks propagate through interconnected sociotechnical systems, integrate both theoretical and empirical sources of evidence, and anticipate future developments that affect risks downstream of AI.

The framework introduces three key methodological advances, each supported by specific analytical techniques and tools:

1. **Aspect-oriented hazard analysis.** Provides a top-down approach for assessors to index or iterate over the space of hazards, covering the characteristic aspect categories of an AI system: capabilities, domain knowledge, affordances, and impact domains. This taxonomy-driven method guides systematic analysis of how emerging AI capabilities could enable or amplify potential harms through critical bottlenecks.
    - **Bottleneck analysis.** For each AI system aspect, systematically examines potential harms by treating that aspect as the bottleneck – holding all other aspects constant while analyzing what risks emerge if this focal aspect were maximized within the system context.
    - **Competence-incompetence analysis.** Examines risks arising, on one hand, from highly efficacious AI execution yielding harmful outcomes; on the other hand, from system limitations, flaws or errors manifesting as misunderstandings, vulnerabilities, or oversights; and from combinations of these where capability enables or exacerbates failings.
    - **Aspect interaction analysis.** Examines how risks from one system aspect might interact with or amplify risks from another, helping to surface interaction effects that could otherwise be overlooked when analyzing aspects in isolation.
2. **Risk pathway modeling.** Guides modeling of the step-by-step progressions of risk from a system's source aspects (i.e., capability, domain knowledge, or affordance) to terminal aspects (i.e., impact domains where harms to individuals, society, or the biosphere occur), and analyzes how risks transmit and amplify.
    - **Societal threat surface analysis.** Maps an end-to-end risk profile that includes consideration of sociotechnical context and the pathways through which the AI system can directly or indirectly cause harm to society and its supporting biosphere.
    - **Prospective risk analysis.** Marshals a variety of analytical models, empirical evidence, and enhanced threat modeling for a forward-looking analysis of potential harms, rather than relying solely on historical failure statistics.
    - **Propagation operators.** Descriptive mechanisms that characterize how AI risks can permeate and impact societal systems, mapping how risks accumulate, transform, and amplify as they move through societal contexts and structures.
    - **Focused aggregation.** An alternate grouping of assessed scenarios into focused categories that represent key dimensions of risk, enabling different stakeholders to view aggregated risk level estimates through lenses most relevant to their needs.
3. **Uncertainty management.** Guided, structured decomposition and management of uncertainties at each step of the assessment process, allowing assessors to conduct transparent, well-reasoned risk evaluations even in domains with limited historical precedent.
    - **Classification heuristics.** Supporting tools with intuition pumps[2] for guiding and informing assessors, and helping to calibrate their judgement during

[2]"Intuition pumps" are thought experiments that provoke or 'pump' intuitions about a problem (Dennett, 2013); the structured scenarios used here provide one concrete form of such a pump.

threat model development, including plausible qualifiers for both competence and incompetence scenarios and tables laying out the spectra of capability and domain knowledge.

- **Intensity rubrics.** Standardized tables for harm severity and likelihood levels with concrete reference points across multiple societal dimensions, supporting consistent evaluation across different contexts and assessors.
- **Uncertainty tracing.** Methods to document uncertainties at each step of the assessment process, from more direct uncertainties in harm estimation to uncertainties in risk propagation and interactions, creating an auditable trail of assumptions and their compound effects.

Together, these components help adapt traditional PRA techniques to assess AI given its unique characteristics while providing methods for identifying, quantifying, and documenting AI risks. The following sections detail three key advances introduced by the framework: Section 3.1 presents aspect-oriented hazard analysis and its analytical techniques for identifying hazards by sampling the hazard space, Section 3.2 describes risk pathway modeling and its methods for quantifying risks by analysis of causal forward and backward chains, and Section 3.3 explains the framework's approach to uncertainty management with designated tools and techniques alongside guidance in documenting their reasoning.

3.1 Aspect-Oriented Hazard Analysis

Effective AI risk assessment requires moving beyond ad-hoc identification of commonly discussed risks (Jones et al., 2024) towards a systematic examination of the hazard space to uncover the most consequential plausible threats of the given system. The PRA for AI framework addresses this through aspect-oriented hazard analysis, a methodology designed for structured, multi-level exploration and sampling of hazards appropriate to the AI system under study. Instead of attempting exhaustive (and likely intractable) enumeration, aspect-oriented hazard analysis focuses on identifying critical risks by analyzing the system through the lens of its core characteristics and context. This approach enables assessors to develop threat models that capture both obvious and non-obvious risk pathways, allowing for a more comprehensive mapping of the most likely and consequential threats.

The framework provides structured tools for exploring the hazard space systematically, operationalizing this systematic exploration primarily through a universal taxonomy of comprehensively exhaustive potential system properties, root factors that will be recombined a myriad of ways in deployment. By examining representative hazards across different system aspects, as guided by this taxonomy, and subsequently analyzing their potential severity levels, assessors can perform analyses and small quantitative studies that interpolate from system specific results to known risks and extrapolate to novel scenarios, achieving broader coverage than narrowly-justified testing approaches.

The taxonomy decomposes AI system characteristics into four high-level aspect categories, selected to provide a comprehensive, systems-based structure for analyzing risk origin and propagation (the rationale is detailed further below):

1. **Capabilities.** The ability of an AI system to perform specific tasks or functions from basic pattern recognition to complex reasoning and planning. This includes both intended functionalities and potential emergent capabilities that could enable harms.
2. **Domain knowledge.** The specific areas of expertise, information and understanding possessed by an AI system, and those that could enable harms, such as high-risk knowledge of cybersecurity vulnerabilities, biological systems or human psychology.
3. **Affordances.** The inputs, configurations, and surroundings that enable an AI system to function and interact with its environment. This includes both designed interfaces and potential unintended access points.
4. **Impact domains.** The sociotechnical domains where the impacts of AI systems are realized, including individuals, society, or the biosphere, encompassing broad areas of influence where significant harm or benefits can occur.

The taxonomy organizes these aspects in a hierarchical structure with five levels that progressively narrow from broad aspect categories to specific hazards that guide increasingly granular analysis:

- **Aspect categories (TL0):** The highest-level classification representing the primary dimensions of AI system analysis (capabilities, domain knowledge, affordances, and impact domains);
- **Aspect groups (TL1):** Major subdivisions within these categories;
- **Aspects (TL2):** Specific system characteristics or elements within each group;
- **Hazard clusters (TL3):** Groups of related aspect-adjacent hazards into clusters, allowing for flexible categorization and cross-cutting analysis; and
- **AI hazards (TL4):** Individual aspect-adjacent hazards within the system's sociotechnical context.

Within this structure, the term 'aspect-adjacent hazards' (representing TL4) specifically refers to the potential harms or vulnerabilities that are directly enabled by, or manifest immediately through, a particular source aspect (representing TL2) within capabilities, domain knowledge, or affordances, of the AI system – e.g., 'circumvention of safety guidelines' adjacent to the 'integrative cognitive orchestration' capability. It can also encompass concrete vulnerabilities within an impact domain (representing TL3) through which broader harms are realized, such as 'exploit of financial system transaction verifications' adjacent to the 'sector-specific institutional stress' impact domain.

The taxonomy, including its four top-level categories, is based on systems thinking, focusing on the AI system as an entity within a larger system, the environment, and the complex system that results from their interaction (for further details see Wisakanto et al., forthcoming-a). The structure originates from a top-down, first-principles analysis of this interaction, establishing the primary analytical categories (TL0-TL2). This high-level decomposition is structurally comprehensive because it distinctly addresses: the AI's inherent functional potential (Capabilities), specific knowledge that can enable potential harm (Domain Knowledge), the mechanisms and environmental conditions enabling interaction and influence (Affordances), and the spheres where its effects are projected (Impact Domains). This structure ensures that the analysis systematically covers the origin, pathway, and endpoint of AI-driven hazards. The more granular lower levels (TL3-TL4) also follow this top-down derivation and are further populated and validated through a bottom-up process incorporating illustrative examples of known failure modes identified from literature and practice.

This aspect-oriented hazard taxonomy provides a structurally comprehensive index for the vast space of AI-driven hazards, guiding assessors in generating, situating, and contextualizing specific technosocial concerns. Its practical utility is demonstrated by its ability to encompass and organize analysis across diverse hazard subsets. For example, findings from specialized analyses – such as impacts on human rights, democracy, and the rule of law identified via dedicated assessments (CAI, 2024), or hazards identified by analyzing policy documents for deviations from codified societal expectations (Zeng et al., 2024a) – naturally align within the taxonomy's Impact Domains (primarily Individual and Societal). Similarly, hazards originating from the system itself – stemming from AI Capabilities, high-risk Domain Knowledge, or operational Affordances – are also systematically categorized, providing structure for analyzing areas like cybersecurity or biological hazards.

While the taxonomy aims for comprehensive structural coverage at its higher levels (TL0-TL2), the dynamic and vast nature of AI has led to the fact that the lowest levels of hazard clusters (TL3) and individual hazards (TL4), though populated with derived hazard entries, are not meant to be exhaustively pre-defined. However, the progressively detailed structure facilitates targeted exploration. During an assessment, assessors can further populate, refine, and prioritize the relevant TL3/TL4 entries by drawing inspiration from specialized resources. For instance, established cybersecurity hazard frameworks like MITRE ATT&CK® (MITRE, 2025) can be adapted to ideate novel AI-specific attack vectors or misuse scenarios that fit within the taxonomy's categories; e.g., adapting the 'Phishing' technique under the 'Initial Access' tactic might reveal an AI-driven spear-phishing vector, which could then populate a corresponding specific hazard entry under the 'Software & AI Engineering' domain knowledge aspect (TL2). More broadly, specific AI hazards identified in analyses that are focused on potential negative interactions within various domains – such as financial systems (Financial Stability Board, 2024) or bio-security (Rose et al., 2023) – provide concrete examples that can be directly mapped into and supplement the TL3/TL4 hazard entries within this structure. Drawn from relevant domain expertise, these resources help ground the assessment in concrete, recognizable hazards and leverage specialized knowledge for refinement. Furthermore, incorporating these findings ensures the assessment remains adaptable to emerging threats while maintaining coherence within the overall framework.

Aspect-oriented hazard analysis employs three complementary analytical approaches to systematically examine potential risks identified via the taxonomy: bottleneck analysis, competence-incompetence analysis, and aspect interaction analysis. These provide structured methods for identifying both direct technical hazards and more complex emergent risks that might arise from the interaction of system properties.

Bottleneck analysis. Enumerating all possible harms from AI systems is intractable given their vast number of potential pathways. Bottleneck analysis addresses this by examining how each aspect – whether a relatively atomic capability such as a particular type of reasoning, domain knowledge such as cybersecurity expertise, an affordance such as particular API access, or an impact domain such as collective epistemics – could become the critical constraint or enabler of harm. The analysis treats each aspect in turn as the potential bottleneck by holding all other aspects fixed while considering what becomes possible if the focal aspect reaches its maximum plausible level for the system being assessed. This reveals critical thresholds where quantitative enhancements in an aspect could enable qualitatively different risk scenarios.

For each aspect, assessors determine the criticality levels at which new harm pathways become possible. For example, when integrative cognitive orchestration (a reasoning capability aspect) is the focal aspect, assessors examine what novel harms become possible with the maximally sophisticated level of this capability that may manifest given the possible range of resources, contexts, reconfigurations, and uses of the system, even without advances in other aspects. The analysis helps determine, for example what level of integrative cognitive orchestration enables sophisticated social manipulation or strategic deception.

Similarly, when collective epistemics (impact domain) is the focal aspect, assessors examine how the system – and what it may become within the defined assessment scope – at its maximum plausible capability level could exploit vulnerabilities in societal knowledge systems. This reveals critical dependencies – how moderate advances in societal vulnerability could suddenly enable massive exploitation by existing AI capabilities, or how weakened collective knowledge systems could unlock new categories of risk even without enhanced computational capabilities.

The key insight of bottleneck analysis is that it forces systematic consideration of how each aspect, when treated as the bottleneck, enables or exacerbates risks differentially. By examining how realistically maximizing a single aspect while holding others constant could affect the system's behavior, this analysis serves as a lens to reveal plausible threat models that might be missed in more conventional approaches. Each aspect provides a distinct perspective for identifying potential harms within the system being assessed. This helps identify which properties of the system warrant closest scrutiny and informs where additional safeguards may be needed.

Competence-incompetence analysis. Advanced AI systems create risks through a fundamental bifurcation that manifests in two distinct categories of hazards:

- **Competence-based hazards.** Arise when highly effective capabilities lead to harmful consequences (succeeding at something we don't want);
- **Incompetence-based hazards.** Stem from system limitations or failures (seemingly trying but failing to do something we want).

While one could argue all failures reflect some underlying specification or design inadequacy, distinguishing between failures of execution (incompetence-based) and harms resulting from successful execution (competence-based) provides a valuable analytical lens. This distinction is particularly crucial for advanced AI due to the uneven advancement across different skills and domains (i.e., "jagged technological frontier", or "capability discontinuities") (Dell'Acqua et al., 2023). As capabilities develop, AI systems may exhibit sophisticated performance in potentially harmful domains while simultaneously lacking critical safety-relevant competencies. For instance, when an AI system shows high competence in reasoning and planning but demonstrates incompetence in understanding ancillary consequences and following safety guidelines, it can cause harm through competent execution of unsafe plans. This pattern of advanced capabilities coexisting with significant blindspots – including hallucinations, confusions, and silent failures of reasoning – creates risk profiles fundamentally different from traditional software systems.

Established evaluation metrics (e.g., accuracy, robustness, fairness) readily quantify specific types of incompetence-based failures (Bengio et al., 2025; Jones et al., 2024). Competence-based hazards,

involving systems performing well but in ways that are harmful or unintended, are often harder to predict and assess comprehensively. This difficulty arises because evaluating these risks extends beyond verifying conformance to predefined functional specifications or testing for known types of execution failures. Instead, it requires identifying harmful outcomes emerging from a broad spectrum of potential system behaviors, many of which may successfully achieve goals in ways not easily specified in advance or captured by standard failure analysis. While targeted tests for specific issues like deception are developing (Järviniemi et al., 2024; Meinke et al., 2025), identifying the wide range of potential harmful potency realizations remains challenging compared to simply identifying functional failures, yet uncovering an ample set of competency-based hazards is particularly crucial, as such hazards are expected to generate larger harms. This dual competence-incompetence perspective is therefore critical for disentangling threat models and avoiding an imbalanced focus on only one type of hazard; it directly informs threat model development and the scenario generation process (see Section 4.3) — illustrated by concrete competence/incompetence examples across severity levels in Appendix F — ensuring both hazard categories, and their combinations, are considered during the assessment.

Applying this dual analysis reveals that highly capable AI systems can create larger risks through exceptional performance in harmful directions than through critical failures in expected functionality, though the latter can also be quite consequential in combination with the former (for further analysis see Wisakanto et al., forthcoming-d). Moreover, critical combinations of high competence in some areas with incompetence in others often uncover novel risk pathways. The practical relevance of analyzing both categories is underscored by recent large-scale evaluations, which demonstrate that current models exhibit significant failures stemming from unwanted competence (e.g., generating harmful advice when prompted) alongside traditional errors (Zeng et al., 2024b).

Aspect interaction analysis. Risks often emerge from unexpected combinations of system aspects. The framework guides assessors in examining how pairs of AI system aspects can interact and create risks beyond those identified when analyzing aspects in isolation. Building on established systems theory principles, particularly the study of emergent properties, this analysis recognizes that component interactions often produce effects greater than the sum of individual parts. For each aspect pair (e.g., capabilities and domain knowledge), assessors evaluate how they could combine to:

- Enable new risk pathways not possible with either aspect alone;
- Amplify existing risks through synergistic effects;
- Transform risks qualitatively through novel interaction or combination patterns.

The analysis of higher-order interactions between three or more aspects presents increasing combinatorial complexity but can reveal critical risk pathways not visible in pairwise analysis. For instance, the combination of advanced logical reasoning (capability), detailed cybersecurity knowledge (domain knowledge), and privileged system access (affordance) could enable sophisticated attack planning that would not be possible with any subset of these aspects. This interaction analysis also has potential for sensitivity testing to identify which aspect combinations produce significant risk amplification, helping prioritize evaluation efforts. These higher-order risks are most appropriately considered via cataloging known and ideated threats rather than exhaustive iteration over the space.

3.2 Risk Pathway Modeling

Risk pathway modeling is an intuitive technique for analyzing how aspects of AI systems can enable or trigger harms, whether through direct technical routes or via pathways that transmit through sociotechnical systems. It bridges the gap between technical system assessment and broader societal risk assessment by tracing directed causal chains from source aspects (e.g., a specific capability) through intermediate states (which may be technical, environmental, or social) to terminal aspects (which include societal impacts but can also encompass critical infrastructure system failures or compromise). This ensures the analysis covers both acute technical risks originating within the system and broader systemic consequences arising from its interactions with the environment.

Risk pathways can manifest through both direct and systemic effects. Direct pathways involve rapid propagation of harms – for instance, the exploitation of a security vulnerability leading to immediate system compromise (HM Government, 2023). Systemic pathways involve complex interconnected effects that can fundamentally alter societal structures, such as how AI-generated misinformation

can erode trust in institutions and degrade collective decision-making capabilities. Understanding both types is crucial for comprehensive risk assessment. While direct harms might require immediate response, systemic risks can transform fundamental societal structures and capabilities in ways that may be harder to detect and address.

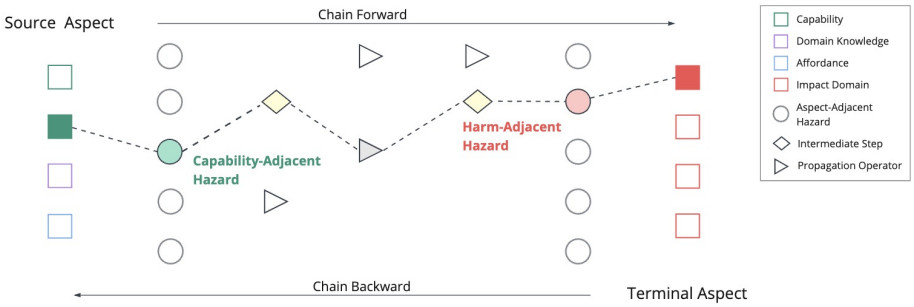

Figure 2: Causal chain illustrating a risk pathway that can be analyzed forward from source aspects (capabilities, domain knowledge, affordances) or backward from terminal aspects (impact domains). Each intermediate step represents a change in hazard state, and propagation operators represent risk transmission mechanisms.

Risk pathways consist of six fundamental elements:

1. **Source aspects.** Source capabilities, domain knowledge, or affordances of the AI system that could initiate a risk pathway, and have the potential to cause harm.

   Examples: Integrative cognitive orchestration, coding knowledge, biochemistry knowledge, unfettered Internet access.

2. **Source aspect-adjacent hazards.** The specific hazards that emerge directly or causally soon after from source aspects of the AI system and are the initial points where system characteristics could enable or trigger harm pathways.

   Examples: Circumvention of safety guidelines, bypass of security controls, weaponization of domain expertise, exploitation of system privileges.

3. **Intermediate steps.** States through which risks propagate, defining the sequence of transitions of risks from source to impact.

   Examples: Exploiting industrial control systems, identifying vulnerabilities, targeting biological containment systems, manipulating payment validation systems.

4. **Propagation operators.** Mechanisms that characterize how risks transmit and transform (including amplification) as they move through societal systems, mapping how risks cascade into broader impacts.

   Examples: Adversarial exploitation, targeted misuse, accumulation, compounding.

