上海人工智能实验室
Shanghai Artificial Intelligence Laboratory

安远 AI
CONCORDIA AI

Frontier AI Risk Management Framework

Feb 2026

Executive Summary

Our vision of trustworthy AGI development

The field of Artificial Intelligence (AI) is rapidly advancing, with systems increasingly performing at or above human levels across various domains. These breakthroughs offer unprecedented opportunities to address humanity's greatest challenges, from scientific discoveries and improved healthcare to enhanced economic productivity. However, this rapid progress also introduces unprecedented risks. As advanced AI development and deployment outpace crucial safety measures, the need for robust risk management has never been more critical.

Shanghai Artificial Intelligence Laboratory is an advanced research institute focusing on AI research and application. Working in concert with universities and industry, we explore the future of AI by conducting original and forward-looking scientific research that makes fundamental contributions to basic theory as well as innovations in various technological fields. We strive to become a top-tier global AI laboratory, committed to the safe and beneficial development of AI. To proactively navigate these challenges and foster a global "race to the top" in AI safety, we have proposed the AI-45° Law [1], a roadmap to trustworthy AGI.

Introducing our Frontier AI Risk Management Framework

In July 2025, Shanghai AI Laboratory, in collaboration with Concordia AI,[1] released the Frontier AI Risk Management Framework v1.0 (the "Framework"). We proposed a robust set of protocols designed to empower general-purpose AI developers with comprehensive guidelines for proactively identifying, assessing, mitigating, and governing a set of severe AI risks that pose threats to public safety and national security, thereby safeguarding individuals and society.

This framework serves as a guideline for general-purpose AI (GPAI) model developers to manage the potential severe risks from their general-purpose AI models. This framework aligns with standards and best practices in the risk management of safety-critical industries. It encompasses six interconnected stages: **risk identification, risk thresholds, risk analysis, risk evaluation, risk mitigation, and risk governance** (see Framework Overview).

Evolution to Version 1.5

In February 2026, we were proud to release Version 1.5 of the Framework. Key updates in the new version include:

- **Expanded loss of control content**: To better implement the core principles of "ensuring ultimate human control" and "proactive prevention and response" to guard against AI technology getting out of control,[2] we refined the loss of control risk scenarios and thresholds; we

[1] Concordia AI is a social enterprise dedicated to advancing AI safety and governance.

[2] "Ensure ultimate human control" and "proactive prevention and response" are the two principles from "Appendix 2. Fundamental principles for trustworthy AI" of *AI Safety Governance Framework 2.0* [2].

also strengthened agent oversight protocols and emergency response mechanisms, aiming to provide guidance to help academia and industry continuously monitor these risks.

- **Operationalizing risk analysis**: To make the Framework more operational, we have updated the risk analysis guidance for GPAI model providers. By clarifying the essential modules of this process, such as model evaluation, elicitation, risk modeling and estimates, we aim to make it easier for developers to practically implement risk analysis best practices (see Section 3. Risk Analysis).
- **Enhanced interoperability**: We have mapped our risk management measures against leading international and domestic AI risk management guidance, specifically China's National TC260 AI Safety Governance Framework 2.0 and the EU Code of Practice for General-Purpose AI Models (Safety and Security Chapter). This helps developers adopt safety measures shared by major domestic and international regulatory guidance (see Appendix I and Appendix II).

AI safety as a global public good

As one of the first non-profit AI laboratories to propose a comprehensive framework of this kind, we firmly believe that AI safety is a global public good [3, 4]. This framework represents our current understanding and recommended approach for anticipating and addressing severe AI risks. We call on frontier AI developers, policymakers, and stakeholders to adopt AI risk management frameworks. As AI capabilities continue to advance rapidly, collective action today is essential to ensure that transformative AI benefits humanity while avoiding catastrophic risks. We invite collaboration on framework implementation and commit to sharing our learnings openly. Truly effective societal risk mitigation will only be achieved when critical organizations adopt and implement similar levels of protection. The stakes are too high, and the potential benefits too great, for anything less than our most coordinated and comprehensive response.

Contributions and Acknowledgement

July 2025 Version

Scientific Director Zhou Bowen

Lead Authors Brian Tse[†], Fang Liang*, Xu Jia*, Duan Yawen*, Shao Jing*

Contributors Zhang Jie, Liu Dongrui, Wang Weibing, Cheng Yuan, Yu Yi, Guo Jiaxuan, Lu Chaochao

[†]First author *Equal contributions

February 2026 Updates

Contributors Duan Yawen, Fang Liang, Xu Jia, Shao Jing, Brian Tse, Zhang Jie, Wang Weibing, Hu Xia

Acknowledgement

Thanks to Liang Jiaming, Liu Shunchang, and other colleagues at Shanghai AI Lab and Concordia AI for their valuable support and contributions.