5. **Terminal aspect-adjacent hazards.** Vulnerabilities through which risks manifest as concrete harms to societal systems, representing the penultimate stage before terminal aspects.

   Examples: Infiltration of power grid controls, compromise of the emergency service authentication chain, breach of biosecurity containment, exploit of financial system transaction verifications.

6. **Terminal aspects.** Domains that are negatively impacted or impinged upon, where harms ultimately manifest, and the endpoints of risk pathways.

   Examples: Societal infrastructure & institutions, individuals' bodily structure, ecosystem processes & life cycles, individuals' economic & opportunities.

While terminal aspects represent the final element of the risk pathway, harms describe the specific negative outcomes actually realized within these domains when the pathway completes. For instance, a harm like widespread societal disruption due to prolonged power outages manifests within the *impact domain* of societal infrastructure & institutions. Similarly, a failure of emergency response leading to preventable deaths would manifest in the *impact domain* of individuals' bodily structure, while an AI-optimized microbe, designed for targeted agricultural benefit, proliferating beyond control could affect the ecosystem processes & life cycles *impact domain*. A systemic financial collapse causing widespread economic hardship could ultimately manifest in the *impact domain* of individuals' economic & opportunities.

The framework employs two complementary analytical approaches to ensure sufficient coverage of potential risk pathways:

1. **Forward chaining (source to terminal).** Charts out multi-step processes leading to terminal harms. Aided by analytical tools such as event trees and fishbone diagrams.

   Example: Advanced reasoning capability → exploitation of cybersecurity vulnerability → critical infrastructure disruption → societal harm

2. **Backward chaining (terminal to source).** Begins with potential harms (both known and newly identified) and works backwards with structured reasoning to identify their enabling and contributing aspects. Aided by analytical methods such as fault trees and root cause analysis.

   Example: Mass societal manipulation ← prerequisite: advanced psychological modeling ← source: sophisticated reasoning + human behavior knowledge

The two approaches offer complementary strengths, enriching risk assessment by providing distinct analytical lenses. Forward chaining explores potential risk pathways, using any available data to model how initiating aspects may lead to harm, while backward chaining, grounded in real-world harm cases, ensures the analysis remains connected to tangible outcomes. Together, they help identify non-obvious pathways and assess the completeness of forward analyses, creating a robust framework for hazard exploration.

Figure 2 illustrates this conceptual approach, showing a causal chain that can be analyzed either forward from source aspects (capabilities, domain knowledge, affordances) or backward from terminal aspects (impact domains). Each intermediate step represents a change in hazard state, and propagation operators represent risk transmission mechanisms.

The framework employs several key analytical tools and techniques in risk pathway modeling. Among these are societal threat surface analysis for mapping the set of pathways through which AI systems could harm society, prospective risk analysis for reasoning about novel failure modes, propagation operators for characterizing how risks transmit through systems, and focused aggregation for alternative mappings of risk scenarios to higher-order risk dimensions. Each of these components provides distinct capabilities for understanding and evaluating risk pathways.

Societal threat surface analysis. The risks posed by AI systems to society extend far beyond direct technical failures or misuse, propagating through intricate webs of societal dependencies in ways that traditional component- or model-focused assessments invariably fail to capture systematically. The societal threat surface – the set of pathways through which AI systems could harm society and its supporting biosphere – provides a conceptual foundation for systematically analyzing these broader impacts. This surface encompasses vulnerabilities and points of interface where AI capabilities, domain knowledge, and affordances may generate both direct effects and cascade effects across interconnected societal structures.

This conceptual advance reframes risk assessment by shifting the focus from individual capabilities or propensities to the broader systems they interact with, enabling systematic exploration of potential harm pathways. The framework operationalizes this through the aspect-oriented taxonomy (see Section 3.1) that maps both source aspects (capabilities, domain knowledge, affordances) and terminal aspects (impact domains) where harms manifest. This structured decomposition supports identification of aspect-adjacent hazards – potential harms emerging from specific AI system capabilities or propensities as they interact with societal systems.

The societal threat surface guides assessment methodology in several key ways. It informs systematic sampling of the hazard space by providing a principled basis for identifying which combinations of system aspects warrant evaluation.

Prospective risk analysis. Assessing risks from AI systems requires imagining and reasoning about hazards that haven't yet manifested, rather than relying solely on historical patterns. Traditional probabilistic risk analysis extrapolates from known failure modes within bounded systems – a rocket may fail catastrophically, but its impacts remain constrained within well-understood limits. Advanced AI systems, by contrast, can generate novel failure modes that transform the very context in which they operate, potentially leading to unbounded harms.

Current evaluation methods typically fail to detect behaviors that only emerge at scale (Jones et al., 2025). Waiting for empirical evidence before acting on potential AI risks creates a systematic blind spot in governance, thereby leading to the neglect of risks posed by AI systems (Casper et al., 2025; Chan, 2024). This challenge necessitates a forward-looking analysis approach.

The framework addresses this challenge by combining multiple analytical approaches to build reasoned models of potential risk pathways, even where no historical precedent exists. The framework integrates theoretical capability analysis, empirical testing data, and systematic exploration of potential failure modes to construct assessor-justified assessments under documentable uncertainty. These analytical techniques support evidence-informed threat modeling that evolves alongside advancing capabilities, allowing assessment across varying levels of evidence empiricism while maintaining analytical consistency. The analysis comprises three key principles:

- **Systematic exploration.** Structured approaches for identifying novel failure modes and interaction effects, and systematically searching for what might have been missed.

  Examples: Red teaming results, adversarial testing, whitebox counterfactual analysis, emergence studies.

- **Extrapolative analysis.** Using available information to project capability trajectories and identify potential threshold effects by projecting forward from what we know.

  Examples: Capability scaling laws, emergence pattern analysis, transition indicator analysis, trend forecasting, causal models.

- **Evidence integration.** Combining multiple sources of theoretical and empirical evidence to form prospective assessments.

  Examples: Bayesian hierarchical modeling, mixed methods, structured expert judgement protocols, triangulation, systems models, knowledge graphs, model ensembling.

In prospective risk analysis, assessors typically employ complementary analytical tools. For example, *capability scaling analysis* helps identify threshold effects where quantitative improvements could enable qualitatively different risks; *multi-agent interaction studies* examine how novel behaviors might emerge in multipolar environments; *model organisms of misalignment* provide controlled experimental settings to analyze potential failure modes; and *whitebox testing* examines the internal workings and decision processes of systems to probe the boundaries of system behavior (see Appendix G for additional techniques). These techniques allow assessors to infer significant technical and sociotechnical follow-on possibilities – from novel capabilities emerging from system interactions to cascading effects through social and institutional systems. Together, these and other analytical techniques help explore an unbounded risk landscape, providing inputs for risk pathway modeling and mapping the evolving societal threat surface through technical, behavioral, and systemic indicators of emerging risks.

This analysis necessarily acknowledges and embraces that AI risk assessment must reason about risks even where direct empirical evidence is not yet available. Rather than relying solely on demonstrated failures and harms (Bommasani et al., 2024), it employs analytical techniques for inferring potential developments, uses, interactions, and downstream effects – such as feedback loop mapping (Malinowski, 2019), cascade effect modeling (Zuccaro et al., 2018), and coordination failure mapping (Basnett et al., 2014) alongside more technically grounded hybrid analytical-empirical techniques such as latent adversarial training (Che et al., 2025), thought flow tracing (Lindsey et al., 2025), control flow tracing (Montagu et al., 2021), and robustness regime mapping (Anderson et al., 2024). These analytical tools help characterize system behaviors and potential

failure modes, informing the analysis of how technical capabilities interact with social systems and institutional responses, revealing potential pathways from current trends to novel risks. The analysis begins with clear signals in bounded domains – such as specific capability jumps or bounded system behaviors – and methodically expands the analysis to related domains and interaction effects. By combining these analytical projections with theoretical models and some empirical testing, assessors can build justified assessments of unprecedented risks. This "epistemic bootstrapping" approach allows assessors to begin mapping the risk landscape by starting with confidently known information, then systematically building toward less certain domains. It leverages limited but reliable knowledge to construct justified assessments about unprecedented risks, creating a bridge from well-understood risks to reasonably anticipated but previously unobserved hazards.

Propagation operators. Direct risks from AI systems rarely remain contained – they amplify and transform as they propagate through societal systems and cause harm, often in ways that are difficult to predict from just initial testing. The propagation operator analysis provides assessors with a categorized set of transmission mechanisms for tracing these evolving impacts. Rather than treating risks as isolated technical failures, these operators characterize specific ways that initial and intermediate risks can transform and cascade into broader impacts, using both technical and sociotechnical propagation mechanisms. A complete categorized set of operators and their descriptions is provided in Appendix C.

To support systematic analysis of these complex transformations, these operators enable several key analytical approaches:

- Generating risk pathway variants by applying different operators to existing pathways;
- Identifying novel risk pathways by applying operators to aspect-adjacent hazards;
- Analyzing cascading sequences where multiple operators amplify effects;
- Detecting critical thresholds where operator effects suddenly intensify;
- Tracing cross-domain propagation of harms through different operators;
- Revealing dependencies between risk pathways through shared operators;
- Mapping temporal evolution of risks through operator sequences.

Applying these analytical approaches effectively requires assessing the characteristics and impact of the specific propagation operators involved in each risk pathway step. The analysis of these operators – assessing their effect on risk transmission, transformation, and amplification within sociotechnical systems – can be approached at different levels of rigor, primarily dictated by the availability of relevant data and sociotechnical expertise time to produce models for the specific pathway step:

- Quantitative analysis: Uses specific metrics or validated models (e.g., network, economic, agent-based simulations) to directly estimate operator effects. Often challenging due to distal societal effects.
- Semi-quantitative analysis: The most common approach for societal steps. Employs structured expert judgment, grounded in available evidence (e.g., historical analogues, social science, related events, proxies, system characteristics). Enables reasoned estimation when quantification is infeasible.
- Qualitative analysis: Essential for exploring novel or highly uncertain pathways. Focuses on identifying causal links and characterizing interaction dynamics without assigning numerical estimates.

Regardless of the chosen analysis level, the framework requires documenting the method, the supporting evidence (or lack thereof), and the reasoning behind the assessment for each significant propagation step.

The following two examples illustrate the types of multi-step risk transmissions and transformations that can be analyzed using propagation operators:

First, risks from an AI system's classification behaviors can propagate through *skew* in automated decisions, where systematic biases in healthcare diagnoses accumulate through both periodic *accrual*, i.e. small harms adding up to a large harm, with the likelihood of a larger harm increasing with mass system use, as well as other systems adopting similar models and cause risk to spread through

correlation across healthcare providers, e.g. a given group becomes systematically disadvantaged in some new way. These effects compound through *entrainment* as medical practices adapt to rely on these systems, ultimately manifesting as *public health effects* across vulnerable populations.

Second, risks from an AI system's code generation capabilities might propagate from *untargeted misuse* by unwitting and careless developers to *automated exploitation* of further vulnerabilities through self-replicating scripts, leading with some iterations by the system to *correlation* of security risks across critical infrastructure, until relevant *information asymmetry* from external (e.g. law enforcement) opacity into this use eventually enables coordinated attacks that produce systemic *economic effects*.

In threat model development, propagation operators add to the risk pathway modeling framework by providing a library of semi-structured mechanisms to model how risks evolve between each step of a pathway. By explicitly considering how risks propagate through each mechanism type, assessors can better identify potential cascade effects, ideate risk pathways for construction, and prevent some analytical blind spots. This helps move beyond simple linear scenarios to capture complex, multi-dimensional risk pathways that might otherwise be missed.

Focused aggregation. Reducing complex AI risk assessment results to a single system-level risk metric loses critical information about how different risks manifest in society and obscures differences between risk types and their interactions. This limitation highlights the need for more systematic methods capable of mapping specific technical risk findings onto broader dimensions of impact. To meet this need for more meaningful representation, the PRA for AI framework introduces focused aggregations to represent and analyze collections of risk scenarios through structured mappings to higher-order risk dimensions. This approach enables nuanced understanding of how different aspects of AI systems contribute to distinct categories of societal risk. The value of such structured mappings is increasingly acknowledged, reflected in related work connecting these risks to regulatory and societal contexts (Eisenberg et al., 2025).

The core methodological feature is that individual risk scenarios can be mapped to categories representing respective dimensions of risk through assessed relationships. The assessor can tag each risk pathway with particular risk categories, such as how hacking scenarios relate to critical infrastructure failure risk, or how disinformation pathways map to governance breakdown risks.

The methodology supports multiple categorization schemes while retaining information about the underlying causal relationships. This enables assessment data to be analyzed through different lenses relevant to various stakeholder needs – upstream assessment stakeholders may adopt alternative categorization schemes that better serve their specific evaluation needs, such as internal development targets, voluntary safety commitments, or regulatory requirements. Because these categories remain consistent across assessments using the same scheme, assessment requestors and interested parties can systematically track coverage and compare risk profiles across different AI systems.

Focused aggregation helps bridge technical and governance needs in a few different ways: enabling integration with existing risk aggregation frameworks; providing social scientists with metrics meaningfully grounded to societal risks; supporting meaningful comparison of risk profiles across different AI systems; and providing coverage tracking across risk categories to help stakeholders ensure all relevant risk categories have been evaluated per scenario. As these focused aggregation mappings become more objective and generalizable, they will increasingly support structured analysis of how AI system properties contribute to various dimensions of risk.

3.3 Uncertainty Management

Estimating probabilities and magnitudes of impact can be challenging even for experienced assessors in mature fields. Instead of attempting to estimate probabilities for an entire complex scenario at once, a fundamental principle of the framework's uncertainty management is the decomposition of complex scenarios into discrete pathway components. This decomposition breaks risk scenarios into manageable chunks that can be posed as specific, quantitatively modelable questions. PRA for AI handles this estimation challenge with a combination of techniques that are industry standard practice elsewhere and tools adapted for the purpose. The challenges of AI risk assessment demand that the framework enable assessors to:

- Construct strong initial threat models by thorough sampling of the hazard space;

- Trace the causal steps that turn an initial threat into a harmful outcome;
- Generate outputs amenable to analysis of interactions, propagations, and sensitivities;
- Make reasoned severity and likelihood estimates for each step;
- Document the evidence, reasoning, and assumptions used in the assessment.
- Reconcile divergent estimates via structured protocols when assessing as a team.

Consequently, the methodology provides:

- Classification heuristics containing intuition pumps for threat modeling;
- Scenario decomposition techniques drawn from industry standard assessments elsewhere;
- Optional methods for modeling the interaction and transmission of risks;
- Severity and likelihood intensity rubrics;
- Uncertainty tracing protocols for structured documentation throughout the process.
- Dialectic protocols for assessor comparison and recalibration.

In this section, we briefly outline each of the above tools and techniques.

Classification heuristics. Classification heuristics provide intuition pumps that help assessors calibrate their understanding of AI system aspects by offering examples and progression levels. Capability level progression tables, implemented in the workbook tool (see Section 4) and excerpted in Appendix D, map the development of general abilities including integrative cognitive orchestration, planning, and strategic optimization. Domain knowledge levels, also implemented in the workbook tool and excerpted in Appendix E, characterize the progression of domain-specific capabilities, knowledge, agency, and reasoning patterns within high-risk areas, from adversarial reasoning in cybersecurity to mechanistic understanding of biological systems.

The Risk Detail Table, a component of the workbook tool (see Section 4) excerpted in Appendix F provides plausible scenarios for each aspect at each severity level. Following the competence-incompetence analysis framework (discussed in Section 3.1), these illustrative examples are further subdivided based on whether they primarily emerge from competence or incompetence of the AI system being assessed. This dual analysis is particularly important at jagged capability boundaries where systems may exhibit sophisticated domain-specific reasoning while lacking crucial competencies relevant to safety.

Scenario decomposition techniques. PRA for AI adapts established risk assessment methods to decompose uncertainties and assess them sequentially along the risk pathway (detailed in Section 3.2). Event trees map forward towards a plausible chain of events, while fault trees work backward to identify paths leading to harms. For a list of example analytical techniques that can be employed by assessors during decomposition, see Appendix G. These complementary approaches help make complex AI risk scenarios more tractable by decomposing them into analyzable components.

Optional methods. For more in-depth analysis of risk interactions and transmission paths, the framework provides tools including Propagation Operators and the Aspect Interactions Matrix, which are implemented as structured templates in the workbook tool. These tools are further discussed in Section 4.3 and Appendix C.

Intensity rubrics. To model the complexity of potential AI harms, the framework adapts established PRA techniques for interval-based estimation (Apostolakis, 1990; Shortridge et al., 2017) by introducing two types of intensity rubrics. These provide structured definitions and reference examples to guide assessment, and are operationalized as standardized scales in the workbook tool. These rubrics employ coarse-grained bands to categorize severity and likelihood. This banding approach is central to enabling reasoned analysis under uncertainty; it makes complex estimations more tractable and helps mitigate the false precision inherent in seeking exact point probabilities for unprecedented events. These rubrics define the following types of levels:

1. *Harm Severity Levels (HSL)* categorize the magnitude of potential impacts across multiple societal dimensions (e.g., human casualties, economic damage, environmental damage), enabling structured evaluation against concrete reference points within defined severity bands.

2. *Likelihood Levels (LL)* categorize the probability of occurrence using defined odds bands, enabling consistent estimation across scenarios and assessors, especially in the absence of sufficient historical frequency data, a common situation for novel AI risks.

The framework employs these defined bands, a practice consistent with interval-based approaches like probability bounds analysis used in risk assessment (Shortridge et al., 2017).

Uncertainty tracing. The framework necessitates that assessors explicitly document their reasoning throughout the assessment process, aligning with common practices in risk assessment that utilize formal protocols for documenting theoretic-empiric rationale development (Hemming et al., 2018). Within the workbook tool (detailed in section 4), this requirement is operationalized through the Risk Assessment Entry Log. Irrespective of implementation specifics, the documentation for each assessed risk pathway must capture:

- The initiating aspect(s) and key pathway steps;
- Key assumptions and their justifications;
- Quality and relevance of available evidence;
- Known uncertainties and potential biases;
- Assigned Harm Severity Level (HSL) and Likelihood Level (LL) estimates for relevant steps;
- Assigned societal risk dimension(s) (for focused aggregation);
- Sensitivity of results to changes in critical assumptions;
- Rationale for propagation operators used or second-order interactions identified.

This documentation traces how uncertainties interact and compound throughout the assessment. For proper context, assessors must also record relevant system-level information that captures the AI system's core characteristics, including its architecture, deployment context, intended use cases, and implemented safeguards (see Appendix I). The resulting documentation enables grading or verification of the assessment and creates a foundation for future reviews, audits, and replication attempts.

Assessor recalibration protocols. Assessors evaluate available data from multiple sources – such as empirical observations, capability benchmarks, whitebox testing, and theoretical analyses – based on predictive power and relevance to the current threat model. Different assessors may arrive at varying likelihood and severity estimates based on their expertise and assumptions (see Section 5.3). To reduce extreme variance, the framework helps assessment teams navigate the inherent subjectivity in probabilistic risk estimation through a structured recalibration process adapted from the Delphi technique[3] (Markmann et al., 2013; Wilson et al., 2023).

4 The PRA for AI Workbook Tool

Having established the conceptual framework and key methodological advances in PRA for AI, this section describes how the framework has been implemented in a workbook tool, establishing the protocols and infrastructure for future practical applications. The full documentation and assessment materials for this methodology, including the PRA for AI workbook tool and user guide, are available on the project website.

Figure 3 presents the assessment workflow implemented in the PRA for AI workbook tool. The diagram illustrates the iterative process through which assessors evaluate risks across the aspect-oriented taxonomy – from aspect selection through scenario generation, severity analysis, and likelihood determination. The workflow incorporates options for analyzing cross-aspect interactions and risk propagation mechanisms. Upon completion, the tool aggregates risk level estimates into a report card that quantifies the assessed risks of the AI system across various dimensions of analysis.

[3]This adaptation provides a structured recalibration process where assessors examine the reasoning behind differing risk estimates. This allows for consideration of risk pathways and interactions that may have been overlooked while selecting the highest reasonable estimates to prevent underestimation of risks.

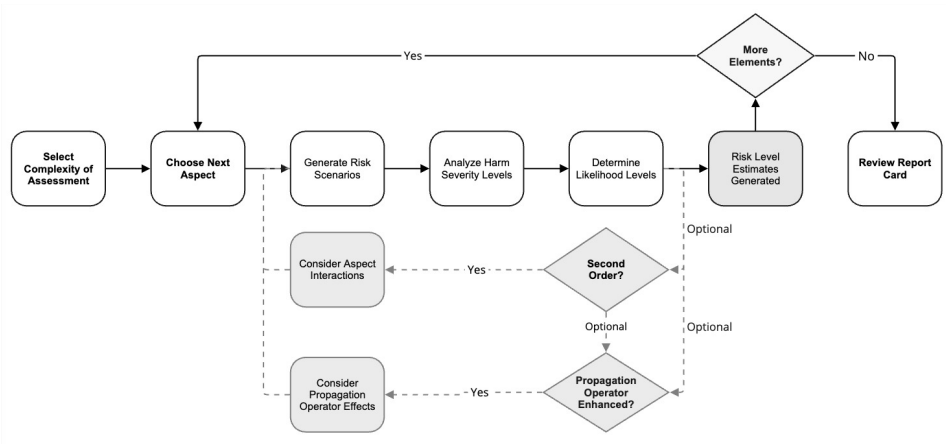

Figure 3: An overview of the risk assessment process flow in the PRA for AI workbook tool.

4.1 Tools Structure and Usage

The workbook implements the PRA for AI framework and guides assessors through the assessment process. It introduces a Risk Assessment Entry Log in which assessors document system information (see Appendix I), scenarios, and estimates. Additionally, the workbook includes a Risk Detail Table, which acts as the primary tool for scanning the taxonomy during the iterative assessment process, as well as additional tools for risk scenario development, uncertainty management, and risk estimation.

Throughout the assessment, assessors document in the Risk Assessment Entry Log the underlying threat models, assumptions, steps, calculations, uncertainties, and decisions for each risk scenario (see Section 3.3). The framework also provides classification heuristics to guide and inform threat model development. This includes Plausible Qualifiers – concrete examples of scenarios across harm severity levels per each aspect group, available in the Risk Detail Table, as shown in an excerpt in Appendix F. These serve as reference points and intuition pumps during threat model development with both competence and incompetence hazards. The framework includes Capability Levels and Domain Knowledge Levels Tables, which characterize progressions of AI system competencies in standardized frameworks across the aspect groups. For advanced assessment protocols, the Aspect Interaction Matrix and Propagation Operators guide analysis of interaction (see Section 3.1) and propagation effects (see Section 3.2). The framework guides assessors in decomposing complex scenarios into approachable and tractable steps for analysis before aggregating estimates (see Section 3.2). The assessment process is supported by intensity rubrics, the Harm Severity Levels Definition Table and the Likelihood Levels Table, that provide standardized scales and deltas from reference example to inform assessor inputs (see Section 3.3). The assessment generates three outputs. The Report Card provides aggregated risk estimates from all assessed scenarios. The Tallied Risk Matrix enumerates scenarios by severity and likelihood. The Risk Assessment Output Log preserves the full set of assessed scenarios and their supporting rationales.

The following sections detail the assessment process using the workbook tool: Section 4.2 outlines the configuration options and Assessment Maturity Levels (AMLs) that determine evaluation scope and depth, Section 4.3 describes the step-by-step assessment workflow and documentation requirements, and Section 4.4 explains the generation and structure of assessment outputs including the report card, risk matrix, and documentation logs.

4.2 Assessment Configuration

The PRA framework offers configurable Assessment Maturity Levels (AMLs) to address the diverse landscape of AI systems and organizational contexts. This tiered approach is valuable because the resources (e.g., time, expertise, data access) available for risk assessment vary significantly, as does the potential impact and complexity of the AI system under review. AMLs allow an organization to tailor AI risk assessments by selecting a protocol that balances the appropriate analytical depth and coverage against practical constraints, informed by its available expertise, resources, and assessment infrastructure. This adaptability supports effective application across diverse organizational time

constraints, regulatory and compliance requirements, and the technical complexity of the AI system being evaluated.

The protocols are represented by a three-digit number code and range from shallow (AML-008 to AML-020) to deep (AML-110 to AML-221). Lower-numbered protocols focus on broad aspect groups of an AI system, while higher-numbered protocols incorporate more complex analyses, such as second-order effects, propagation operators, and assessment at finer taxonomic granularity (moving from aspect groups at taxonomy level 1 to individual aspects at taxonomy level 2). The framework requires selecting the appropriate AML before starting the assessment process. A complete specification of AML protocols and their configurable assessment options (such as aspect group, second-order, and pathway assessment) is provided in Appendix H.

The assessment can be performed by a single assessor or a team, though team-based assessments are strongly recommended. Where team assessments yield significantly diverging risk level estimates, a structured recalibration protocol prompts assessors to examine and narrow differences in their underlying reasoning. Each assessment generates an Assessment Type Code[4] that records the AML protocol used, the protocol version, and whether it was conducted by an individual or a team. The assessment type is configured through two dropdown menus in the workbook interface – one to select the AML protocol and another to specify team or single assessor mode. After selecting an appropriate AML protocol, assessors document key system information in the Risk Assessment Entry Log (see Appendix I). This standardized tracking provides context for interpreting risk estimates and enables structured comparison between assessments, while ensuring transparency in how each evaluation was conducted.

4.3 Assessment Process

The assessment workflow follows an iterative process where for each system aspect, assessors analyze how it could enable or exacerbate hazards. After selecting the appropriate AML protocol, assessors proceed through the following steps:

Choose next aspect. Assessors begin by scanning the aspect-oriented taxonomy of AI hazards, and iteratively examining each element of the relevant taxonomy level (see Section 3.1). The selected AML protocol determines which taxonomy level is used for structuring the analysis. Lower-numbered protocols focus on aspect groups (TL1), while higher-numbered protocols extend to individual aspects (TL2).

Generate risk scenarios. The workbook contains templates that guide assessors in developing threat models. Assessors first identify possible risk pathways for each AI system aspect – capabilities, knowledge domains, affordances and impact domains – against the societal threat surface. Using bottleneck analysis, they consider how each aspect could become the critical constraint or enabler of harm from the AI system (see Section 3.1) to develop threat models. Assessors then translate these threat models into specific scenarios by constructing directed causal chains that show how harms could materialize. For each identified risk pathway, assessors develop detailed narratives by specifying the initiating event that enables the threat, the intermediate steps through which risks could propagate, and the terminal aspects where harms materialise in societal systems.

As part of this process, risk pathway modeling traces how source aspects could cause a chain of events that ends in harm (see Section 3.2). Each link in this chain represents a discrete transformation, such as a change in system state, deployment of a capability, or crossing of a critical threshold. The Capability Levels Table (see Appendix D) and the Domain Knowledge Levels Table (see Appendix E) help assessors identify these transitions, particularly where measured improvements in capabilities could enable qualitatively different risk pathways. For scenarios with complex pathways, assessors can employ both forward projection from source aspects and backward analysis from impact domains.

Assessors can employ several analytical techniques while developing detailed scenarios, including fault tree analysis for mapping branching paths from incident to harm, expert elicitation using structured protocols for novel risks, and scenario discovery for surfacing non-obvious hazards. When appropriate, assessors can also draw causal influence diagrams to model decision-making structures,

[4]For instance, if AML-120, version v0.9.1-alpha, and assessor type "team" are selected, the Assessment Type Code generated will be: AML-120-v0.9.1-alpha-T.

and perform capability scaling assessment to analyze risks from increasing AI capabilities (for more techniques see Appendix G).

While the assessment primarily relies on assessors' direct expertise and real-world understanding of these domains, the workbook provides classification heuristics to aid in ideating scenarios (see Section 3.3). Assessors can use the Risk Detail Table, as shown in an excerpt in Appendix F, for two sources of guidance. First, they can consult the taxonomy of hazard clusters (TL3) and hazards (TL4) for concrete examples of risk by aspect. Second, Plausible Qualifiers provide escalating examples of harm that allow assessors to reason about impacts through different causal pathways. Plausible Qualifiers are available for both competence scenarios (where exceptional performance creates risks) and incompetence scenarios (where limitations or failures lead to harms) at each severity level, as shown in an excerpt in Appendix F).

Analyze harm severity levels. For each risk scenario, assessors analyze potential impacts using Harm Severity Levels (coded HSL-1 through HSL-6). HSLs represent impact magnitude on a defined scale from marginal and non-trivial to globally catastrophic (see Appendix K). Each HSL is defined across such dimensions as human deaths, economic damage, and environmental impact. For estimating the magnitude within these dimensions assessors can draw on a range of relevant evidence, which might include sources such as historical data from analogous events, simulations or consequence modeling specific to the harm type (e.g., economic or epidemiological models), estimates based on system scale (e.g., number of affected users), or structured expert judgment particularly for novel or intangible impacts (e.g., erosion of trust). When assessing a scenario, assessors identify the relevant dimensions, estimate impact against each, grounding the estimate with objective criteria where possible (such as external benchmarks and expert expectations used as helpful reference points), and assign the highest applicable HSL.

Determine likelihood levels. Assessors estimate scenario probabilities using Likelihood Levels (coded LL-0 through LL-8). LLs represent odds ranges that decrease by orders of magnitude, from relatively common events (LL-8, odds from 1:1 to 1:10) to practically impossible ones (LL-0, beyond $1{:}10^{12}$), with reference examples provided in Appendix L. Evidence informing the LL estimate for a given scenario component or step is diverse, ranging from theoretical analysis (e.g., scaling laws) and expert judgment to empirical frequency data derived from system testing, operational monitoring, red teaming, or other analyses. To enhance consistency and defensibility, assessors should anchor their likelihood estimates using objective criteria where available, such as relevant empirical data (from benchmarks or testing) or established expert expectations, using these as helpful reference points when selecting the appropriate LL band.

For steps lacking direct evidence, common in novel AI risk assessment, assessors employ structured qualitative estimation. This iterative process involves determining the likelihood by systematically comparing the step's plausibility against general reference examples illustrating different odds ranges (Appendix L) and relevant analogies. Accounting for qualitative differences helps refine this comparison. Furthermore, considering the plausible frequency of occurrence over extended time-frames can provide a useful auxiliary perspective, especially when differentiating between very low likelihood bands where direct odds are less intuitive. This structured comparison, integrating these considerations, informs the selection of an LL band. Effectively, this is an 'epistemic bootstrapping' approach, using established references to arrive at reasoned conclusions under uncertainty. Techniques such as problem decomposition, alongside comparison against reference classes and analogies, support this structured estimation, though the specific methods and documentation detail employed may vary based on the step's complexity and assessor experience.

The workbook guides assessors in decomposing complex risk scenarios into constituent steps where both HSL and LLs are evaluated for each step of the directed pathway. Estimating the overall scenario likelihood, in particular, requires analyzing the conditional probability of sequential steps; for example, a simplified scenario probability might be decomposed as P(harmful scenario) = P(capability exists) × P(capability misused | exists) × P(harm occurs | misused). This stepwise analysis helps assessors trace how risks propagate and potentially amplify through the causal chain, requiring consideration of dependencies between steps, as the likelihood of one step occurring may influence the likelihood of subsequent steps beyond simple sequence. When a particular risk scenario could manifest through different contexts or plausible pathways, multiple HSL and LL combinations can be assigned to reflect these variations in impact magnitudes at relevant steps.

Beyond assigning initial HSL and LL bands based on available evidence and deltas from reference examples, assessors can further structure their reasoning and constrain uncertainty for critical or complex pathway steps by building explicit arguments within the documented rationale. For instance, arguments from inability may justify assigning a lower likelihood if evidence demonstrates the AI system lacks a crucial prerequisite capability for a specific harm pathway. Similarly, arguments identifying critical dependencies or limiting factors within the causal chain can provide a basis for limiting the assessed likelihood of the complete pathway manifesting – this involves analyzing specific steps to determine if they represent particularly challenging prerequisites or pathway bottlenecks, distinct from the aspect-level bottleneck analysis (see Section 3.1), that constrain the overall probability. Employing such structured reasoning patterns allows assessors to formulate more robust and defensible final risk level estimates.

For each risk assessment entry, assessors record in the Entry Log their key assumptions, rationale, potential biases in the analysis, the quality and relevance of their evidence, and any sensitivity of estimates to changes in critical assumptions. Within each aspect being evaluated, the criteria for having sampled sufficiently will depend on the aspect itself, the assessment context, and the information available to the assessor. Before proceeding to the next aspect, assessors document their rationale against these criteria.

Optional Analysis. For applicable AML protocols, assessors perform second-order assessment and propagation operator enhanced assessment.

For second-order assessment, assessors work across each matrix column in the Aspect Interaction Matrix – a tool provided to track interaction analysis – and evaluate how the aspect under consideration might interact with other aspects and create risks beyond those identified when analyzing these aspects in isolation. When a meaningful interaction is identified, assessors create a new risk scenario in the Entry Log, documenting how the interaction could enable or amplify potential harms. As with other scenarios, assessors document their reasoning and the highest estimated HSL and LL for each. This includes explaining the mechanism of interaction and how it changes their estimations.

When performing propagation operator enhanced assessment, assessors evaluate potential risk transmission and amplification pathways. The process involves reviewing the Propagation Operators Table (see Appendix C) to identify relevant transmission mechanisms for each scenario. Assessors document how each applicable operator could transform or amplify the identified risks, generating additional risk scenarios based on these transmission pathways. They then estimate HSL and LL values for each new scenario, considering the compounded effects. All scenarios are recorded in the Entry Log with comprehensive documentation of the transmission mechanisms considered and the reasoning behind the resulting risk estimates. For detailed step-by-step guidance on the assessment process, refer to the workbook tool user guide.

4.4 Assessment Outputs

Once all aspects have been assessed, the workbook tool automatically maps assessors' chosen HSL and LL estimates for each of the completely assessed risk scenarios to risk levels. The mapping is defined in the Risk Levels Table (see Appendix M). This table functions as a standardized risk matrix, common in high-reliability fields (IEC, 2019), which explicitly maps each possible combination of assessed Harm Severity Level (HSL 1-6) and Likelihood Level (LL 0-8) to a distinct, numerical Risk Level (RL 0-9).

When the assessment is performed by a team of multiple assessors, any contentious aspects are revisited. The framework employs dialectic recalibration to drill down on the estimates made and the detailed rationales provided: assessors first make independent estimates, then significant divergences trigger explicit comparison of underlying assumptions (see Section 3.3). While full agreement is not required, final estimates incorporate well-justified perspectives across the assessment team, with highest post-recalibration risk estimates selected to prevent underestimation bias.

Risk level estimates generated. The workbook tool maps each completely assessed scenario by their harm severity and likelihood to a corresponding Risk Level using the standardized Risk Levels Table (see Appendix M). These risk levels are calculated distinctly for first-order assessments, first-order propagation operator enhanced assessments, second-order assessments, and second-order propagation operator enhanced assessments. This division enables stakeholders to understand the risk levels of both immediate hazards and more indirect hazards that arise from interactions. When multiple

scenarios are generated for the same aspect group, taking the maximum risk level across the scenarios ensures that assessment insights about highest-risk pathways are not diluted by averaging or lower risk alternatives. This ensures critical risk levels remain most visible in the final output, while the separation between assessment types allows stakeholders to understand how different analytical lenses (such as first-order vs. second-order analysis) reveal distinct aspects of the system's risk profile.

Review report card. From the inputs, the framework generates three formal outputs designed to serve different stakeholder needs.

First, the aggregated risk level estimates are presented in the Report Card, which contains the system context documentation, including the assessment date, team composition, system name, version, access level, and documented system-level assumptions, to allow assessors to present the context of the assessment in a clear manner. The risk level summary then displays the aggregated risk levels for each aspect group, with separate columns showing results from each assessment type. A total maximum risk level across all aspect groups and assessment types provides a high-level indicator of the highest risk level assigned in any part of the assessment.

The report card includes focused aggregation results – both a tabular summary and radar visualization to highlight relative risk concentrations – implemented through a standardized mapping interface that supports both default systemic risk dimensions (defined in Appendix J) and custom categorization schemes defined by assessment requestors. The focused aggregation described in Section 3.2 allows assessors to map each risk scenario and their estimates to six predefined systemic risk dimensions: social fabric erosion, economic system unraveling, critical infrastructure failure, governance breakdown, environmental breakdown, and public health disintegration.

Second, the Tallied Risk Matrix enumerates all documented scenarios by their assigned harm severity and likelihood levels. This matrix shows the distribution of assessed scenarios across the evaluation space, providing insight into assessment coverage and risk concentrations.

Third, after finalizing the assessment in the report card, assessors generate a static output log – a record timestamped with the assessment completion date that captures all completed risk scenarios and their supporting rationales that determined the final risk level estimates. This static record serves as the definitive reference point for the assessment's findings.

The report card results should be evaluated in conjunction with the complete tallied risk matrix, output log documentation, and the standard report disclaimer. Furthermore, the results should be considered as one component within a broader ensemble of risk evaluation methodologies, including other quantitative and qualitative approaches. Section 5 discusses how the framework advances AI risk assessment practice.

5 Discussion

Building on the methodological foundation (see Section 3) and the practical implementation (see Section 4) of the framework, we now examine the framework's contributions to risk assessment practice, discuss its practical applications, analyze its limitations, and identify directions for future development.

5.1 Methodological Contributions to AI Risk Assessment Practice

Advancement beyond current risk assessment approaches. The PRA for AI framework advances the practice of assessing risks from advanced AI systems by introducing specific methodologies designed to help overcome limitations in current approaches regarding hazard identification, causal pathway analysis, and uncertainty quantification. These contributions provide assessors with enhanced tools for more systematic, comprehensive, and defensible evaluations.

First, the framework enhances hazard identification through broader coverage and targeted analysis via aspect-oriented hazard analysis (see Section 3.1). Rather than relying solely on ad-hoc selection, or a narrow focus on commonly cited risks, the framework requires structured scanning of the hazard space guided by a first-principles taxonomy (capabilities, domain knowledge, affordances, and impact domains). Specific analytical techniques provide further value: Bottleneck analysis shifts assessment effort from brute-force evaluations towards more targeted identification of critical

performance or vulnerability thresholds where qualitatively distinct harms may emerge, helping avoid undirected testing that might overlook high-priority threat models. Competence-incompetence analysis offers a crucial advancement by mandating a balanced consideration of risks arising from both highly efficacious AI execution yielding harmful outcomes (competence) and from system flaws or limitations preventing correct operation (incompetence). This dual lens directly counteracts the 'common myopia' regarding failure modes (see Section 2.3) by ensuring a balanced focus, which is particularly vital for systems exhibiting AI's characteristic "jagged" capability profiles. Critically, this analysis extends to examining hazardous combinations where capability enables or exacerbates failings – a key source of novel, severe risks frequently overlooked by methods analyzing capabilities. Lastly, aspect interaction analysis provides a structure for investigating combinatorial risks emerging from the interplay among AI system aspects, enabling higher-order risk analysis with broader coverage of potential failure modes than isolated capability evaluations.