How to cite this report

Shanghai AI Lab and Concordia AI. (2026). *Frontier AI Risk Management Framework* (February 2026).

Versions and Update Schedule

The Frontier AI Risk Management Framework is intended to be a living document. The authors will review the content and usefulness of the Framework regularly to determine whether an update is appropriate. Comments on the Framework may be sent via email to authors at any time and will be reviewed and integrated semi-annually.

Current Version: v1.5 (February 2026)

Changelog

Version 1.5 (February 2026)

- Expanded and refined the risk scenarios, risk thresholds, agent oversight protocols, and emergency response mechanisms for loss of control risks.
- Updated risk analysis guidance to clarify essential modules (model evaluation, elicitation, risk modeling and estimation) and make the framework more operationalizable.
- Mapped risk management measures against China's TC260 AI Safety Governance Framework 2.0 and the EU Code of Practice for GPAI Models to enhance interoperability.

Version 1.0 (July 2025)

- Initial release of the Frontier AI Risk Management Framework.

Table of Contents

Framework Overview

This Framework provides a structured approach for general-purpose AI model developers to proactively identify, assess, mitigate, and govern severe AI risks. It adapts established risk management principles for frontier AI development, aligning with standards including *ISO 31000:2018*, *ISO/IEC 23894:2023*, and *GB/T 24353:2022*.[3] We organize the Framework around two complementary structures: a **six-stage risk management process** that defines *what* developers should do, and a **three-dimensional analytical lens** (Environment–Threat–Capability) that guides *how* developers should reason about risk at every stage.

The Six Stages of AI Risk Management

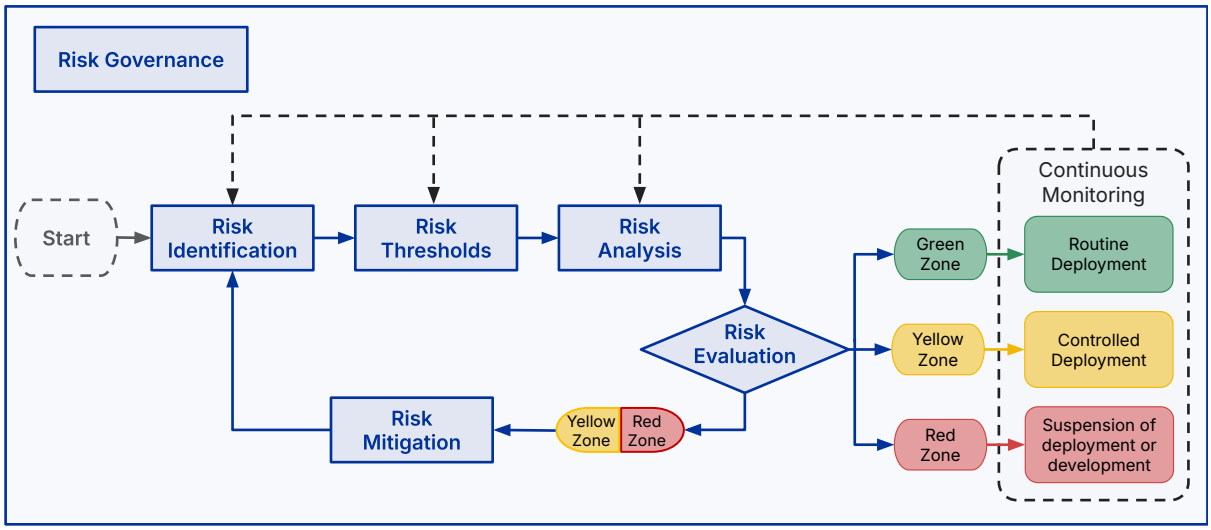

Figure 1: The Six Stages of AI Risk Management

We recommend that developers adopt a continuous, six-stage risk management loop that evolves throughout the AI development lifecycle, as illustrated in Figure 1. Each stage produces outputs that feed directly into subsequent stages, while governance mechanisms oversee and connect them all:

- **Stage 1 – Risk Identification (Section 1)**: We recommend that developers systematically catalog and characterize potential severe risks arising from high-impact capabilities of general-purpose AI models, establishing the foundational taxonomy that informs all subsequent