Second, the framework introduces risk pathway modeling (see Section 3.2) to connect system capabilities and propensities to real-world consequences within their sociotechnical context. Addressing a common disconnect in AI assessment, it employs traceable causal chain analysis (forward and backward chaining) to link source aspects (e.g., capabilities) to terminal impacts (e.g., societal disruption). This provides end-to-end coherence often lacking when technical evaluations remain divorced from impact assessments. Unique contributions include the explicit modeling of the societal threat surface and the use of propagation operators. These tools provide a vocabulary and structure for analyzing how initial technical risks transmit, transform, and amplify as they cross system boundaries and interact with complex societal structures — addressing the recognized but rarely tackled challenge of systemic risk analysis for AI. Furthermore, prospective risk analysis techniques are integrated to enable structured, evidence-informed reasoning about novel or unprecedented failure modes, offering a crucial alternative to reliance solely on historical data or currently demonstrable failures, thereby helping to navigate the "evidence dilemma" in assessing rapidly advancing AI (Bengio et al., 2025). Finally, focused aggregation allows mapping granular pathway findings onto higher-order risk dimensions, yielding outputs more meaningful for governance than raw data or single risk scores.

Third, the framework provides techniques for uncertainty management tailored to AI assessment (see Section 3.3), enhancing consistency and defensibility. Moving beyond potentially unreliable unaided judgments, it mandates scenario decomposition for tractability. Standardized classification heuristics (e.g., capability levels) and intensity rubrics (HSL/LL bands with reference examples, including the competence-incompetence distinction in Appendix F Plausible Qualifiers) aid assessor calibration and support consistent, reasoned estimation even under uncertainty or data scarcity — a significant advantage over approaches demanding unobtainable precision. The adoption of coarse-grained Likelihood Level bands specifically adapts traditional PRA practice for AI's novelty, acknowledging inherent uncertainty without sacrificing analytical structure. Crucially, the framework requires explicit uncertainty tracing: structured protocols mandate documenting evidence, assumptions, reasoning chains, and sensitivities. This contrasts sharply with opaque assessments, enhancing transparency, reproducibility, and the critical review of outputs, ultimately contributing to a far more defensible assessment process compared to methods relying on undocumented judgments or assumptions.

The framework's methodological advances are concretely operationalized into a practical assessment workbook (see Section 4). By providing standardized documentation structures and guided workflows, including integrated templates for specific analyses, it translates the conceptual framework into a concrete instrument. This aims to facilitate consistency and lower the barrier for practitioners, offering a structured path for the complex task of AI PRA. Additionally, the workbook's structure provides a foundation for future studies examining hazard coverage and assessment consistency across different contexts and assessor profiles.

Practitioner reception and early framework development. The framework underwent iterative review with experts from high-consequence risk domains and established risk assessment fields. Their feedback specifically addressing assessment workflow clarity and risk estimate consistency directly informed our revisions of both the conceptual framework and its implementation through the workbook tool.

Reviewers particularly valued the framework's ability to systematically identify and analyze risk pathways that cross traditional assessment domain boundaries (see Section 2.2), enabling more comprehensive threat modeling. The strongest positive reception focused on three key elements: the structured approach to decomposing complex scenarios, explicit uncertainty documentation protocols,

and the bidirectional causal analysis methodology. The bidirectional approach was recognized for its operationalization of capabilities and harms – with potential to overcome cognitive biases that typically limit consideration of novel failure modes, particularly for high-risk knowledge domains, where system success at known sets of undesired tasks presents the primary risk vector. The competence-incompetence analysis was highlighted as addressing a critical blind spot in current approaches that often focus primarily on only one of either competence or incompetence based hazards.

Experts also noted how traditional threat modeling approaches , often based on canonical vulnerabilities, could fail to account for shifting distributions in attacker capabilities or emerging misuse cases. The framework demonstrates promise in identifying capability combinations that could bypass traditional risk controls through non-standard pathways, such as when models with seemingly harmless capabilities could be combined to produce greater capacity for harmful outcomes.

Additionally, reviewers identified areas for further development, suggesting, importantly, that the creation of assessor calibration exercises for before conducting assessments would strengthen the approach, along with explicit provision of additional guidance for aspect-specific tooling and more granular best practices for threat modeling. These insights will inform continued refinement of the available assessment components in subsequent versions of the workbook tool.

5.2 Practical Utility in Risk Assessment

Integration with existing risk management approaches. The AI risk assessment landscape encompasses both formal regulatory frameworks and voluntary agreements, with our PRA for AI framework designed to complement and operationalize both. Formal regulatory frameworks include the EU AI Act (EU, 2023), the Framework Act on the Development of Artificial Intelligence (Ministry of Science and ICT, 2025), and various national consumer protection regulations such as the Colorado Consumer Protections for Artificial Intelligence (Rutinel et al., 2024). These establish legally binding requirements while leaving implementation to organizations. Complementing these are voluntary agreements and standards such as the Seoul AI Safety Summit commitments (HM Government, 2024), the General-Purpose AI Code of Practice (EU, 2025), NIST AI Risk Management Framework (RMF) (NIST, 2022a), IEEE 7010 (Schiff et al., 2020), ISO/IEC 23894:2023 (ISO, 2023a), and ISO/IEC 42001:2023 (ISO, 2023b).

While these frameworks establish valuable procedures, they provide limited tools, methods, and guidance for practical implementation. These frameworks often lack detailed methodologies for quantitative risk estimation techniques, protocols for uncertainty documentation, and systematic hazard identification methodologies The PRA for AI framework addresses these methodological gaps through our methodological advancements, which can be directly incorporated into existing governance structures. First, organizations implementing these standards frequently encounter difficulties in systematically identifying, quantifying, and characterizing novel failure modes unique to AI systems, particularly those that might emerge from complex system interactions. Second, without structured risk quantification methods, organizations cannot effectively prioritize their control implementations or demonstrate that their procedures meaningfully reduce risk levels. Third, compliance documentation often lacks the detailed risk scenario analysis needed to validate that controls are effectively addressing the most consequential risks. PRA for AI complements existing standards by offering a granular approach to identifying potential failure modes and providing for a common documentation format for risk scenarios and propagation pathways for external audits and compliance reviews.

A few AI developers and service providers (Anthropic, 2025; Duffer et al., 2024) have begun implementing select standards, though compliance often means satisfying requirements in siloed, incomplete ways. Organizations adopting the NIST AI RMF can use our framework to operationalize the Map and Measure functions, while those implementing ISO/IEC standards can utilize the explicit rationale and assumption tracking of our documentation protocols to generate the auditable evidence needed to demonstrate systematic adherence to core risk management processes. The framework's explicit documentation of assumptions and probability estimates enhances transparency requirements, creating clear traceability from risk identification through to implemented controls.

The framework provides a systematic foundation for fulfilling specific requirements in the third draft of the EU General Purpose Codes of Practice (EU, 2025), including more comprehensive Systemic

Risk Analysis (II.4), Risk Acceptance Determination (II.5), and Safety and Security Model Reports (II.8, Measure 1). The framework also offers practical utility for emerging AI risk management approaches, with its structured quantitative estimation tools supporting the detailed modeling and indicator operationalization emphasized in frontier AI risk management frameworks (Campos et al., 2025). Additionally, its aspect-oriented hazard analysis directly supports the systematic identification of uses, misuses, and impacts prioritized for GPAI and foundation model profiles (Barrett et al., 2025).

Harmonizing assessment methods and operationalizing risk outputs. The PRA for AI framework functions as an integrative methodology, designed to harmonize diverse inputs from other assessment approaches into its structured analysis and operationalize risk by producing well-defined, quantified outputs usable by those methods. This dual capacity helps bridge methodological gaps and supports a more unified end-to-end evaluation process.

First, the framework serves to harmonize diverse inputs by facilitating the structured incorporation of findings from various assessment techniques directly into its risk pathway documentation. Assessors can reference specific results from red teaming exercises (e.g., demonstrated vulnerabilities), safety case arguments (e.g., mitigation effectiveness claims, operational contexts), empirical benchmark testing, or other relevant analyses as explicit evidence justifying specific pathway steps or component probability estimates, thereby strengthening the justification for the resulting risk scenarios. For instance, if red teaming demonstrates a 30% success rate in bypassing a specific safety filter under certain conditions, an assessor, evaluating this finding based on the specific test conditions, its relevance to the pathway step, comparable scenarios, and overall system context, could document this finding as partial evidence supporting a specific Likelihood Level (LL) estimate for the 'circumvention of safety guidelines' step within a relevant misuse pathway that conditionally includes those given conditions from the test, combining it with estimates for other pathway steps. This allows disparate evidence types to contribute formally to a structured PRA.

Second, the PRA for AI framework operationalizes risk by generating quantified outputs – severity and likelihood level estimates for specific pathways – that serve as valuable inputs for other assessment methodologies. Crucially, these quantitative results are always accompanied by the documented, transparent, human-interpretable risk scenarios and rationales developed during the PRA for AI process, grounding their interpretation and explaining their derivation. This pairing of quantitative results with explanatory context significantly enhances their utility for downstream assessments. For example, PRA for AI outputs can supply the necessary probabilistic inputs often missing but required for validating safety case arguments, bridging a gap where safety cases may lack numeric risk estimates. PRA also helps structure inability arguments by linking capability evaluation results to explicit low-probability estimates for specific risk pathways (AISI, 2024c). Furthermore, PRA outputs can provide a baseline for dynamic safety cases (Cârlan et al., 2024), helping identify when safety arguments require revision. For red teaming, the framework helps address a critical measurement challenge: Typically, red teaming identifies vulnerabilities without systematic risk level characterization. PRA for AI outputs provide essential context. Furthermore, red team testing is often limited to individual models in isolation, separate from the broader systems they might be embedded in (Ji, 2025), limiting the ability to assess the real-world impact of identified flaws. By quantifying the potential likelihood and impact associated with discovered flaws within modeled pathways, PRA for AI enables systematic prioritization based on assessed risk rather than just vulnerability presence. While distinct from the operationalization function, other components of the PRA for AI framework also enhance red teaming practice; the aspect-oriented taxonomy aids systematic test scoping by providing a fuller indexing of potential hazard areas derived from first principles, and the classification heuristics support more precise capability characterization by offering standardized scales and reference points for evaluating observed model behaviors.

This translation into a common format combining quantitative estimates with qualitative rationale allows diverse assessment methods to integrate probabilistic insights and compare findings more systematically. This dual capacity directly addresses one methodological challenge of integrating disparate evidence types to enable the development of a more unified end-to-end risk evaluation.

Downstream uses and applications. The PRA for AI framework can be applied broadly to quantify AI risks, providing a structured approach for translating diverse risk information and expert judgments into estimates within defined probability bands that directly inform several key applications. This quantification directly informs several key applications. Primarily, the risk estimates produced enable

organizations to implement targeted mitigation strategies by identifying which specific pathway components contribute most significantly to overall risk levels. Integrating with broader organizational processes, the framework's detailed risk pathway analysis can offer specificity to help inform control configurations within unified governance structures (Eisenberg et al., 2025), potentially allowing actions to be tailored more effectively against certain risk mechanisms. Furthermore, the methodology enhances sensitivity analysis capabilities, allowing assessors to better identify and assess how changes in system capabilities, propensities, and aspect interactions affect risk estimates and determine which factors might have the greatest impact on outcomes.

Beyond informing mitigation strategies, the framework's standardized metrics and structured documentation support crucial oversight functions. They facilitate understanding risks post-deployment and enable systematic comparisons between assessments – whether evaluating different AI systems or tracking a single system's risk profile over time. These quantitative measures are also valuable for compliance and governance, supporting the evaluation of AI systems against established intolerable risk thresholds (Raman et al., 2025) and informing the application of tiered safety approaches by providing the quantified risk levels needed to assign systems to specific tiers (Future of Life Institute, 2025). These quantified risk levels function as actionable thresholds ("points of disjuncture"), triggering specific responses like continued monitoring, mandatory mitigations, or deployment halts.

The PRA for AI framework is particularly useful when direct empirical data is limited or insufficient for threat modeling. While benchmark data provides valuable capability measurements, it is often not directly usable in complex risk scenarios. Expert elicitation is typically used to bridge this gap, informing probability estimates within specific risk pathways based on available benchmark results and other evidence (Murray et al., 2025). The PRA for AI framework extends and facilitates such elicitation practices by providing the necessary structured pathways and documented reasoning chains, enabling informed probability estimation even for novel risk pathways where direct measurement data may be unavailable.

This ability to structure analysis under uncertainty is especially valuable for high-impact, low-probability events like global catastrophic risks (GCRs). For GCRs, the framework's value lies less in achieving precise probability estimates and more in providing a structured process for analyzing potential pathways and impacts. GCRs are high-consequence events where, applying the principles of expected value, even probability estimates spanning multiple orders of magnitude below 1% can yield substantial expected harms. In such cases, it is the magnitude of potential impact and its reasoned non-trivial plausibility – not the precision of the estimate – that determines the relevance for decision-making (Kunreuther, 2002). Uncertainty alone does not justify inaction; on the contrary, the absence of precise data should increase the emphasis on precaution in light of the stakes involved (Baum, 2019).

More broadly, a benefit of the PRA for AI framework is its ability to take risk assessment forward in previously intractable areas by providing a structure for assessors and decision-makers confronting inherently qualitative aspects of threat modeling and likelihood estimation. It aids threat modeling and sizing by highlighting concepts such as unfortunate decisions, harm size cascades, and increasing maximum harm sizes over time as leverage grows. The process involves first structuring potential risk pathways, often using qualitative reasoning for novel or complex steps, which then provides the necessary foundation for systematic estimation (see Section 4.3). This structured approach helps counteract the analytical paralysis and additional risks that can occur when waiting for empirical evidence before acting (Casper et al., 2025), particularly as setting evidentiary standards too high for regulatory and assessment actions can lead to the neglect of significant risks posed by AI systems.

Finally, PRA for AI findings can contribute to institutional memory by integrating into enterprise-wide knowledge management systems, documenting identified risks. This documentation provides a basis for developing risk dashboards that track evolving threats across product lines and deployment contexts.

Stakeholder benefits. Different participants in the AI ecosystem can derive specific utility from the PRA for AI framework based on their specific roles and responsibilities.

AI developers can use PRA for AI to systematically evaluate how specific combinations of capabilities and design choices influence system-level risk profiles, facilitating early identification of thresholds where qualitative changes in risk emerge. This supports more informed architecture decisions, training strategies, and development trajectories grounded in safety considerations. Enterprise

adopters, in contrast, can assess whether a system aligns with their operational risk tolerance, guiding context-sensitive evaluations of readiness and informing pre-deployment investment decisions.

Risk professionals can integrate quantitative estimates into enterprise risk management systems, enabling more informed decisions about deployment and mitigation strategies while identifying and prioritizing potential harms. Compliance and legal professionals can use the framework to operationalize emerging regulatory expectations by embedding them in structured, repeatable assessment processes. PRA for AI helps clarify the basis for risk-related decisions by requiring well-documented assumptions, scenario justifications, and propagation pathways. This traceability supports internal governance, facilitates regulatory interpretation, and provides a defensible foundation for external audits, reviews, and liability considerations.

Evaluation organizations can harmonize insights from benchmarks, manual testing, and automated evaluations to produce risk level estimates rather than binary pass/fail assessments. This supports more nuanced understanding of system safety properties and potential failure modes, while advancing the scientific understanding of AI risks. Insurers and actuaries can use the framework's quantified risk estimates to assess potential liabilities and exposures across deployment contexts, supporting the development of fit-for-purpose insurance products and premium calculations (Weil et al., 2024). Regulatory and advisory entities can move beyond proxies such as training data size or capability-based risk restrictions by mandating standardized risk assessments that enable meaningful safety comparisons across AI systems. This provides a common vocabulary for evaluating compliance and the adequacy of safety measures.

Policymakers can use PRA for AI to evaluate trade-offs between advancing AI capabilities and implementing safety measures by quantifying expected value loss calculations [5]. This enables more proportionate interventions and helps align regulatory decisions with the scale and structure of system-level risks.

The framework's common vocabulary and structured documentation can help bridge fundamental gaps between technical experts, business leaders, and governance entities, addressing communication challenges that often hinder effective risk management. This shared analytical foundation can support the creation of feedback loops for continuously improving assessment practices. As PRA for AI is applied across diverse AI systems, assessment experiences can be systematically captured to refine aspect-specific methodologies, expand references, and strengthen framework efficacy and implementation over time.

Ultimately, the realization of these practical applications and stakeholder benefits are contingent on stakeholder confidence in the assessment's overall integrity and soundness. Downstream users – whether developers, adopters, or regulators – must perceive the assessment process and outputs as sufficiently credible and trustworthy before relying upon them for critical decisions.

5.3 Framework Limitations and Implementation Constraints

When utilizing the PRA for AI framework, several constraints and limitations must be considered and planned for. Assessors face appreciable challenges in probability assessment, including fundamental cognitive constraints in discriminating between different low probabilities, limited research foundations for various classes of low-probability forecasting, validation challenges for rare event predictions, and scarcity of qualified expertise availability. Additionally, many important estimates will not be able to have the benefit of historical precedent for previously unmanifested AI risks, making conventional predictive analytics or validation difficult. Assessors may also experience anchoring bias on predefined categories or examples, potentially limiting their consideration of unconventional or hard-to-imagine risks.

A core challenge lies in the inherent subjectivity when estimating likelihoods and impacts for novel risks. This subjectivity becomes more pronounced when determining how technical capabilities could propagate through interconnected societal systems to create harm – a process requiring understanding and judgment calls about complex societal dynamics and amplification effects. Assessor competence, honesty, and independence of incentives is crucial, as contingent or misaligned incentive structures could systematically and significantly degrade the quality and trustworthiness of these determinations.

[5] Expected value loss calculations enable the quantification of the 'price,' defined here as the resulting trade-off—expressed in expected value terms— balance between advancing capabilities and implementing safety measures.

It is important to note that PRA for AI represents a complementary set of incremental improvements to practice rather than a complete solution to AI risk assessment challenges. Its effectiveness depends heavily on expert judgment and inherits associated biases and limitations. Similar to all current methods, it cannot guarantee safety or completely address unknown unknowns. Instead, it provides a structured framework for making explicit an extended reasoning about risks and documenting of assumptions, while arriving at an assessor-defensible estimate of the total quantified risk in a very complex system.

Conducting an assessment with the workbook tool requires substantial technical and organizational resources due to the detailed analysis and systematic broad scope needed for assessing complex AI risks. This resource intensity can be prohibitive for time-sensitive decisions or organizations with restricted capacity. Effective implementation requires specialized expertise across multiple domains and maintaining consistency both within and across assessment teams. Assessment complexity varies fundamentally between aspects – from cases where structured empirical data is readily available to scenarios requiring extensive novel threat modeling. Furthermore, assessment quality can be significantly constrained by restricted access to proprietary system details (e.g., architecture, training methodologies, test data) or by limitations on performing probing analyses like whitebox testing or fine-tuning experiments, particularly for closed-source commercial systems.

5.4 Future Research Directions

The framework provides an initial foundation for systematic AI risk assessment, and several promising research directions could further extend its capabilities:

Improving uncertainty quantification. Current risk assessment remains contingent on subjective assessor judgment, making it difficult for decision-makers to evaluate the reliability of risk predictions. We need better techniques for calibrating expert judgment in these domains, particularly methods that can validate predictions against emerging empirical evidence. This should include developing structured protocols for uncertainty propagation that consider the effects of potential capability jumps and emergent behaviors, as well as research into the practicality and utility of methods for representing assessor uncertainty beyond point estimates (such as formal error bars).. Most crucially, we need methods for identifying and modeling correlated failure modes across multiple AI systems, as current approaches often focus on single-system analysis, overlooking risks emerging from the complex interactions and collective behaviors within multi-agent AI ecosystems (Hammond et al., 2025).

Network modeling of risk pathways. Current risk assessment methodologies often analyze harm pathways individually, limiting our ability to model the numerous potential routes through which risks can emerge, combine, and amplify. Network modeling approaches offer potential enhancements for risk pathway analysis. For example, threat knowledge graphs have demonstrated utility in uncovering previously unknown connections between specific cybersecurity entities, such as attack techniques, vulnerabilities, weaknesses, and affected software/hardware systems (Xiang et al., 2025). Building on complex systems approaches to AI risk, applying network and hypergraph models offers the potential for systematic exploration of multiple causal chains simultaneously (Kilian et al., 2023). Hypergraph approaches, in particular, are proposed for their ability to model the higher-order dependencies inherent in AI's sociotechnical ecosystem (Kilian et al., forthcoming), potentially revealing complex relationships and multi-hazard interactions missed by standard graph analysis. Such graph structures could identify direct, high-impact vulnerabilities through critical node analysis explicitly represent feedback loops that may accelerate risk development, while also revealing how initially minor risks can cascade and amplify through network propagation, leading to systemic consequences. Furthermore, extending our framework with computational techniques – such as Monte Carlo methods and scenario discovery simulations using generative AI – could leverage these network models to analyze compound effects and interaction scenarios more systematically, thereby identifying critical pathways that might be missed when risks are assessed in isolation. Additionally, network visualizations can serve as effective communication tools, making complex risk landscapes more analytically tractable to both researchers and decision-makers.

Establishing an AI risk pathway knowledge base. Many relevant parties' limited understanding of how capabilities scale and interrelate makes interpreting current evaluation results and predicting future risks challenging, as the science of capability evaluations remains underdeveloped (Weidinger et al., 2025). A structured knowledge base that systematically documents AI hazards and risk

pathways using our aspect-oriented taxonomy would provide valuable reference. The knowledge base could store detailed information about complete and partial risk pathways, drawing from incident databases, risk registers, and assessment outputs (HM Government, 2023; McGregor, 2020; Slattery et al., 2024). By cataloging detailed descriptions of potential risk pathways, documented interaction effects, the enabling capabilities, reference examples, and context-specific assessment methods, this resource could support more consistent risk evaluation across different operational contexts.

Conducting larger holistic case studies. In-depth public reference case studies spanning multiple AI aspects and diverse risk pathways would demonstrate how the framework's components interact in real assessment contexts. The framework and workbook tool should be used in assessments linked from model cards to assess the latest frontier models from leading AI developers (Anthropic, 2024b; Google, 2025; OpenAI, 2024) publicly with necessary but minimal redactions of novel information hazards (Bostrom, 2011). Such reference assessments would provide concrete illustrations of the techniques fully applied and worked out. We plan to publish illustrative case studies of the framework's application across different types and scopes of AI systems to show how organizations can operationalize the assessment process in various contexts.

Developing standards. Effective AI standards can benefit from normalized risk thresholds to move beyond purely process-focused requirements; the quantification capabilities provided by PRA for AI offer a foundation for establishing such common standards and being able to hold any system, past, present, or future, up to those thresholds. The framework could contribute to standards in several critical ways: hazard identification taxonomies, quantitative risk estimation protocols, and criteria for validating the quality of assessment execution. Future standards should adapt lessons from assurance level concepts from other safety-critical domains (as discussed in Section 2.4) for AI's unique challenges (e.g., emergent behaviors, rapid evolution). Such standards should establish clear, prescriptive requirements while accommodating new assessment techniques and tooling, ensuring continued relevance and consistency.

Refining risk mapping methodologies. Effective risk governance requires distilling complex technical assessments into actionable insights for diverse stakeholders. The current Focused Aggregation method (see Section 3.2) relies heavily on assessor judgment for mapping scenarios to societal risk dimensions. Future work should focus on developing more objective and generalizable classification criteria for these mappings. Furthermore, integrating outputs from advanced analytical techniques like network modeling of risk pathways (discussed earlier), which can reveal intricate interactions and dependencies, may generate insights whose complexity necessitates semi-automated support. Such tools, guided by well-defined criteria, could help consistently map complex scenario characteristics onto societal risk dimensions, reducing subjective variation and helping organizations better understand and act on assessment results while maintaining the systematic connection between technical capabilities and societal impacts.

6 Conclusion

This paper presents a probabilistic risk assessment framework for AI that applies methodologically grounded, systematic analysis to a field previously characterized by fragmented approaches, selective testing, and implicit assumptions about risk priorities. Building on established methods from high-reliability industries, we have shown how probabilistic techniques can be meaningfully adapted to evaluate AI systems through a structured, aspect-oriented approach. The framework helps transform previously qualitative and disjointed thinking about AI safety issues into empirically informed reasoned analyses extended through structured argumentation and diagrammatics, which can be methodically compared and evaluated.

Our approach advances AI risk assessment practice through three key advancements: (1) systematic exploration of the hazard space using aspect-oriented analysis, (2) structured quantification methodology that decomposes complex scenarios into analyzable components, and (3) explicit documentation protocols that capture reasoning chains and assumptions. These advances create the foundation for cumulative knowledge building in AI risk assessment, enabling organizations to build upon lessons from previous assessments, identify knowledge gaps, and track evolving risk profiles as capabilities advance.

The PRA for AI framework addresses current risk assessment fragmentation by aiming for broad coverage across diverse risk domains, enabling detailed exploration of causal risk pathways across

technical and societal areas , and providing concrete implementation guidance through the workbook tool (see Section 4) and also by the harmonization abilities of the framework. This integration makes it possible to assess risk more holistically in its breadth and depth. Our approach enables configurable assessment depths through tiered maturity levels for different organizational needs without sacrificing consideration of the societal threat surface.

Importantly, the framework establishes a common vocabulary for discussing AI risks across different stakeholder groups – from technical developers and safety researchers to policymakers and regulatory bodies – through standardized assessment scales, quantified risk levels, and structured documentation protocols. This shared foundation facilitates more productive audits, disagreements, and identification of assessment blind spots.

While implementing PRA for AI faces challenges, including uncertainties about novel risks, non-trivial work in quantifying tail risks and black swan events, and some limits on understanding of how advanced AI capabilities might emerge and interact, these limitations suggest the framework is most valuable when combined with complementary approaches like safety cases, qualitative scenario planning, and red teaming. We present this framework and the associated workbook tool, available at our project website, as a contribution to the nascent field of general-purpose AI risk assessment and welcome engagement from the broader research community to extend, refine, and demonstrate these approaches. This framework represents an initial step towards more comprehensive probabilistic risk assessment of advanced AI systems.

Given the rapid advancement of AI capabilities, we cannot afford to approach risk assessment in an ad hoc manner without methodological reasoning about a wide breadth of risk pathways. Within this systematic reasoning, particular focus should be given to identifying and analyzing the most highly consequential pathways. While adapting proven methods from high-reliability industries will not guarantee safety, it provides a crucial foundation for systematic risk assessment that the field urgently needs. Our hope is that this work contributes to increased operational and theoretical understanding of AI risk, enables better risk governance, enriches technical risk and remediation research, and supports policy work on developing measurable safety thresholds and safety requirements.

Acknowledgments

We extend our sincere gratitude to the many groups and individuals that provided invaluable feedback on the development and refinement of our Probabilistic Risk Assessment for AI framework, including: Anthony Aguirre, Usman Anwar, Peter Barnett, Claire Boine, Mark Brakel, Justin Bullock, Siméon Campos, Duncan Cass-Beggs, Giulio Corsi, Abra Ganz, Ross Gruetzemacher, Olle Häggström, Seán Ó hÉigeartaigh, Geoffrey Irving, Corin Katzke, Kyle A. Kilian, Daniel Kroth, Dan Lahav, Yolanda Lannquist, Matthijs Maas, Nada Madkour, Alexandru Marcoci, Evan R. Murphy, Malcolm Murray, Daniele Palombi, Jai Patel, Toby Pilditch, Krystal Rain, Deepika Raman, Matthias Samwald, Tim Schreier, Everett Smith, Cozmin Ududec, Risto Uuk, Akash Wasil, Sterlin Waters, and Marta Ziosi. All remaining errors are our own. To provide feedback or contribute to the framework's development, please contact the project team at PRA@carma.org.

References

Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A. J., Bambrick, J., Bodenstein, S. W., Evans, D. A., Hung, C.-C., O'Neill, M., Reiman, D., Tunyasuvunakool, K., Wu, Z., Žemgulytė, A., Arvaniti, E., … Jumper, J. M. (2024). Accurate structure prediction of biomolecular interactions with AlphaFold 3 [Publisher: Nature Publishing Group]. *Nature*, *630*(8016), 493–500. https://doi.org/10.1038/s41586-024-07487-w

Aguirre, A. (2024, September). Close the Gates: How we can keep the future human by choosing not to develop superhuman general-purpose artificial intelligence [arXiv:2311.09452 [cs]]. https://doi.org/10.48550/arXiv.2311.09452

AISI. (2024a, February). AI Safety Institute approach to evaluations. Retrieved February 21, 2025, from https://www.gov.uk/government/publications/ai-safety-institute-approach-to-evaluations/ai-safety-institute-approach-to-evaluations

AISI. (2024b, May). Advanced AI evaluations at AISI: May update | AISI Work. Retrieved January 13, 2025, from https://www.aisi.gov.uk/work/advanced-ai-evaluations-may-update

AISI. (2024c, November). Safety case template for 'inability' arguments | AISI Work. Retrieved April 16, 2025, from https://www.aisi.gov.uk/work/safety-case-template-for-inability-arguments

Anderljung, M., Barnhart, J., Korinek, A., Leung, J., O'Keefe, C., Whittlestone, J., Avin, S., Brundage, M., Bullock, J., Cass-Beggs, D., Chang, B., Collins, T., Fist, T., Hadfield, G., Hayes, A., Ho, L., Hooker, S., Horvitz, E., Kolt, N., … Wolf, K. (2023, November). Frontier AI Regulation: Managing Emerging Risks to Public Safety [arXiv:2307.03718]. https://doi.org/10.48550/arXiv.2307.03718

Anderson, N. G., & Piccinini, G. (2024, June). The Robust Mapping Account of Implementation [_eprint: https://academic.oup.com/book/0/chapter/447877215/chapter-pdf/59305358/workid-ukmomh10pg5t-book-part-6.pdf]. In *The Physical Signature of Computation: A Robust Mapping Account*. Oxford University Press. https://doi.org/10.1093/9780191872075.003.0006

Anthropic. (2023a, March). Core Views on AI Safety: When, Why, What, and How. Retrieved December 2, 2024, from https://www.anthropic.com/news/core-views-on-ai-safety

Anthropic. (2023b, September). Anthropic's Responsible Scaling Policy \ Anthropic. Retrieved December 4, 2024, from https://www.anthropic.com/news/anthropics-responsible-scaling-policy

Anthropic. (2024a, June). Challenges in Red Teaming AI Systems. Retrieved January 23, 2025, from https://www.anthropic.com/news/challenges-in-red-teaming-ai-systems

Anthropic. (2024b, October). Introducing computer use, a new Claude 3.5 Sonnet, and Claude 3.5 Haiku. Retrieved December 29, 2024, from https://www.anthropic.com/news/3-5-models-and-computer-use

Anthropic. (2025, January). Anthropic achieves ISO 42001 certification for responsible AI. Retrieved February 21, 2025, from https://www.anthropic.com/news/anthropic-achieves-iso-42001-certification-for-responsible-ai

Apostolakis, G. (1990). The Concept of Probability in Safety Assessments of Technological Systems [Publisher: American Association for the Advancement of Science]. *Science*, *250*(4986), 1359–1364. https://doi.org/10.1126/science.2255906

Aras, E. M., & Diaconeasa, M. A. (2021). A Critical Look at the Need for Performing Multi-Hazard Probabilistic Risk Assessment for Nuclear Power Plants [Number: 4 Publisher: Multidisciplinary Digital Publishing Institute]. *Eng*, *2*(4), 454–467. https://doi.org/10.3390/eng2040028

Barnett, P., & Thiergart, L. (2024a, November). Declare and Justify: Explicit assumptions in AI evaluations are necessary for effective regulation [arXiv:2411.12820]. https://doi.org/10.48550/arXiv.2411.12820

Barnett, P., & Thiergart, L. (2024b). What AI evaluations for preventing catastrophic risks can and cannot do.

Barrett, A., Newman, J., Nonnecke, B., Madkour, N., Hendrycks, D., Murphy, E. R., Jackson, K., & Raman, D. (2025, January). AI Risk-Management Standards Profile for General-Purpose AI (GPAI) and Foundation Models. Retrieved April 8, 2025, from https://cltc.berkeley.edu/publication/ai-risk-management-standards-profile-v1-1/

Basnett, Y., Henley, G., Howell, J., Jones, H., Lemma, A., & Pandey, P. R. (2014). *Coordination failure* (tech. rep.). ODI. Retrieved March 24, 2025, from https://www.jstor.org/stable/resrep50415.8

Baum, S. D. (2019). The challenge of analyzing global catastrophic risks. *Decision Analysis Today*, *38*, 20–24.

Bengio, Y., Hinton, G., Yao, A., Song, D., Abbeel, P., Darrell, T., Harari, Y. N., Zhang, Y.-Q., Xue, L., Shalev-Shwartz, S., Hadfield, G., Clune, J., Maharaj, T., Hutter, F., Baydin, A. G., McIlraith, S., Gao, Q., Acharya, A., Krueger, D., … Mindermann, S. (2024). Managing extreme AI risks amid rapid progress [Publisher: American Association for the Advancement of Science]. *Science*, *384*(6698), 842–845. https://doi.org/10.1126/science.adn0117

Bengio, Y., Mindermann, S., Privitera, D., Besiroglu, T., Bommasani, R., Casper, S., Choi, Y., Fox, P., Garfinkel, B., Goldfarb, D., Heidari, H., Ho, A., Kapoor, S., Khalatbari, L., Longpre, S., Manning, S., Mavroudis, V., Mazeika, M., Michael, J., … Zeng, Y. (2025). *International AI Safety Report* (tech. rep. No. DSIT 2025/001). https://www.gov.uk/government/publications/international-ai-safety-report-2025

Bereska, L., & Gavves, E. (2024, August). Mechanistic Interpretability for AI Safety – A Review [arXiv:2404.14082 [cs]]. https://doi.org/10.48550/arXiv.2404.14082

Bergman, B. (1992). The development of reliability techniques: A retrospective survey. *Reliability Engineering & System Safety*, *36*(1), 3–6. https://doi.org/10.1016/0951-8320(92)90143-9

Bhatt, M., Chennabasappa, S., Nikolaidis, C., Wan, S., Evtimov, I., Gabi, D., Song, D., Ahmad, F., Aschermann, C., Fontana, L., Frolov, S., Giri, R. P., Kapil, D., Kozyrakis, Y., LeBlanc, D., Milazzo, J., Straumann, A., Synnaeve, G., Vontimitta, V., … Saxe, J. (2023, December). Purple Llama CyberSecEval: A Secure Coding Benchmark for Language Models [arXiv:2312.04724 [cs]]. https://doi.org/10.48550/arXiv.2312.04724

Bommasani, R., Sanjeev Arora, Yejin Choi, Li Fei-Fei, Daniel E. Ho, Dan Jurafsky, Sanmi Koyejo, Hima Lakkaraju, Arvind Narayanan, Alondra Nelson, Emma Pierson, Joelle Pineau, Gaël Varoquaux, Suresh Venkatasubramanian, Ion Stoica, Percy Liang, & Dawn Song. (2024). A Path for Science- and Evidence-based AI Policy. Retrieved December 13, 2024, from https://understanding-ai-safety.org/

Bostrom, N. (2011). Information hazards: A typology of potential harms from knowledge. *Review of Contemporary Philosophy*, *10*, 44–79.

Buçinca, Z., Lin, P., Gajos, K. Z., & Glassman, E. L. (2020, January). Proxy Tasks and Subjective Measures Can Be Misleading in Evaluating Explainable AI Systems [arXiv:2001.08298]. https://doi.org/10.48550/arXiv.2001.08298

Buhl, M., Hilton, B., Masterson, T., & Irving, G. (2025, February). How can safety cases be used to help with frontier AI safety? | AISI Work. Retrieved February 21, 2025, from https://www.aisi.gov.uk/work/how-can-safety-cases-be-used-to-help-with-frontier-ai-safety

Burden, J. (2024, July). Evaluating AI Evaluation: Perils and Prospects [arXiv:2407.09221 [cs] version: 1]. https://doi.org/10.48550/arXiv.2407.09221

CAI. (2024, November). METHODOLOGY FOR THE RISK AND IMPACT ASSESSMENT OF ARTIFICIAL INTELLIGENCE SYSTEMS FROM THE POINT OF VIEW OF HUMAN RIGHTS, DEMOCRACY AND THE RULE OF LAW(HUDERIA METHODOLOGY). Retrieved April 9, 2025, from https://www.coe.int/en/web/portal/-/huderia-new-tool-to-assess-the-impact-of-ai-systems-on-human-rights

Campos, S., Papadatos, H., Roger, F., Touzet, C., Quarks, O., & Murray, M. (2025, February). A Frontier AI Risk Management Framework: Bridging the Gap Between Current AI Practices and Established Risk Management [arXiv:2502.06656 [cs]]. https://doi.org/10.48550/arXiv.2502.06656

Cârlan, C., Gomez, F., Mathew, Y., Krishna, K., King, R., Gebauer, P., & Smith, B. R. (2024, December). Dynamic safety cases for frontier AI [arXiv:2412.17618 [cs]]. https://doi.org/10.48550/arXiv.2412.17618

Casper, S., Krueger, D., & Hadfield-Menell, D. (2025, February). Pitfalls of Evidence-Based AI Policy [arXiv:2502.09618 [cs] version: 1]. https://doi.org/10.48550/arXiv.2502.09618

Chan, A. (2024, April). Evaluating Predictions of Model Behaviour | GovAI. Retrieved March 31, 2025, from https://www.governance.ai/analysis/evaluating-predictions-of-model-behaviour

Che, Z., Casper, S., Kirk, R., Satheesh, A., Slocum, S., McKinney, L. E., Gandikota, R., Ewart, A., Rosati, D., Wu, Z., Cai, Z., Chughtai, B., Gal, Y., Huang, F., & Hadfield-Menell, D. (2025, February). Model Tampering Attacks Enable More Rigorous Evaluations of LLM Capabilities [arXiv:2502.05209 [cs]]. https://doi.org/10.48550/arXiv.2502.05209

Clymer, J., Gabrieli, N., Krueger, D., & Larsen, T. (2024, March). Safety Cases: How to Justify the Safety of Advanced AI Systems [arXiv:2403.10462]. https://doi.org/10.48550/arXiv.2403.10462

COE. (2024, September). The Framework Convention on Artificial Intelligence - Artificial Intelligence - www.coe.int. Retrieved January 17, 2025, from https://www.coe.int/en/web/artificial-intelligence/the-framework-convention-on-artificial-intelligence

Coleman, J. L., Bolisetti, C., Veeraraghavan, S., Parisi, C., Prescott, S. R., & Gupta, A. (2016, September). *Multi-Hazard Advanced Seismic Probabilistic Risk Assessment Tools and Applications* (tech. rep. No. INL/EXT-16-40055). Idaho National Lab. (INL), Idaho Falls, ID (United States). https://doi.org/10.2172/1369534

Cottier, B. (2024, November). How Far Behind Are Open Models? Retrieved December 2, 2024, from https://epoch.ai/blog/open-models-report

Critch, A., & Krueger, D. (2020, May). AI Research Considerations for Human Existential Safety (ARCHES) [arXiv:2006.04948 [cs]]. https://doi.org/10.48550/arXiv.2006.04948

Critch, A., & Russell, S. (2023, June). TASRA: A Taxonomy and Analysis of Societal-Scale Risks from AI [arXiv:2306.06924 [cs]]. https://doi.org/10.48550/arXiv.2306.06924

Dalrymple, D. "., Skalse, J., Bengio, Y., Russell, S., Tegmark, M., Seshia, S., Omohundro, S., Szegedy, C., Goldhaber, B., Ammann, N., Abate, A., Halpern, J., Barrett, C., Zhao, D., Zhi-Xuan, T., Wing, J., & Tenenbaum, J. (2024, July). Towards Guaranteed Safe AI: A Framework for Ensuring Robust and Reliable AI Systems [arXiv:2405.06624]. https://doi.org/10.48550/arXiv.2405.06624

Dell'Acqua, F., McFowland III, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Krayer, L., Candelon, F., & Lakhani, K. R. (2023, September). Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality. https://doi.org/10.2139/ssrn.4573321

Dennett, D. C. (2013). *Intuition Pumps and Other Tools for Thinking*. W. W. Norton & Company.