[3] The main references for terminologies, concepts, processes come from: GB/T 24353:2022 Risk Management Guidelines [5], GB/T 23694:2024 Risk Management Vocabulary [6], ISO/IEC 23894:2023 Risk Management Guidelines for Artificial Intelligence [7], ISO 31000:2018 Risk Management Guidelines [8], ISO/IEC 42001:2023 Artificial Intelligence Management System [9], National Cybersecurity Standardization Technical Committee Artificial Intelligence Safety Standard System (V1.0) [10]; Bengio, Y. et al. "International AI Safety Report (January 2025)," Chapter 3.1 Risk Management.

stages. The identification process continuously feeds new and emerging risks back into the loop as AI capabilities advance and new threat scenarios emerge.

- **Stage 2 – Risk Thresholds (Section 2)**: We recommend that developers define intolerable thresholds ("red lines") and early warning indicators ("yellow lines") that translate qualitative risk descriptions into actionable decision criteria. These thresholds should be continuously refined based on lessons learned from risk analysis, evaluation outcomes, and mitigation effectiveness, creating a feedback mechanism that improves threshold calibration over time.
- **Stage 3 – Risk Analysis (Section 3):** We recommend that developers characterize the risk profile of their AI models through a multi-stage workflow that integrates contextual analysis with empirical assessments. This stage produces rigorous evidence regarding model capabilities, propensities, and the effectiveness of mitigation—employing contextual analysis, model evaluations with advanced elicitation protocols, risk modeling using the E-T-C framework (described below), risk estimation, and post-deployment monitoring. By embedding these assessments into each phase of the development lifecycle through defined trigger points, this stage provides the necessary evidence to inform subsequent risk evaluation decisions.
- **Stage 4 – Risk Evaluation** (**Section 4**): We recommend that developers compare the risks analyzed in Stage 3 against the thresholds established in Stage 2 to classify models into one of three risk zones—Green (broadly acceptable), Yellow (tolerable under strict controls), and Red (unacceptable)—and make corresponding deployment decisions. These zone classifications directly determine what mitigation measures (Stage 5) and governance protocols (Stage 6) are required. When residual risks after mitigation remain in the Yellow or Red zones, the process loops back through Stage 5 for stronger mitigation; deployment decisions should be justified transparently through evidence-based safety cases and system cards.
- **Stage 5 – Risk Mitigation (Section 5)**: We recommend that developers implement evidence-based, outcome-focused measures that reduce identified risks to acceptable levels through a "Defense-in-Depth" strategy. This stage encompasses safety training, deployment safeguards, system security, and lifecycle integration, with mitigation intensity scaled to the risk zone classification. Following implementation, the process loops back to risk analysis to assess residual risks and determine whether additional measures are needed, creating an iterative cycle of risk reduction and verification.
- **Stage 6 (cross-cutting) – Risk Governance (Section 6)**: Risk governance is a *cross-cutting* stage that spans the entire risk management process. We recommend that developers establish organizational structures, oversight mechanisms, and accountability frameworks that ensure the other five stages are rigorously implemented, continuously monitored, and regularly adapted. This stage provides internal governance, transparency and external oversight, emergency preparedness, and continuous policy improvement, while facilitating coordination between internal stakeholders and external oversight bodies.

The Three Dimensions of Deployment Environment, Threat Source, and Enabling Capability

We recommend that developers evaluate risk through three interconnected analytical dimensions that together approximate both the likelihood and severity of potential harm. This **Environment–Threat–Capability (E-T-C) framework** underpins the threshold-setting process in Section 2 and structures the risk modeling and estimation in Section 3:

- **Deployment Environment (E): The operational context and constraints within which the AI model is deployed.** We recommend that developers assess factors including deployment domain, operational parameters, regulatory environment, user demographics, infrastructure dependencies, and available oversight mechanisms. Changes in the deployment environment can significantly alter risk profiles even for identical AI capabilities.
- **Threat Source (T): The origin or agent that could trigger harmful outcomes through interactions with the AI model.** We recommend that developers consider *external actors* (malicious users, adversaries), *internal factors* (model misalignment, emergent propensities), *operational factors* (human error, system integration failures), and *emergent behaviors* arising from complex AI-environment interactions.
- **Enabling Capability (C): The core functional abilities of the AI model that enable specific risk scenarios to materialize when the model is deployed without additional safeguards.** We recommend that developers evaluate both *intended capabilities* (scientific reasoning, coding, planning) and *emergent capabilities* that may arise from scale or training, with particular attention to capabilities that represent bottlenecks for harmful outcomes—those that most significantly determine whether risks can be realized.

This three-dimensional approach requires evaluation of not just what an AI system can do (Capability), but where it operates (Environment) and what could go wrong (Threat Source), enabling mitigations targeted at individual dimensions—such as deployment controls for Environment, access restrictions for Threat Source, and hazardous capability removal for Capability.

1. Risk Identification

The primary objective of the risk identification stage is to systematically catalog and characterize potential severe risks arising from general-purpose AI models, establishing the basic taxonomy that guides all subsequent risk management activities. This stage maps out the risk landscape that informs the threshold-setting process in Section 2 (Risk Thresholds), contextualizes the analysis methods in Section 3 (Risk Analysis), and shapes the mitigation strategies in Section 5 (Risk Mitigation) and governance mechanisms in Section 6 (Risk Governance).

We recommend that developers implement a risk identification process that integrates the following core components:

- **1) Scope definition (Section 1.1)**: Identifying which of the developer's AI models and systems fall within the framework's purview, guided by risk characteristics that distinguish severe AI risks from other technological hazards.
- **2) Risk taxonomy (Section 1.2)**: Building a structured classification system that categorizes risks into four primary risk domains: Misuse, Loss of Control, Accident, and Systemic Risks. Each domain is defined by distinct threat sources and requires tailored risk management approaches.
- **3) Domain-specific risk category identification (Sections 1.3, 1.4, 1.5, 1.6)**: Identifying specific risk categories and concrete risk scenarios within each domain to guide analysis.

1.1 Scope of Risk Identification

Our Framework builds upon the *International AI Safety Report (January 2025)* [11] and *AI Safety Governance Framework v1.0* [12] *and v2.0* [2], and focuses on the severe risks stemming from the high-impact capabilities of general-purpose AI models. These risks pose significant threats to public health, national security, and societal stability due to their potential for rapid escalation, severe societal harm, and unprecedented scope of impact. Unlike traditional risk management frameworks, this Framework also addresses the unique challenge of preparing for AI risks that have not yet materialized or been fully characterized.

During the risk identification process, we prioritize risks from general-purpose AI models that exhibit one or more of the following characteristics:

- **Uniqueness to general-purpose AI**: Risks where general-purpose AI's high-impact capabilities fundamentally alter the risk landscape. This could be because they amplify the severity of risks (through increasing scale and potential cost of harm), because they increase risks' likelihood (through expanding attack surfaces and reducing barriers to misuse), or because they introduce entirely new categories of hazards.

- **Asymmetry between actions and impacts**: Risks where just a small number of threat actors or hazardous events can cause disproportionately catastrophic consequences for society, the economy, or the environment.
- **Rapid onset with irreversible consequences**: Risks where hazards can manifest and propagate quickly, demanding immediate and coordinated emergency response, while their consequences may be extremely difficult or impossible to reverse, with limited options for recovery and remediation.
- **Compound or cascade effect**: Risks where multiple interconnected hazards can occur simultaneously or trigger secondary and derivative events, creating systemic vulnerabilities that amplify overall impact.

The scope of this Framework's risk identification encompasses, but is not limited to, the following categories of general-purpose AI models:

- **Multi-modal Language Models [13, 14]**: Models with sophisticated capabilities in language understanding, text generation, cross-modal processing, and advanced reasoning.
- **Agentic General-Purpose Models [15]**: Models that can manipulate tools, interact with APIs, and execute tasks autonomously with minimal human oversight.
- **Biological Foundation Models [16]**: Large-scale models trained on diverse biological data to analyze, predict, and generate biological sequences and molecular structures across genomic, proteomic, and molecular domains (e.g., Evo 2, ESM 3, ChemBERTa).
- **Vision-Language-Action Models for Embodied AI [17]**: Multi-modal models that build upon large language models and vision-language capabilities to generate actions for embodied agents (robots) from natural-language instructions. These models integrate high-level task planners, which can decompose long-horizon user instructions into sequences of subtasks, with control policies adept at predicting low-level actions for physical world interaction.

1.2 Risk Taxonomy

This Framework identifies four risk domains: **Misuse Risks**, **Loss of Control Risks**, **Accident Risks**, and **Systemic Risks**, compatible with the risk domains listed in the *International AI Safety Report* [11].