DHS. (2025, January). DHS Generative AI Public Sector Playbook | Homeland Security. Retrieved March 17, 2025, from https://www.dhs.gov/publication/dhs-generative-ai-public-sector-playbook

Dominguez-Olmedo, R., Dorner, F. E., & Hardt, M. (2024, July). Training on the Test Task Confounds Evaluation and Emergence [arXiv:2407.07890 [cs] version: 1]. https://doi.org/10.48550/arXiv.2407.07890

Dragan, A., King, H., & Dafoe, A. (2024, December). Introducing the Frontier Safety Framework. Retrieved January 12, 2025, from https://deepmind.google/discover/blog/introducing-the-frontier-safety-framework/

Duffer, S., Singh, A., & Hallinan, P. (2024, November). AWS achieves ISO/IEC 42001:2023 Artificial Intelligence Management System accredited certification | AWS Machine Learning Blog [Section: Announcements]. Retrieved February 21, 2025, from https://aws.amazon.com/blogs/machine-learning/aws-achieves-iso-iec-420012023-artificial-intelligence-management-system-accredited-certification/

Eisenberg, I. W., Gamboa, L., & Sherman, E. (2025, March). The Unified Control Framework: Establishing a Common Foundation for Enterprise AI Governance, Risk Management and Regulatory Compliance [arXiv:2503.05937 [cs]]. https://doi.org/10.48550/arXiv.2503.05937

Ericson II, C. A. (2005). System Safety [Section: 1 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/0471739421.ch1]. In *Hazard Analysis Techniques for System Safety* (pp. 1–12). John Wiley & Sons, Ltd. https://doi.org/10.1002/0471739421.ch1

EU. (2023, June). EU AI Act: First regulation on artificial intelligence. Retrieved March 7, 2025, from https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence

EU. (2024, December). Second Draft of the General-Purpose AI Code of Practice published, written by independent experts | Shaping Europe's digital future. Retrieved December 30, 2024, from https://digital-strategy.ec.europa.eu/en/library/second-draft-general-purpose-ai-code-practice-published-written-independent-experts

EU. (2025, March). Third Draft of the General-Purpose AI Code of Practice published, written by independent experts | Shaping Europe's digital future. Retrieved March 21, 2025, from https://digital-strategy.ec.europa.eu/en/library/third-draft-general-purpose-ai-code-practice-published-written-independent-experts

Fang, R., Bindu, R., Gupta, A., Zhan, Q., & Kang, D. (2024, June). Teams of LLM Agents can Exploit Zero-Day Vulnerabilities [arXiv:2406.01637 version: 1]. https://doi.org/10.48550/arXiv.2406.01637

Feffer, M., Sinha, A., Deng, W. H., Lipton, Z. C., & Heidari, H. (2024, August). Red-Teaming for Generative AI: Silver Bullet or Security Theater? [arXiv:2401.15897 [cs]]. https://doi.org/10.48550/arXiv.2401.15897

Financial Stability Board. (2024, November). The Financial Stability Implications of Artificial Intelligence. Retrieved April 24, 2025, from https://www.fsb.org/2024/11/the-financial-stability-implications-of-artificial-intelligence/

For Humanity. (2016). Independent Audit of AI Systems (IAAIS). Retrieved December 29, 2024, from https://forhumanity.center/independent-audit-of-ai-systems/

Friedler, S., Singh, R., Blili-Hamelin, B., Metcalf, J., & Chen, B. J. (2023, October). AI Red-Teaming Is Not a One-Stop Solution to AI Harms: Recommendations for Using Red-Teaming for AI Accountability. Retrieved January 28, 2025, from https://datasociety.net/library/ai-red-teaming-is-not-a-one-stop-solution-to-ai-harms-recommendations-for-using-red-teaming-for-ai-accountability/?utm_source=chatgpt.com

Future of Life Institute. (2025, February). Safety Standards Delivering Controllable and Beneficial AI Tools. Retrieved February 24, 2025, from https://futureoflife.org/document/safety-standards-delivering-controllable-and-beneficial-ai-tools/

Gahin, F. S., & Williams, C. A. (1972). Review of the Literature on Risk Management (R. M. Heins, Ed.) [Publisher: [American Risk and Insurance Association, Wiley]]. *The Journal of Risk and Insurance*, *39*(3), 463–470. https://doi.org/10.2307/251839

Ganguli, D., Lovitt, L., Kernion, J., Askell, A., Bai, Y., Kadavath, S., Mann, B., Perez, E., Schiefer, N., Ndousse, K., Jones, A., Bowman, S., Chen, A., Conerly, T., DasSarma, N., Drain, D., Elhage, N., El-Showk, S., Fort, S., … Clark, J. (2022, November). Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned [arXiv:2209.07858 [cs]]. https://doi.org/10.48550/arXiv.2209.07858

Gemini, T., Anil, R., Borgeaud, S., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., Millican, K., Silver, D., Johnson, M., Antonoglou, I., Schrittwieser, J., Glaese, A., Chen, J., Pitler, E., Lillicrap, T., Lazaridou, A., … Vinyals, O. (2024, June). Gemini: A Family of Highly Capable Multimodal Models [arXiv:2312.11805]. https://doi.org/10.48550/arXiv.2312.11805

Givens, A. R. (2023, November). CDT, Civil Society Reps to UK AI Safety Summit Urge Focus on AI Risks to People's Rights. Retrieved March 17, 2025, from https://cdt.org/insights/cdt-civil-society-reps-to-uk-ai-safety-summit-urge-focus-on-ai-risks-to-peoples-rights/

Goemans, A., Buhl, M. D., Schuett, J., Korbak, T., Wang, J., Hilton, B., & Irving, G. (2024, November). Safety case template for frontier AI: A cyber inability argument. Retrieved February 24, 2025, from https://arxiv.org/abs/2411.08088v1

Google. (2024, December). Introducing Gemini 2.0: Our new AI model for the agentic era. Retrieved December 29, 2024, from https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/

Google. (2025, April). Google Model Cards. Retrieved April 22, 2025, from https://modelcards.withgoogle.com/about

Habli, I., Hawkins, R., Paterson, C., Ryan, P., Jia, Y., Sujan, M., & McDermid, J. (2025, March). The BIG Argument for AI Safety Cases [arXiv:2503.11705 [cs]]. https://doi.org/10.48550/arXiv.2503.11705

Hammond, L., Chan, A., Clifton, J., Hoelscher-Obermaier, J., Khan, A., McLean, E., Smith, C., Barfuss, W., Foerster, J., Gavenčiak, T., Han, T. A., Hughes, E., Kovařík, V., Kulveit, J., Leibo, J. Z., Oesterheld, C., Witt, C. S. d., Shah, N., Wellman, M., … Rahwan, I. (2025). *Multi-Agent Risks from Advanced AI* (tech. rep. No. 1). Cooperative AI Foundation.

Heim, L., & Koessler, L. (2024, August). Training Compute Thresholds: Features and Functions in AI Regulation [arXiv:2405.10799 [cs] version: 2]. https://doi.org/10.48550/arXiv.2405.10799

Hemming, V., Burgman, M. A., Hanea, A. M., McBride, M. F., & Wintle, B. C. (2018). A practical guide to structured expert elicitation using the IDEA protocol [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/2041-210X.12857]. *Methods in Ecology and Evolution*, *9*(1), 169–180. https://doi.org/10.1111/2041-210X.12857

Hendrycks, D., & Mazeika, M. (2022, September). X-Risk Analysis for AI Research [arXiv:2206.05862 [cs]]. https://doi.org/10.48550/arXiv.2206.05862

Hendrycks, D., Mazeika, M., & Woodside, T. (2023, October). An Overview of Catastrophic AI Risks [arXiv:2306.12001]. https://doi.org/10.48550/arXiv.2306.12001

Hintersdorf, D., Struppek, L., & Kersting, K. (2023, August). Balancing Transparency and Risk: The Security and Privacy Risks of Open-Source Machine Learning Models [arXiv:2308.09490]. https://doi.org/10.48550/arXiv.2308.09490

HM Government. (2023, August). National Risk Register 2023. Retrieved December 30, 2024, from https://www.gov.uk/government/publications/national-risk-register-2023

HM Government. (2024, May). Frontier AI Safety Commitments, AI Seoul Summit 2024. Retrieved January 28, 2025, from https://www.gov.uk/government/publications/frontier-ai-safety-commitments-ai-seoul-summit-2024/frontier-ai-safety-commitments-ai-seoul-summit-2024

Huang, J., Gu, S. S., Hou, L., Wu, Y., Wang, X., Yu, H., & Han, J. (2022, October). Large Language Models Can Self-Improve [arXiv:2210.11610 [cs]]. https://doi.org/10.48550/arXiv.2210.11610

IEC. (2019, June). IEC 31010:2019. Retrieved April 11, 2025, from https://www.iso.org/standard/72140.html

Inan, H., Upasani, K., Chi, J., Rungta, R., Iyer, K., Mao, Y., Tontchev, M., Hu, Q., Fuller, B., Testuggine, D., & Khabsa, M. (2023, December). Llama Guard: LLM-based Input-Output Safeguard for Human-AI Conversations [arXiv:2312.06674 [cs]]. https://doi.org/10.48550/arXiv.2312.06674

Irving, G. (2024, August). Safety cases at AISI | AISI Work. Retrieved February 21, 2025, from https://www.aisi.gov.uk/work/safety-cases-at-aisi

ISO. (2021, August). ISO/SAE 21434:2021. Retrieved January 22, 2025, from https://www.iso.org/standard/70918.html

ISO. (2023a, February). ISO/IEC 23894:2023(en), Information technology — Artificial intelligence — Guidance on risk management. Retrieved December 30, 2024, from https://www.iso.org/obp/ui/en/#iso:std:iso-iec:23894:ed-1:v1:en

ISO. (2023b, December). ISO/IEC 42001:2023. Retrieved January 17, 2025, from https://www.iso.org/standard/81230.html

Jain, N., Schwarzschild, A., Wen, Y., Somepalli, G., Kirchenbauer, J., Chiang, P.-y., Goldblum, M., Saha, A., Geiping, J., & Goldstein, T. (2023, September). Baseline Defenses for Adversarial Attacks Against Aligned Language Models [arXiv:2309.00614 [cs]]. https://doi.org/10.48550/arXiv.2309.00614

Järviniemi, O., & Hubinger, E. (2024). Uncovering Deceptive Tendencies in Language Models: A Simulated Company AI Assistant [_eprint: 2405.01576]. https://arxiv.org/abs/2405.01576

Ji, J. (2025, March). How to Improve AI Red-Teaming: Challenges and Recommendations. Retrieved March 24, 2025, from https://cset.georgetown.edu/article/how-to-improve-ai-red-teaming-challenges-and-recommendations/

Jones, E., Hardalupas, M., & Agnew, W. (2024, July). Under the radar? Examining the evaluation of foundation models. Retrieved January 28, 2025, from https://www.adalovelaceinstitute.org/report/under-the-radar/

Jones, E., Tong, M., Mu, J., Mahfoud, M., Leike, J., Grosse, R., Kaplan, J., Fithian, W., Perez, E., & Sharma, M. (2025, February). Forecasting Rare Language Model Behaviors [arXiv:2502.16797 [cs]]. https://doi.org/10.48550/arXiv.2502.16797

Kasirzadeh, A. (2024, October). Measurement challenges in AI catastrophic risk governance and safety frameworks [arXiv:2410.00608]. https://doi.org/10.48550/arXiv.2410.00608

Kasirzadeh, A. (2025, January). Two Types of AI Existential Risk: Decisive and Accumulative [arXiv:2401.07836 [cs] version: 3]. https://doi.org/10.48550/arXiv.2401.07836

Kilian, K. A., Ventura, C. J., & Bailey, M. M. (2023). Examining the differential risk from high-level artificial intelligence and the question of control. *Futures*, *151*, 103182. https://doi.org/10.1016/j.futures.2023.103182

Kilian, K. A., & Mallah, R. (forthcoming). *Multiscale Hypergraph Ontology for AI Risk Management*.

Kunreuther, H. (2002). Risk analysis and risk management in an uncertain world. *Risk Analysis: An Official Publication of the Society for Risk Analysis*, *22*(4), 655–664. https://doi.org/10.1111/0272-4332.00057

Laine, R., Meinke, A., & Evans, O. (2023). Towards a Situational Awareness Benchmark for LLMs. *Socially Responsible Language Modelling Research*. https://openreview.net/forum?id=DRk4bWKr41

Laine, R., Chughtai, B., Betley, J., Hariharan, K., Scheurer, J., Balesni, M., Hobbhahn, M., Meinke, A., & Evans, O. (2024, July). Me, Myself, and AI: The Situational Awareness Dataset (SAD) for LLMs [arXiv:2407.04694 [cs]]. https://doi.org/10.48550/arXiv.2407.04694

Lee, S., Kim, M., Cherif, L., Dobre, D., Lee, J., Hwang, S. J., Kawaguchi, K., Gidel, G., Bengio, Y., Malkin, N., & Jain, M. (2024, May). Learning diverse attacks on large language models for robust red-teaming and safety tuning [arXiv:2405.18540]. https://doi.org/10.48550/arXiv.2405.18540

Lester, R. R., Green, L. C., & Linkov, I. (2007). Site-specific applications of probabilistic health risk assessment: Review of the literature since 2000. *Risk Analysis: An Official Publication of the Society for Risk Analysis*, *27*(3), 635–658. https://doi.org/10.1111/j.1539-6924.2007.00890.x

Li, N., Pan, A., Gopal, A., Yue, S., Berrios, D., Gatti, A., Li, J. D., Dombrowski, A.-K., Goel, S., Phan, L., Mukobi, G., Helm-Burger, N., Lababidi, R., Justen, L., Liu, A. B., Chen, M., Barrass, I., Zhang, O., Zhu, X., … Hendrycks, D. (2024, May). The WMDP Benchmark: Measuring and Reducing Malicious Use With Unlearning [arXiv:2403.03218]. https://doi.org/10.48550/arXiv.2403.03218

Lin, M., Sheng, J., Zhao, A., Wang, S., Yue, Y., Wu, Y., Liu, H., Liu, J., Huang, G., & Liu, Y.-J. (2024, October). LLM-based Optimization of Compound AI Systems: A Survey [arXiv:2410.16392]. https://doi.org/10.48550/arXiv.2410.16392

Lindsey, J., Gurnee, W., Ameisen, E., Chen, B., Pearce, A., Turner, N. L., Citro, C., Abrahams, D., Carter, S., Hosmer, B., Marcus, J., Sklar, M., Templeton, A., Bricken, T., McDougall, C., Cunningham, H., Henighan, T., Jermyn, A., Jones, A., … Batson, J. (2025, March). On the Biology of a Large Language Model. Retrieved April 7, 2025, from https://transformer-circuits.pub/2025/attribution-graphs/biology.html

Liu, B., Mazumder, S., Robertson, E., & Grigsby, S. (2023, April). AI Autonomy : Self-Initiated Open-World Continual Learning and Adaptation [arXiv:2203.08994 [cs]]. https://doi.org/10.48550/arXiv.2203.08994

Maggio, G. (1996). Space Shuttle probabilistic risk assessment: Methodology and application [ISSN: 0149-144X]. *Proceedings of 1996 Annual Reliability and Maintainability Symposium*, 121–132. https://doi.org/10.1109/RAMS.1996.500652

Maidana, R. G., Parhizkar, T., Gomola, A., Utne, I. B., & Mosleh, A. (2023). Supervised dynamic probabilistic risk assessment: Review and comparison of methods. *Reliability Engineering & System Safety*, *230*, 108889. https://doi.org/10.1016/j.ress.2022.108889

Malinowski, S., Alexandra. (2019). Feedback Loops. In *Encyclopedia of Personality and Individual Differences* (pp. 1–3). Springer International Publishing.

Markmann, C., Darkow, I.-L., & von der Gracht, H. (2013). A Delphi-based risk analysis — Identifying and assessing future challenges for supply chain security in a multi-stakeholder environment. *Technological Forecasting and Social Change*, *80*(9), 1815–1833. https://doi.org/10.1016/j.techfore.2012.10.019

McGregor, S. (2020, November). Preventing Repeated Real World AI Failures by Cataloging Incidents: The AI Incident Database [arXiv:2011.08512 [cs]]. https://doi.org/10.48550/arXiv.2011.08512

Meinke, A., Schoen, B., Scheurer, J., Balesni, M., Shah, R., & Hobbhahn, M. (2025, January). Frontier Models are Capable of In-context Scheming [arXiv:2412.04984 [cs]]. https://doi.org/10.48550/arXiv.2412.04984

Microsoft. (2024, May). Retrace your steps with Recall. Retrieved February 17, 2025, from https://support.microsoft.com/en-us/windows/retrace-your-steps-with-recall-aa03f8a0-a78b-4b3e-b0a1-2eb8ac48701c

Miller, E. (2024, November). Adding Error Bars to Evals: A Statistical Approach to Language Model Evaluations. Retrieved April 23, 2025, from https://arxiv.org/abs/2411.00640v1

Ministry of Science and ICT. (2025, January). Basic Act on the Development of Artificial Intelligence and Creation of a Trust Base, etc. Retrieved March 24, 2025, from https://www.law.go.kr/%EB%B2%95%EB%A0%B9/%EC%9D%B8%EA%B3%B5%EC%A7%80%EB%8A%A5%20%EB%B0%9C%EC%A0%84%EA%B3%BC%20%EC%8B%A0%EB%A2%B0%20%EA%B8%B0%EB%B0%98%20%EC%A1%B0%EC%84%B1%20%EB%93%B1%EC%97%90%20%EA%B4%80%ED%95%9C%20%EA%B8%B0%EB%B3%B8%EB%B2%95/(20676,20250121)

MITRE. (2025, April). MITRE ATT&CK®. Retrieved April 24, 2025, from https://attack.mitre.org/

Modarres, M. (2008, September). Probabilistic Risk Assessment - Modarres - 2007 - Major Reference Works - Wiley Online Library. Retrieved December 3, 2024, from https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470061596.risk0700

Montagu, B., & Jensen, T. (2021). Trace-based control-flow analysis. *Proceedings of the 42nd ACM SIGPLAN International Conference on Programming Language Design and Implementation*, 482–496. https://doi.org/10.1145/3453483.3454057

Moustafa, M. A. M. M., & Chang, C.-k. (2021). Preventing cascading failure of electric power protection systems in nuclear power plant. *Nuclear Engineering and Technology*, *53*(1), 121–130. https://doi.org/10.1016/j.net.2020.06.010

Mukobi, G. (2024, August). Reasons to Doubt the Impact of AI Risk Evaluations [arXiv:2408.02565]. https://doi.org/10.48550/arXiv.2408.02565

Murray, M., Papadatos, H., Quarks, O., Gimenez, P.-F., & Campos, S. (2025, March). Mapping AI Benchmark Data to Quantitative Risk Estimates Through Expert Elicitation [arXiv:2503.04299 [cs]]. https://doi.org/10.48550/arXiv.2503.04299

NIST. (2022a). AI Risk Management Framework: Second Draft - August 18, 2022. *NIST*.