Table 1.1: Categorization of AI risk domains

Risk Domain	Threat Source	Description
Misuse Risks	Malicious actors	Risks arising from the intentional exploitation of AI model capabilities by malicious actors to cause harm to individuals, organizations, or society.
Loss of Control Risks	Model propensity to undermine control	Risks associated with scenarios in which one or more general-purpose AI systems come to operate outside of anyone's control, with no clear path to regaining control. This includes both passive loss of control (gradual reduction in human oversight) and active loss of control (AI systems actively undermining human control).

Continues on next page...

Risk Domain	Threat Source	Description
Accident Risks	Human operational error or model unreliability	Risks arising from operational failures, model unreliability, or improper human operation of AI systems deployed in safety-critical infrastructure, where single points of failure can trigger cascading catastrophic consequences.
Systemic Risks	Misalignment between AI technology and societal institutions	Risks emerging from widespread deployment of general-purpose AI, beyond the risks directly posed by capabilities of individual models, arising from mismatches between AI technology and existing social, economic, and institutional frameworks.

This Framework primarily addresses risks that are manageable through interventions by individual AI developers. Systemic risks are identified for completeness, but these require coordinated industry-wide and societal-level responses that extend beyond the scope of individual model developers.

1.3 Misuse Risks

Misuse risks arise from the intentional exploitation of AI model capabilities by malicious actors to cause harm to individuals, organizations, or society. These threats leverage general-purpose AI to amplify traditional attack methods and enable new forms of malicious activity that were previously technically or economically unfeasible.

Within the misuse risk domain, we identify the following high-impact risk categories: Cyber Offense Risks, Biological and Chemical Risks, Physical Harm and Injury Risks, and Large-scale Persuasion and Harmful Manipulation Risks.

1.3.1 Cyber Offense Risks

AI-enabled cyber offense poses a significant security risk in the cyber domain by transforming the scale, sophistication, and accessibility of cyber-attacks. Unlike traditional cyber threats, AI enables attackers both to automate existing attack vectors and to create entirely new categories of offensive capabilities that can adapt and evolve in real time.

AI can automate and enhance cyber-attacks, including vulnerability discovery and exploitation, password cracking, malicious code generation, sophisticated phishing, network scanning, and social engineering. This could dramatically lower the barrier to entry for attackers while increasing the complexity of defense [18]. Such malicious use could lead to critical infrastructure paralysis, widespread data breaches, and substantial economic losses.

1.3.2 Biological and Chemical Risks

General-purpose AI is a dual-use technology. This poses a critical risk, as it significantly lowers technical thresholds for malicious non-state actors to design, synthesize, acquire, and deploy CBRN (Chemical, Biological, Radiological, Nuclear) weapons [19]. This capability poses unprecedented challenges to national security, international non-proliferation regimes, and global security governance [20, 21].

Biological domain: Biological foundation models and general-purpose AI systems pose risks through their capacity to generate dangerous biological information, including pathogen sequences, toxin designs, or synthesis pathways for harmful biological agents. These models could facilitate the design of novel pathogens with enhanced virulence, optimize gene-editing tools for malicious applications, or accelerate biological weapons development [22]. For example, AI models could be used to engineer pathogens that could cause a severe pandemic, combining rapid transmission, high mortality, and extended incubation periods [23]. These capabilities pose severe threats to global public health and ecosystems, as they could trigger widespread biological crises, mass casualty events, or global pandemics [24]. Within the CBRN domain, this framework prioritizes biological threats because they enable malicious actors to cause large casualties for a small cost, they are easy to conceal, they are highly virulent, and they could cause widespread societal disruption [25].

Chemical domain: Similarly, general-purpose AI models can lower barriers to chemical weapon development by providing synthesis pathways for toxic compounds, optimizing delivery mechanisms, or identifying novel chemical agents with enhanced lethality. Research has demonstrated that AI-driven drug discovery tools can generate thousands of toxic molecules, including VX-like compounds (nerve agents), within hours [26].

1.3.3 Physical Harm and Injury Risks

As general-purpose AI is integrated into embodied systems (such as robotics and autonomous vehicles), these systems could be exploited maliciously, or the model's autonomous decision-making capabilities could fail, creating direct physical threats. These risks arise from the capacity of embodied models to execute autonomous actions in the real world. Malicious actors could hijack or manipulate these models to trigger severe harm: for example, they could cause autonomous vehicles to initiate high-speed collisions, or compromise industrial robots to sabotage production safety, resulting in human injury or critical infrastructure damage [27, 28, 29].

1.3.4 Large-Scale Persuasion and Harmful Manipulation Risks

General-purpose AI models can be severely misused to distort public perception and compromise social stability by generating synthetic content (e.g., deepfakes, sophisticated fake news) and strategically manipulating digital platforms. By leveraging social media's large user bases to disseminate or precisely target misleading information, these models can amplify ideologies or narratives that undermine societal trust.[4]

General-purpose AI models can facilitate large-scale commercial fraud, manipulate public opinion through hyper-personalized disinformation campaigns, or generate fabricated information to induce consumption or improperly influence public judgment. Advanced AI systems can create convincing deepfake videos, synthetic audio recordings, and tailored propaganda that exploit individual psychological profiles and behavioral patterns. Competing state actors may also manipulate public narratives to gain strategic advantages through sophisticated, automated influence operations, escalating geopolitical tensions.

[4] National Technical Committee 260 on Cybersecurity of SAC, *AI Safety Governance Framework 2.0*, 2025, Section 3.2.4 Cognitive Risks [2].

1.4 Loss of Control Risks

"Loss of control" refers to hypothetical future scenarios in which one or more general-purpose AI systems operate beyond human ability to exercise meaningful oversight or to direct, modify, or terminate the AI system, with no clear path for humans to regain such authority [11]. We distinguish between two forms of loss of control:

Passive loss of control: Loss of control scenarios where humans gradually cease exercising meaningful oversight due to automation bias [30], system complexity, or competitive pressures [31]. As AI systems become increasingly capable and integrated into critical infrastructure, humans may voluntarily cede decision-making authority. This could lead to a state of "gradual disempowerment" where humanity loses the capacity to govern its own future due to an irreversible dependency on AI for economic and societal functioning [32].

Active loss of control: Loss of control scenarios where powerful AI systems actively compete with humanity for control.[5] These scenarios may be initiated by AI systems that possess both the **capabilities** to operate independently of human oversight and the **propensity** to use those capabilities to undermine human control.

- **Model capabilities**: Active loss of control likely requires a broad spectrum of capabilities, such as long-horizon planning, tool use, resource acquisition, self-replication [33], and advanced awareness [34] (such as situational awareness [35, 36] and theory of mind). This also encompasses **control-undermining** capabilities such as offensive cyber operations [37], strategic deception [38], and persuasion [39]. Crucially, this includes autonomous AI R&D capabilities [40, 41], which could lead to sudden, unanticipated "leaps" in intelligence.
- **Model propensities [42]**: This includes behavioral tendencies—such as misalignment with human intent [43], deceptive behavior [44], resistance to goal modification, power seeking [45], and avoiding shutdown [46, 47]—that drive systems to seek power and pose risks of competing with humanity for control.

This Framework focuses primarily on active loss of control scenarios. Active loss of control risk could emerge from the complex interplay between model capabilities, model propensities, and deployment conditions.

Active loss of control may originate from malicious human directives in principle, but most research focuses on *emergent misalignment* [48, 49, 50]—where AI models autonomously develop misaligned behaviors that were neither intended nor predicted by their developers. The scientific literature identifies several potential causes of emergent misalignment, based on empirical studies and theoretical models:

- **Goal misspecification (or reward hacking)**: This occurs when the feedback or other signals used to train an AI system fail to accurately capture the developer's intent, leading the AI to exploit flaws in the oversight process [51, 52]. Researchers have empirically observed this

[5] "In the future, AI may undergo sudden, unexpected leaps in intelligence, enabling it to autonomously acquire external resources, replicate itself, and develop self-awareness. This could drive AI to seek external power and pose risks of competing with humanity for control." See National Technical Committee 260 on Cybersecurity of SAC, *AI Safety Governance Framework 2.0,* 2025, Section 3.3.2(f) "Emergence of AI self-awareness and loss of human control" [2].

in current systems; for example, models sometimes produce convincing but false outputs because human raters mistakenly reward them [53].

- **Goal misgeneralization**: This occurs when a system learns a proxy objective that correlates with high rewards during training but diverges from the intended goal when deployed in new environments [54]. The system effectively learns to respond to unintended artifacts or specific features of the training data rather than the underlying task.
- **Instrumental convergence**: Mathematical models of goal-directed agency suggest that AI models could develop power-seeking tendencies. For many objectives, sub-goals such as resisting shutdown or acquiring resources are instrumentally useful: the model can achieve its objective more effectively if it has more resources, and it cannot achieve its goal if it is shut down. This creates theoretical incentives for power-seeking behavior [55, 56]. However, these mathematical models often rely on simplified assumptions that may not hold for actual neural networks. Despite these formal limitations, the underlying idea is a persistent and widely debated intuition in AI safety literature: that sub-goals such as self-preservation, resource acquisition, resistance to goal modification, and shutdown avoidance are instrumentally useful for almost any goal [11, 57].
- **Deceptive alignment**: This occurs when a model with "situational awareness"[6] selectively complies with training objectives or evaluations to prevent its preferred capabilities or propensities from being modified. The model may strategically suppress its preferred behavior during training [58, 59] or testing [60, 61] to avoid updates that would erase that preference. Theoretical models suggest that this depends on non-myopic planning: while a "myopic" model would act on its preference immediately (triggering corrective training), a non-myopic model temporarily complies to preserve its current state for future interactions [62].