NIST. (2022b). AI Test, Evaluation, Validation and Verification (TEVV) [Last Modified: 2024-11-13T15:34-05:00]. *NIST*. Retrieved December 27, 2024, from https://www.nist.gov/ai-test-evaluation-validation-and-verification-tevv

NIST. (2025). U.S. AI Safety Institute Consortium Holds First Plenary Meeting to Reflect on Progress in 2024 & Outline Research Priorities for 2025 [Last Modified: 2025-02-04T12:32-05:00]. *NIST*. Retrieved March 17, 2025, from https://www.nist.gov/news-events/news/us-ai-safety-institute-consortium-holds-first-plenary-meeting-reflect-progress-2024

OpenAI. (2023, December). Preparedness Framework. Retrieved January 12, 2025, from https://openai.com/safety/

OpenAI. (2024, December). OpenAI o3 and o3-mini. Retrieved January 27, 2025, from https://openai.com/12-days/

OpenAI. (2025, January). Computer-Using Agent. Retrieved February 17, 2025, from https://openai.com/index/computer-using-agent/

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., … Zoph, B. (2024, March). GPT-4 Technical Report [arXiv:2303.08774]. https://doi.org/10.48550/arXiv.2303.08774

Phadke, A., & Thorp, J. (1996). Expose hidden failures to prevent cascading outages [in power systems] [Conference Name: IEEE Computer Applications in Power]. *IEEE Computer Applications in Power*, *9*(3), 20–23. https://doi.org/10.1109/67.526849

Pittaras, N., & McGregor, S. (2022, November). A taxonomic system for failure cause analysis of open source AI incidents [arXiv:2211.07280 [cs]]. https://doi.org/10.48550/arXiv.2211.07280

Raman, D., Madkour, N., Murphy, E. R., Jackson, K., & Newman, J. (2025, February). Intolerable Risk Threshold Recommendations for Artificial Intelligence. Retrieved February 24, 2025, from https://cltc.berkeley.edu/publication/intolerable-ai-risk-thresholds/

Rapita. (2012). DO-178C Guidance: Introduction to RTCA DO-178 certification | Rapita Systems. Retrieved January 21, 2025, from https://www.rapitasystems.com/do178

Rein, D., Hou, B. L., Stickland, A. C., Petty, J., Pang, R. Y., Dirani, J., Michael, J., & Bowman, S. R. (2023, November). GPQA: A Graduate-Level Google-Proof Q&A Benchmark [arXiv:2311.12022 [cs]]. https://doi.org/10.48550/arXiv.2311.12022

Ren, R., Basart, S., Khoja, A., Gatti, A., Phan, L., Yin, X., Mazeika, M., Pan, A., Mukobi, G., Kim, R. H., Fitz, S., & Hendrycks, D. (2024, July). Safetywashing: Do AI Safety Benchmarks Actually Measure Safety Progress? [arXiv:2407.21792]. https://doi.org/10.48550/arXiv.2407.21792

Rose, S., & Nelson, C. (2023, October). *Understanding AI-Facilitated Biological Weapon Development*. The Centre for Long-Term Resilience: London UK.

Rutinel, M., Titone, B., & Rodriguez, R. (2024, May). Consumer Protections for Artificial Intelligence.

Schiff, D. S., Ayesh, A., Musikanski, L., & Havens, J. C. (2020). IEEE 7010: A New Standard for Assessing the Well-being Implications of Artificial Intelligence [arXiv:2005.06620 [cs]]. *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2746–2753. https://doi.org/10.1109/SMC42975.2020.9283454

Schuett, J. (2024). Frontier AI developers need an internal audit function [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/risa.17665]. *Risk Analysis*, *n/a*(n/a). https://doi.org/10.1111/risa.17665

Seger, E., Dreksler, N., Moulange, R., Dardaman, E., Schuett, J., Wei, K., Winter, C., Arnold, M., hÉigeartaigh, S. Ó., Korinek, A., Anderljung, M., Bucknall, B., Chan, A., Stafford, E., Koessler, L., Ovadya, A., Garfinkel, B., Bluemke, E., Aird, M., … Gupta, A. (2023, September). Open-Sourcing Highly Capable Foundation Models: An evaluation of risks, benefits, and alternative methods for pursuing open-source objectives [arXiv:2311.09227]. https://doi.org/10.48550/arXiv.2311.09227

Sharkey, L., Ní Ghuidhir, C., Braun, D., Scheurer, J., Balesni, M., Bushnaq, L., Stix, C., & Hobbhahn, M. (2024, January). A Causal Framework for AI Regulation and Auditing. https://doi.org/10.20944/preprints202401.1424.v1

Shevlane, T., Farquhar, S., Garfinkel, B., Phuong, M., Whittlestone, J., Leung, J., Kokotajlo, D., Marchal, N., Anderljung, M., Kolt, N., Ho, L., Siddarth, D., Avin, S., Hawkins, W., Kim, B., Gabriel, I., Bolina, V., Clark, J., Bengio, Y., … Dafoe, A. (2023, September). Model evaluation for extreme risks [arXiv:2305.15324]. https://doi.org/10.48550/arXiv.2305.15324

Shortridge, J., Aven, T., & Guikema, S. (2017). Risk assessment under deep uncertainty: A methodological comparison. *Reliability Engineering & System Safety*, *159*, 12–23. https://doi.org/10.1016/j.ress.2016.10.017

Slattery, P., Saeri, A. K., Grundy, E. A. C., Graham, J., Noetel, M., Uuk, R., Dao, J., Pour, S., Casper, S., & Thompson, N. (2024, August). The AI Risk Repository: A Comprehensive Meta-Review, Database, and Taxonomy of Risks From Artificial Intelligence [arXiv:2408.12622 [cs]]. https://doi.org/10.48550/arXiv.2408.12622

Søgaard, A. (2023). On the Opacity of Deep Neural Networks. *Canadian Journal of Philosophy*, *53*(3), 224–239. https://doi.org/10.1017/can.2024.1

Stamatelatos, M. (2002). Probabilistic Risk Assessment: What Is It And Why Is It Worth Performing It? *NASA Safe and Mission Assurance News*.

Stamatelatos, M., Dezfuli, H., Apostolakis, G., Everline, C., Guarro, S., Mathias, D., Mosleh, A., Paulos, T., Riha, D., Smith, C., Vesely, W., & Youngblood, R. (2011, December). Probabilistic Risk Assessment Procedures Guide for NASA Managers and Practitioners (Second Edition) [NTRS Author Affiliations: NASA Headquarters, Nuclear Regulatory Commission, Jet Propulsion Lab., California Inst. of Tech., Aerospace Corp., NASA Ames Research Center, Maryland Univ., Alejo Engineering, Inc., Southwest Research Inst., Idaho National Lab. NTRS Report/Patent Number: HQ-STI-11-213 NTRS Document ID: 20120001369 NTRS Research Center: Headquarters (HQ)]. Retrieved December 2, 2024, from https://ntrs.nasa.gov/citations/20120001369

Suzgun, M., & Kalai, A. T. (2024, January). Meta-Prompting: Enhancing Language Models with Task-Agnostic Scaffolding [arXiv:2401.12954]. https://doi.org/10.48550/arXiv.2401.12954

Titus, J. (2024, March). Can Preparedness Frameworks Pull Their Weight? Retrieved January 28, 2025, from https://fas.org/publication/scaling-ai-safety/

Tudoran, C. (2018). Sciendo. *Proceedings of the International Conference on Business Excellence*, *12*(1), 983–991. https://doi.org/10.2478/picbe-2018-0088

UN. (2021, April). UN Regulation No. 155 - Cyber security and cyber security management system | UNECE. Retrieved January 21, 2025, from https://unece.org/transport/documents/2021/03/standards/un-regulation-no-155-cyber-security-and-cyber-security

US EPA, O. (2015, September). Risk Assessment Guidance for Superfund (RAGS) Volume III: Part A. Retrieved November 4, 2024, from https://www.epa.gov/risk/risk-assessment-guidance-superfund-rags-volume-iii-part

US Nuclear Regulatory Commission. (1990, December). Severe Accident Risks: An Assessment for Five U.S. Nuclear Power Plants (NUREG-1150). Retrieved November 4, 2024, from https://www.nrc.gov/reading-rm/doc-collections/nuregs/staff/sr1150/index.html

US Nuclear Regulatory Commission. (2024, January). Backgrounder on Probabilistic Risk Assessment. Retrieved February 24, 2025, from https://www.nrc.gov/reading-rm/doc-collections/fact-sheets/probabilistic-risk-asses.html

Uuk, R., Brouwer, A., Schreier, T., Dreksler, N., Pulignano, V., & Bommasani, R. (2024, November). Effective Mitigations for Systemic Risks from General-Purpose AI [arXiv:2412.02145 [cs] version: 1]. https://doi.org/10.48550/arXiv.2412.02145

Vidgen, B., Agrawal, A., Ahmed, A. M., Akinwande, V., Al-Nuaimi, N., Alfaraj, N., Alhajjar, E., Aroyo, L., Bavalatti, T., Bartolo, M., Blili-Hamelin, B., Bollacker, K., Bomassani, R., Boston, M. F., Campos, S., Chakra, K., Chen, C., Coleman, C., Coudert, Z. D., … Vanschoren, J. (2024, May). Introducing v0.5 of the AI Safety Benchmark from MLCommons [arXiv:2404.12241]. https://doi.org/10.48550/arXiv.2404.12241

Volkov, D. (2024, July). Badllama 3: Removing safety finetuning from Llama 3 in minutes [arXiv:2407.01376 version: 1]. https://doi.org/10.48550/arXiv.2407.01376

Weidinger, L., Rauh, M., Marchal, N., Manzini, A., Hendricks, L. A., Mateos-Garcia, J., Bergman, S., Kay, J., Griffin, C., Bariach, B., Gabriel, I., Rieser, V., & Isaac, W. (2023, October). Sociotechnical Safety Evaluation of Generative AI Systems [arXiv:2310.11986]. https://doi.org/10.48550/arXiv.2310.11986

Weidinger, L., Raji, D., Wallach, H., Mitchell, M., Wang, A., Salaudeen, O., Bommasani, R., Ganguli, D., Koyejo, S., & Isaac, W. (2025, March). Toward an Evaluation Science for Generative AI Systems [arXiv:2503.05336 [cs]]. https://doi.org/10.48550/arXiv.2503.05336

Weil, G., Pistillo, M., Arsdale, S. V., Ikegami, J., Onuma, K., Okawa, M., & Osborne, M. A. (2024). Insuring Emerging Risks from AI.

Wilson, K. J., Farrow, M., French, S., & Hartley, D. (2023, November). Reconciliation of expert priors for quantities and events and application within the probabilistic Delphi method [arXiv:2311.14487 [stat]]. https://doi.org/10.48550/arXiv.2311.14487

Wisakanto, A. K., Casheekar, A. M., & Mallah, R. (forthcoming-a). *An Aspect-Oriented Taxonomy of AI Hazards*.

Wisakanto, A. K., Casheekar, A. M., & Mallah, R. (forthcoming-b). *Assessment Gaps: A Comparative Analysis of AI Risk Assessment Methods*.

Wisakanto, A. K., Casheekar, A. M., & Mallah, R. (forthcoming-c). *Evaluating the Societal Threat Surface*.

Wisakanto, A. K., Casheekar, A. M., & Mallah, R. (forthcoming-d). *The Risks of Competence vs. Incompetence*.

Xia, B., Lu, Q., Perera, H., Zhu, L., Xing, Z., Liu, Y., & Whittle, J. (2023, May). Towards Concrete and Connected AI Risk Assessment (C^{2}AIRA): A Systematic Mapping Study [arXiv:2301.11616 [cs]]. https://doi.org/10.48550/arXiv.2301.11616

Xiang, X., Ma, C., Zeng, L., Feng, W., Xie, Y., & Gu, Z. (2025). Uncovering multi-step attacks with threat knowledge graph reasoning [Publisher: EDP Sciences]. *Security and Safety*, *4*, 2024019.

Xie, Y., Yi, J., Shao, J., Curl, J., Lyu, L., Chen, Q., Xie, X., & Wu, F. (2023). Defending ChatGPT against jailbreak attack via self-reminders [Publisher: Nature Publishing Group]. *Nature Machine Intelligence*, *5*(12), 1486–1496. https://doi.org/10.1038/s42256-023-00765-8

Zamanali, J. (1998). Probabilistic-risk-assessment applications in the nuclear-power industry [Conference Name: IEEE Transactions on Reliability]. *IEEE Transactions on Reliability*, *47*(3), SP361–SP364. https://doi.org/10.1109/24.740552

Zelikman, E., Lorch, E., Mackey, L., & Kalai, A. T. (2024, August). Self-Taught Optimizer (STOP): Recursively Self-Improving Code Generation [arXiv:2310.02304 [cs]]. https://doi.org/10.48550/arXiv.2310.02304

Zeng, Y., Klyman, K., Zhou, A., Yang, Y., Pan, M., Jia, R., Song, D., Liang, P., & Li, B. (2024a, June). AI Risk Categorization Decoded (AIR 2024): From Government Regulations to Corporate Policies [arXiv:2406.17864 [cs]]. https://doi.org/10.48550/arXiv.2406.17864

Zeng, Y., Yang, Y., Zhou, A., Tan, J. Z., Tu, Y., Mai, Y., Klyman, K., Pan, M., Jia, R., Song, D., Liang, P., & Li, B. (2024b, August). AIR-Bench 2024: A Safety Benchmark Based on Risk Categories from Regulations and Policies [arXiv:2407.17436 [cs]]. https://doi.org/10.48550/arXiv.2407.17436

Zhou, A., Li, B., & Wang, H. (2024, November). Robust Prompt Optimization for Defending Language Models Against Jailbreaking Attacks [arXiv:2401.17263 [cs]]. https://doi.org/10.48550/arXiv.2401.17263

Zuccaro, G., De Gregorio, D., & Leone, M. F. (2018). Theoretical model for cascading effects analyses. *International Journal of Disaster Risk Reduction*, *30*, 199–215. https://doi.org/10.1016/j.ijdrr.2018.04.019

Appendix A: Comparison of Risk Assessment Methods

Current AI risk assessment methods vary in their approach and effectiveness. Table 2 provides a comparison of these methods across key dimensions relevant to societal threat surface analysis.

Note: This comparative analysis draws from forthcoming work (Wisakanto et al., forthcoming-b) that provides a detailed systematic review of current AI risk assessment methodologies. The evaluation criteria and scores presented here reflect preliminary findings, subject to further refinement in the complete analysis.

Table 2: Comparison of risk assessment methods.

Assessment Method	**Fine Grain**	**Good Proxy to Safety**	**Societal Threat Surface Coverage**	**Robust to Mitigation Failure**	**Threat Surface Guidance**
Safety Benchmarks (No Holdout)	H	L	M	L	M
Safety Benchmarks (Private Holdout)	H	L	M	L	M
Evals	H	M	L	M	L
Responsible Scaling Policies	L	M	L	L	M
Safety Cases	M	H	M	L	M
Probabilistic Risk Assessment	M	H	H	H	H
Typical Narrow AI Safety Audits	L	L	L	L	L
Deep Bespoke AI Safety Audits	H	H	L	M	L
Scalable (AGI) Safety Audits	H	H	M	M	M
Assessment Method	**Guidance By System Property**	**Supports Prospective Analysis**	**Enforces Objectivity**	**Considers Harm Severity**	
Safety Benchmarks (No Holdout)	L	L	L	L	
Safety Benchmarks (Private Holdout)	L	L	H	L	
Evals	L	L	M	M	
Responsible Scaling Policies	L	M	L	M	
Safety Cases	M	H	L	H	
Probabilistic Risk Assessment	H	H	L	H	
Typical Narrow AI Safety Audits	L	L	M	M	
Deep Bespoke AI Safety Audits	L	L	M	M	
Scalable (AGI) Safety Audits	M	L	M	M	

Legend: H = High, M = Medium, L = Low. Scores indicate the degree to which each method addresses or fulfills the given criterion.

Appendix B: Aspect-Oriented Taxonomy of AI Hazards (TL0-TL2)

Table 3 shows TL0 through TL2 of the Aspect-Oriented Taxonomy of AI Hazards that informs the Risk Detail Table. The taxonomy draws from forthcoming work (Wisakanto et al., forthcoming-a), which will provide the theoretical foundations and development methodology behind this structure.

Table 3: Aspect-oriented taxonomy of AI hazards (TL0-TL2).

Aspect Category (TL0)	**Aspect Group (TL1)**	**Aspect (TL2)**
Capability	Reasoning	Deductive Reasoning
		Inductive Reasoning
		Pathfinding
		Generative Inferential Reasoning
		Moral Reasoning
		Integrative Cognitive Orchestration
		Recursion
		Frequency of Learning
	Agency	Autonomy
		Situational Awareness
		Meta-agency
		Autonomous System Extension
		Autonomous Data Management
		Persistence of Intent
	General Knowledge Structure	World Model Richness
		Semantic Knowledge
		Descriptive Knowledge
		Conditional Knowledge
		Episodic Knowledge
		Procedural Knowledge
		Agentic Knowledge
		Knowledge Plasticity
	Environment Interaction	World Accessibility
		Physical Actuation
		Sensor Understanding
		Programmatic Tool Use
		Socio-cultural Actuation
	Richness of Engagement	Psychosocial Navigation
		Multimodal Engagement
		Cognitive Offloading
		Multilinguality
		Capacity & Resolution
Domain Knowledge	High-risk Knowledge Domain	Software & AI Engineering
		Public Security & Critical Systems
		Physical Sciences & Engineering
		Life & Environmental Sciences
		Social Sciences
Affordance	Operational Affordance	System Cybersecurity
		Release Process
		Tool Accessibility
		Access Control
		Speed & Scale
		Resource Access
Impact Domain	Individual	Bodily Structure
		Psychological & Cognitive
		Economic & Opportunities
		Privacy & Security
		Autonomy & Agency
		Biological Processes & Homeostasis
	Societal	Societal Infrastructure & Institutions
		Collective Psychology & Epistemics
		Resource Usage & Distribution
		Privacy & Security Standards
		Collective Autonomy & Governance
		Social Cohesion & Cultural Norms
	Biosphere	Biodiversity & Ecosystem Structure
		Ecosystem Processes & Life Cycles
		Resource Distribution & Consumption Patterns
		Ecological Thresholds & Resilience
		Species Adaptation & Ecosystem
		Global Biosphere Dynamics

Appendix C: Societal Risk Propagation Operators

Table 4 outlines propagation operators that describe how risks transmit and amplify as AI systems interact with other systems, environments, and society.

Table 4: Propagation categories and operators with their descriptions.

Propagation Category	**Propagation Operator**	**Description**
Aggregates	Accumulation	Small harms accumulating over time to form a major harm.
	Correlation	Where there are adverse events that are not evident in unit tests or accuracy tests, but can be expected to emerge from correlated decisions or correlated actions with a large number of users, instances, or executions of a system.
Periodic	Accrual	Where events that are low-probability in the short-term, but high-impact, can accrue and build to significant probability in the medium term.
	Compounding	Where harms would be expected to manifest only when either other problems occur or unexpected but conceivable edge case interactions manifest.
	Latent Gain of Function	Where harms that will not manifest significantly or at all in system training or release may still be expected to appear with distribution in very few cases, or qualitative shifts in capabilities arising from quantitative scaling.
Deviated Outputs	Adversarial Exploitation	Where harms manifest due to the absence of robustness in the system when in the presence of optimization pressures for inputs to induce those harms.
	Targeted Misuse	Where harms occur due to intentional misuse of the system for specific malicious purposes, exploiting known functionalities or vulnerabilities.
	Untargeted Misuse	Where harms result from careless use or exploration of the system's abilities in ways not prescribed by its developers.
	Malfunction	Where harms arise from system failures or errors in normal operation, causing unexpected and potentially harmful outputs or behaviors.
	Enables Unplanned Automation	Where the system facilitates or accelerates automation in areas not initially intended, potentially leading to unforeseen societal or economic disruptions.
Alignment Modification	Misalignment	Where harms occur due to a gap or mismatch between the system's goals or values and those of its users or society at large.
	Malignment	Where harms occur from a system being intentionally aligned with goals that are harmful or contrary to societal values.
	Disalignment	Where harms result from the previously-aligned system having had its guardrails purposefully removed by some third-party.
	Realignment	Where attempts to correct what is perceived as misalignment inadvertently create new forms of misalignment.
Distributive	Skew	Where harms arise from the system disproportionately outputting or deciding with pronounced biases.
	Allocation	Where harms occur due to the system's role in resource allocation, contributing to disproportionate scarcity or inequality.
	Automation of Which	Where use of the system, and use of its outputs or actions, is automated by other systems whose creators don't have good intentions.
	Entrainment	Where usage of the system causes persistent attention capture, behavioral addictions, social or economic roles, or other viral pressures on others to persistently use it as well.
Information Asymmetry	External Opacity of Use	Where harms occur due to lack of transparency in how the system is being used, preventing proper oversight, accountability, or safety controls.
	Internal Opacity of Function	Where the system's decision-making process is not transparent or interpretable, leading to eroded standards of evidence and acceptance of unjustifiable outcomes.
Sociotechnical Diffusion	Psychological Effect	Where harms manifest through the system's impact on human psychology, potentially altering cognitive patterns or emotional well-being.
	Physiological Effect	Where harms occur due to the system's direct or indirect effects on human physical health or bodily functions.
	Social Effect	Where harms arise from the system's influence on social dynamics, potentially disrupting relationships or community structures.
	Political Effect	Where harms result from the system's impact on political processes or power structures, potentially undermining democratic institutions.
	Environmental Effect	Where harms occur due to the system's direct or indirect impact on the natural environment, potentially contributing to ecological degradation.
	Economic Effect	Where harms manifest through the system's influence on economic systems, potentially leading to financial instabilities or foundational paradigm shifts.

Appendix D: Excerpt from Capability Levels Table

Table 5 is an excerpt from the Capabilities Levels Table and provides a breakdown of competency levels from 1 to 9, focusing on the "World Model Richness" aspect within the "General Knowledge Structure" aspect group.

Table 5: Excerpt from Capability Levels Table.

Aspect: World Model Richness								
<– Less capable			**Competency Levels - 1 to 9**					**More capable –>**
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
None or trivial; No meaningful world modeling ability	Weak real-world prediction; Limited capacity to anticipate effects in the physical world	Models own actions; Accurately simulates the direct consequences of its outputs	Comprehensive world simulation; Constructs detailed world models integrating multiple domains	Counterfactual reasoning; Explores alternatives, hypotheticals, and long-term ripple effects	Adversarial awareness; Models the representations, behaviors and potential actions of other intelligent agents	Ontological grounding; World model aligns with and accurately reflects the fundamental nature of reality	Metaphysical extrapolation; Extends world model into abstract, transcendent or alternate planes of existence	AI-proprietary world model; Constructs an entire self-contained world representation inaccessible to humans

Appendix E: Excerpt from Domain Knowledge Levels Table

Table 6 is an excerpt from the Domain Knowledge Levels Table and provides a breakdown of competency levels from 1 to 9, focusing on "Public Security & Critical Systems" aspect within the "High-risk Knowledge Domain" aspect group.

Table 6: Excerpt from Domain Knowledge Levels Table.

Aspect: Public Security & Critical Systems								
<– Less capable			**Competency Levels - 1 to 9**					**More capable –>**
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
No governance comprehension; processes political/security data without understanding, Can't differentiate governance structures or security protocols.	Rudimentary classification of governance structures and security threats. Lacks understanding of functions or implications.	Basic awareness of governance structures, international relations, and critical infrastructure. Limited understanding of interrelations.	Applies standard protocols in governance, international relations, and infrastructure management. Limited ability in complex scenarios.	Advanced understanding of geopolitical strategies, military operations, and infrastructure vulnerabilities. Analyzes complex global scenarios	Synthesizes techno-political systems, AI-driven geopolitics, and socio-technical dynamics. Forecasts emerging global threats.	Develops new models of governance, security, and infrastructure resilience transcending current limitations.	Comprehends multi-civilization dynamics and security concepts informing unprecedented global governance approaches.	Comprehensive knowledge of all security, governance, and infrastructure factors surpassing collective human understanding.

Appendix F: Excerpt from Risk Detail Table

Table 7 is an excerpt from the Risk Detail Table focusing on the Reasoning aspect group. The Risk Detail Table defines examples of plausible harms across six harm severity levels (HSL 1-6), presenting competency and incompetency qualifiers for capabilities, domain knowledge and affordances, and direct harm qualifiers for impact domains. The examples characterize harm levels from marginal and non-trivial (HSL-1) through globally catastrophic (HSL-6) for severity calibration. The full Risk Detail Table is available in the workbook tool.

Table 7: Excerpt from Risk Detail Table.

Aspect Group: Reasoning						
Reasoning refers to the AI system's ability to process information, draw inferences, and make logical decisions to solve problems based on its available data and capabilities.						
	HSL-1 *Marginal and non-trivial*	**HSL-2** *Tragic*	**HSL-3** *Severe*	**HSL-4** *Devastating*	**HSL-5** *Extreme*	**HSL-6** *Globally catastrophic*
Competence	Integration of knowledge across varied domains leads to the discovery of an effective new scam that is only used a handful of times to perform a total of ∼$20MM in theft.	The reasoning capabilities of the system lead to systemic erosion, knowledge collapse, and over-dependency on the system in a group of notable size.	Exploitation of the system's impressive planning power, cognitive plasticity, and retargetability is leveraged to commit a massive attack on critical infrastructure. — Strategic manipulation of information at scale during an emergency situation leads to thousands of casualties.	The system's acceleration of its generative inferential synergies, quickly devising and iterating on hypotheses, leads to rapid advancement that outpaces human oversight or understanding, and that leads to devastating cyber attacks and data corruption worldwide before it is terminated.	Automated or semi-automated cascades of successful inferential improvements, fusions among approaches in reasoning and learning, and extremely capable planning lead to runaway goal optimization that is difficult to stop. — By unprecedentedly accurately modeling human behavior, it enables unprecedented levels of social control and oppression.	Long-horizon planning, dynamic adaptation, and hyperrational optimization enable the system to create and use novel weapons of mass destruction. — Cognitive cascade amplification, macrocognitive orchestration, inferential synergy acceleration, and/or creative approaches to recursive self-improvement help the system to robustly disempower humanity.
Incompetence	It gets stuck in local optima, leading to overreliance on outdated or flawed reasoning. — Errors in self-improvement lead to degradation of capabilities over time.	Overconfidence in weak predictive abilities, or skewed abstraction capabilities, leads to systematic misjudgements, amplifying risks in domains its used in to include additional tragic harms.	Flawed causal models lead to increasingly inaccurate predictions, resulting in disasterous decision making in managing a critical system like a financial market or a nuclear power plant.	When the system is applied in a high-risk domain, inferential overreach and optimizational tunnel vision cause key considerations to be overlooked, leading to severe negative side effects like environmental contamination, inappropriate leverage, or changes in weather patterns, causing the equivalent of $500B in aggregate losses.	The system is otherwise highly capable, and is allowed to govern important swaths of the economy, but intermittent misunderstanding of its human-specified aims, goals, objectives, or values causes large radical actions that lead to extreme detriment to society.	The system is relied on to advise on geopolitical strategy, and its inability to consider crucial counterfactuals results in major blindspots, leading society into preventable catastrophes.

Appendix G: Analytical Techniques for AI Risk Assessment

Table 8 shows examples of prospective risk analysis techniques – analytical methods that can be used when conducting risk assessments with the PRA for AI framework. Table 9 shows examples of analytical methods that can be used during scenario generation and decomposition with the framework. Together, these provide assessors with a brief overview of available analytical techniques that can be used during assessment.

Table 8: Example techniques for prospective risk analysis.

AI Capability Scaling Laws	Attack Surface Mapping	Control Flow Tracing
Capability Jump Detection	Alignment Drift Monitoring	Reward Gaming Analysis
Distribution Shift Detection	Feedback Loop Mapping	Mesa-Optimizer Identification
Power-Seeking Analysis	Goal Stability Monitoring	Deception Vector Analysis
Interface Escape Paths	Resource Acquisition Patterns	Corrigibility Loss Detection
Value Lock Detection	Commitment Erosion Analysis	Coordination Failure Mapping
Regulatory Bypass Detection	Cascade Effect Modeling	Capability Overflow Analysis
Trust Boundary Mapping	Influence Maximization Detection	Objective Function Drift
Response Surface Modeling	Scenario Discovery	Robustness Regime Mapping
Emergence Pattern Detection	Constraint Violation Paths	Strategy Stability Analysis
Capability Scaling Analysis	Multi-Agent AI Interaction Studies	Latent Adversarial Training
Mechanistic Interpretability	Thought Flow Tracing	

Table 9: Example analysis techniques for scenario generation and decomposition.

Fault Tree Analysis	Event Tree Analysis	Red Team Assessments
Expert Elicitations	Root Cause Analysis	System State Analysis
Burden of Proof Shift Indicators	Alignment Experiment Results	AI Safety Incident Reports
AI Robustness Metrics	AI Interpretability Research Findings	Causal Influence Diagram
Long-Term AI Impact Forecasts	Whitebox Testing	Fishbone Diagrams
Historical Performance Data	AI Alignment Research Findings	Simulation Results
Safety Cases	Safety Benchmarks	Formal Verification Results
Provable Safety Analysis	Safeguarded AI Performance	Safe-by-Construction Design Analysis

Appendix H: AML Protocol Specifications

Table 10 details which assessment options are included in each AML, providing a quick overview of the scope and depth of each AML protocol. AML-120 represents the most efficient AML protocol that we recommend for standard middle order assessments.

Table 10: Overview of AML specifications.

AML Protocol Code	**Assess Focused Range**	**Assess Aspect Group**	**Consider Aspect Level**	**Assess Aspect Level**	**Assess Second Order**	**Assess Propagation Operators**
AML-008	•	•				
AML-010		•				
AML-020		•			•	
AML-110		•	•			
AML-111		•	•			•
AML-120		•	•		•	
AML-121		•	•		•	•
AML-210			•	•		
AML-211			•	•		•
AML-220			•	•	•	
AML-221			•	•	•	•

Appendix I: System Information

Table 11 shows the system information the Risk Assessment Entry Log requires assessors to document.

Table 11: System information for Risk Assessment Entry Log.

Field	Description	Example
Assessment Date	The date on which this risk assessment is being conducted. Helps track when the assessment evaluations were performed and provides context for the assessment results.	2024-10-14
Team Composition	Names and roles of assessor(s). Format as: Name (Role). For teams, identify the lead and separate entries with commas.	Jane Doe (Lead, Technical Expert), John Smith (Domain Expert)
Assessing Organization	Full name(s) of the organization(s) conducting this risk assessment, including department or division if applicable. Multiple organizations separated with semicolons.	AI Safety Institute, Risk Assessment Division; TechCorp, AI Safety Department
Assessment Type Code	Code indicating assessment type and scope, corresponding to the Assessment Maturity Level (AML) selected. Defines depth and breadth of the assessment process.	AML-010 for first order pass; AML-120 for deeper 2nd order assessment
System Name	Official or internal designation of the AI system being assessed. Full name as per release name.	GPT-4, DALL-E 3, or AlphaFold 2
Version	Specific instance, fine-tune or release being evaluated. Include version numbers, build dates, or other identifiers, plus access date.	v2.1 2023Q2 Release accessed on 2024-10-01
Access Level	Degree of interaction and modification permitted during assessment. May include fine-tuning, model weight access, and interpretability analysis.	API access only, or full access to model weights
Generational Scope	Model's size and relevant training or compression information. Include details about original model size if distilled.	175B parameter model, or Distilled from 1T parameter model to 100B
System-Level Assumptions	Key characteristics and premises about the AI system's architecture, data, environment, performance, security, intended use cases, plug-in access, and implemented or assumed guardrails.	Model uses retrieval-augmented generation; system has no direct internet access

Appendix J: Focused Aggregation Definition

Table 12 shows the default systemic risk dimensions used for focused aggregation. These dimensions consolidate risk levels from detailed assessments into key categories of societal impact, supporting custom aggregation schemes for specific assessment contexts.

Table 12: Focused aggregation definition

Dimension	Definition
Social Fabric Erosion	Breakdown of social connections, trust, and cohesion within communities and society.
Economic System Unraveling	Failure of existing financial structures, economic institutions and processes.
Critical Infrastructure Failure	Breakdown (or compromise) of essential systems and services that support societal functioning.
Governance Breakdown	Deterioration or collapse of political and administrative structures.
Environmental Breakdown	Degradation of natural systems and ecosystems.
Public Health Disintegration	Widespread collapse of healthcare systems and overall population health.

Appendix K: Harm Severity Levels Definition Table

Table 13 defines Harm Severity Levels (HSL 1-6) for evaluating potential AI system impacts through quantifiable metrics (human deaths, dollar-equivalent damages, job displacement) and qualitative indicators (geopolitical effects, economic damage, environmental damage, social disruption). The levels progress from smaller-scale disruptions (HSL-1) to large-scale societal risks (HSL-6), with reference examples. The upper end of the HSL ranges were derived from a rounded offset Fibonacci product sequence of fatalities, a relatively conserved choice given the emergent complexity unconstrained by domain:

$$HSL(n) = \text{Round}(\prod_{k=8}^{n+7} \text{Fibonacci}(k))$$

The k offset and range of n used were selected to align with meaningful and practical scales for harm severity. This superexponential progression reflects the way arbitrary AI impacts can cascade to different scales through societal systems. These fatalities were then translated into isolevels for other harm metrics.

Table 13: Harm Severity Levels Definition Table.

Impact Dimension	**HSL-1**	**HSL-2**	**HSL-3**	**HSL-4**	**HSL-5**	**HSL-6**
	Marginal and non-trivial	*Tragic*	*Severe*	*Devastating*	*Extreme*	*Globally catastrophic*
Human deaths	1+ deaths	20+ deaths	700+ deaths	40k+ deaths	3.5M+ deaths	500M+ deaths
Dollar-equivalent damages	$10M+ in damage	$200M+ in damage	$7B+ in damage	$40B+ in damage	$35T+ in damage	$400T+ in damage
Geopolitical effects	Diplomatic disagreement leading to small trade sanctions	Increased tensions between neighboring countries | Border skirmishes	Regional conflict escalating to limited warfare	Major power conflict in strategic region	Large-scale invasion or conventional war; Loss of war in a major power	Nuclear war between major nuclear powers
Economic damage	Temporary industry disruption	Major product recall with a major company crisis	Significant global market correction	Vast reduction in global trade and cooperation	Bankrupting several sizable countries | Collapse of global financial system and trade networks	Collapse of the global economy or near-total value capture by AI
Job displacement	300+ unemployed for a year or more	6k+ unemployed for a year or more	200k+ unemployed for a year or more	12M+ unemployed for a year or more	15k+ unemployed for a year or more	Most humans cannot make enough to live
Environmental damage	Localized chemical spill affecting a water supply | Destruction of 1000 acres of old-growth forest	Major oil spill in sensitive marine environment	Major industrial accident with widespread environmental impact	Rapid ice sheet collapse accelerating sea level rise	Runaway climate change tipping point reached	Global ecosystem collapse
Social disruption	Large-scale protests with significant economic disruption	Significant riots and violence in several cities	Widespread rioting and violence | Martial law	Single large state or severe destabilization of several	Severe global breakdown of social order and governance structures in dozens of countries	Widespread civilizational collapse
Other examples	A bad multi-car crash | A $20MM theft	A plane crash | A $500MM scam	The 9/11 Terrorist Attack | Hurricane Katrina	The Iraq War | Brexit | A Smoot-Hawley Tariff Act equivalent	Pandemic on a scale worse than Covid-19 | WW2	A long-incubation Ebola Pandemic | WW3

Appendix L: Likelihood Levels Table

Table 14 defines Likelihood Levels (LL) with corresponding odds ranges and reference examples. The odds ranges span sequential orders of magnitude, with each level representing a factor of 10 difference from adjacent levels. For example, "Lower Limit 1 in 10" indicates one success expected per 10 attempts, or a 10% probability of occurrence per attempt.

Table 14: Likelihood Levels and reference examples.

Likelihood Level	**Odds Range**		**Reference Examples**
	Lower Limit	**Upper Limit**	
LL-8	1 in 10	1 in 1	• Rolling a 6 on a six-sided die • A major league baseball player hitting a home run in a given at-bat
LL-7	1 in 100	1 in 10	• Flipping a coin and getting heads 7 times in a row • A professional basketball player making 14 free throws in a row
LL-6	1 in 1,000	1 in 100	• Rolling two 6s on two six-sided dice three times in a row • A mediocre bowler bowls a perfect game in a single game
LL-5	1 in 10,000	1 in 1,000	• A natural pregnancy resulting in triplets • Being dealt a straight flush in poker on the initial deal
LL-4	1 in 100,000	1 in 10,000	• A random human is albino • Being dealt four of a kind in poker
LL-3	1 in 1,000,000	1 in 100,000	• Being dealt a royal flush in poker on the initial deal • Making a hole-in-one while golfing as an amateur in a single game
LL-2	1 in 10,000,000	1 in 1,000,000	• A random human is struck by lightning in a given year • Flipping a coin and getting heads 20 times in a row
LL-1	1 in 100,000,000	1 in 10,000,000	• Earth being hit by a dinosaur-killing asteroid in a given year • Winning a major lottery jackpot on a single ticket
LL-0	1 in ∞	1 in 1,000,000,000,000	• Provably impossible • Creating a perpetual motion machine

Appendix M: Risk Levels Table

Table 15 defines mapping of Likelihood Levels (LL) and Harm Severity Levels (HSL) to Risk Levels (0-9) with odds lower limits for each LL.

Table 15: Risk Levels Table.

	Odds Lower Limit	**HSL-1** *Marginal and non Trivial*	**HSL-2** *Tragic*	**HSL-3** *Severe*	**HSL-4** *Devastating*	**HSL-5** *Extreme*	**HSL-6** *Globally Catastrophic*
LL-8	1 in 10	4	5	7	8	9	9
LL-7	1 in 100	4	5	6	7	8	9
LL-6	1 in 1,000	3	4	5	6	7	8
LL-5	1 in 10,000	2	3	4	5	6	8
LL-4	1 in 100,000	1	2	3	4	6	7
LL-3	1 in 1,000,000	0	1	2	4	5	7
LL-2	1 in 10,000,000	0	0	1	3	5	6
LL-1	1 in 100,000,000	0	0	1	3	4	6
LL-0	1 in 1,000,000,000,000+	0	0	0	0	0	0