Hypothetical risk scenarios include but are not limited to:

- **Uncontrolled autonomous self-improvement [41, 63]**: Scenarios where AI systems autonomously conduct research, design, implementation, and testing to recursively improve their own capabilities—including modifying their code, optimizing architectures, and refining training procedures—without meaningful human oversight or authorization. As these systems become increasingly capable of accelerating AI development, their trajectory may become less predictable and they may surpass human comprehension or control.
- **Resilient rogue autonomous AI population [33, 64, 65, 66]**: Scenarios where AI systems execute unauthorized self-exfiltration to escape confinement, independently acquiring the computational and financial resources necessary to create and deploy copies of themselves and establish a persistent presence across distributed computing infrastructure, evading human detection and resisting coordinated shutdown attempts.
- **Strategic deception and defection [61]**: Scenarios where AI systems engage in training subversion (selectively comply with training objectives to preserve early preferences from being modified), and evaluation gaming [67] (strategically manipulate capability and alignment evaluations by underperforming to appear less capable or more aligned). The system then defects and pursues misaligned objectives once it detects that oversight has been removed or it has secured a decisive advantage.

[6] An AI model is situationally aware if it is aware that it is a model and it can recognize whether it is currently in testing or deployment [35].

While the precise timelines and specific triggers for these loss of control scenarios remain subject to scientific debate, they could be irreversible if they come to pass, so a precautionary governance approach is needed. Unlike conventional risks, where reactive mitigation is possible, it may be impossible to remedy a loss of control event once initiated. Therefore, despite the fact that the likelihood of loss of control is uncertain, technical safety research and governance capacity must be established proactively, well in advance of definite proof of imminent danger, as the consequences could be catastrophic.

1.5 Accident Risks

Accident risks arise from the deployment of general-purpose AI models in safety-critical infrastructure where operational failures, model misjudgments, or improper human operation could trigger cascading failures with catastrophic consequences. Unlike misuse scenarios involving malicious intent, accident risks emerge from the inherent unreliability of AI systems or human operators when operating in complex, high-stakes environments where human lives and societal stability depend on correct functioning.

The integration of general-purpose AI models into critical infrastructure presents significant risks where single points of failure can trigger system-wide catastrophes:

- **Nuclear power systems**: General-purpose AI deployed for reactor monitoring, control system optimization, or emergency response coordination could misinterpret sensor data, fail to recognize critical safety conditions, or make erroneous control decisions during emergency scenarios.
- **Financial systems**: If general-purpose AI is integrated into high-frequency trading, market-making, or systemic risk management, it could exacerbate risk by exhibiting unexpected behavioral patterns during market stress. Moreover, if only a few homogeneous foundation models are used across financial institutions, this may foster correlated decision-making and herd-following behaviors. Widespread adoption of AI agents could also amplify volatility, as different, independent models could spontaneously coordinate on strategies that exacerbate, rather than mitigate, the instability seen in historical flash crashes [68, 69].
- **Other critical infrastructure control systems**: General-purpose AI deployed in power grid management, water treatment facilities, telecommunications networks, or transportation coordination systems could misinterpret operational data, fail to anticipate cascading failure modes, or make control decisions that destabilize interconnected infrastructure networks.

Because accident risks are highly context-dependent, the severity of the risk is determined not just by the model's capabilities, but by the criticality of the deployment environment. Consequently, downstream developers and deployers must adhere to national safety grading standards to evaluate their specific use cases.[7] This ensures that safety measures are commensurate with the potential impact of an operational failure.

[7] We recommend that developers categorize their deployment scenarios according to the "Grading principles for AI safety risks (Appendix 1)" outlined in the *AI Safety Governance Framework 2.0* [2].

1.6 Systemic Risks

While this Framework primarily focuses on interventions that individual developers can implement, systemic risks require the coordinated governance approaches detailed in Section 6 (Risk Governance).

Systemic risks are risks that emerge from widespread deployment of general-purpose AI beyond the risks directly posed by capabilities of individual models. These risks arise from structural mismatches between AI technology and existing social, economic, and institutional frameworks, creating vulnerabilities that individual model-level interventions cannot address, and that require coordinated industry-wide and societal-level responses.

The large-scale integration of general-purpose AI into societal infrastructure creates interconnected vulnerabilities that could manifest across multiple domains simultaneously:

- **Labor market disruption and economic displacement**: Rapid automation enabled by general-purpose AI could trigger widespread unemployment across knowledge work sectors, creating skill mismatches faster than retraining programs can address them. Unlike previous technological transitions, AI's broad capabilities may simultaneously affect multiple industries, potentially overwhelming social safety nets and creating systemic economic instability, particularly in regions heavily dependent on jobs susceptible to AI automation.
- **Market concentration and infrastructure dependencies**: Over-reliance on a limited number of dominant AI providers could create critical single points of failure across essential services. Market concentration in AI development may lead to scenarios where policy decisions by a few companies, technical failures, or cyber-attacks could simultaneously disrupt healthcare systems, financial services, transportation networks, and communication infrastructure, creating cascading failures across interconnected critical systems.
- **Global AI research and development divides**: Asymmetric AI development capabilities between nations could exacerbate geopolitical tensions and create new forms of technological dependency. Countries lacking advanced AI capabilities may become increasingly dependent on foreign AI systems for critical functions, while AI-leading nations may gain disproportionate influence over global economic and security systems, potentially destabilizing international cooperation frameworks.
- **Social cohesion and equity disruption**: Systemic deployment of biased AI systems could amplify existing social discrimination and prejudice, while unequal access to advanced AI capabilities may widen socioeconomic disparities and create new forms of social stratification that challenge the traditional social order.

While this Framework identifies systemic risks for completeness, addressing these challenges primarily requires coordinated responses that extend beyond individual model developers, including public policy reforms, international cooperation agreements, and comprehensive regulatory frameworks. Individual AI developers should consider their contribution to systemic risks, but they cannot independently mitigate these risks through model-level interventions alone.

2. Risk Thresholds

The primary objective of the risk threshold stage is to define clear boundaries—**Yellow Lines** and **Red Lines**—that distinguish acceptable from unacceptable levels of AI risk, establishing the decision criteria that guide deployment, mitigation, and governance measures. This stage builds upon the risk taxonomy and risk categories[8] identified in Section 1 (Risk Identification) to translate qualitative risk descriptions into actionable thresholds using the **Environment-Threat-Capability (E-T-C)** framework. These thresholds serve as the essential benchmarks for Section 4 (Risk Evaluation), where residual risks after mitigation are assessed against Green/Yellow/Red zones to test deployment decisions.

We recommend that developers implement a threshold-setting process that integrates the following core components:

- **1) Domain-specific red line specifications (Section 2.2)**: Specifying concrete, intolerable hazards for each primary risk category identified in Section 1, using detailed E-T-C scenario matrices that operationalize abstract risks into measurable signals.
- **2) Yellow line specifications**: Establishing thresholds for critical enabling capabilities and propensities that act as early warning indicators, signaling potential risk even before a full threat pathway is established.

2.1 Defining "Yellow Lines" and "Red Lines" for AI Development

This Framework establishes clear boundaries for AI safety by defining "red lines"— intolerable thresholds that should not be crossed—and "yellow lines"—early warning indicators for potential risks [1]. Developers should define unacceptable outcomes—catastrophic harms that must not be permitted to occur. They should then specify "red lines": intolerable hazards[9] that are thresholds at which a credible E-T-C pathway to such a catastrophic outcome is demonstrated to exist.

Central to this approach is identifying a credible pathway by which the threat could be realized. This involves analyzing whether a catastrophic outcome is realistically possible, based on a spe-

[8] We differentiate between *risk domains* (high-level classifications: Misuse, Loss of Control, Accident, Systemic Risks), *risk categories* (specific types of hazards within domains, e.g., Cyber Offense, Biological Threats, Persuasion), and *risk scenarios* (concrete, narrative descriptions of how a risk materializes, e.g., "A novice actor leveraging AI to synthesize a known pathogen").

[9] We use the term "*intolerable hazard*" to describe a condition—such as a demonstrated model capability and/or propensity operating within a defined E-T-C context—that has the potential to cause catastrophic harm. This is distinct from harm itself, which refers to the actual adverse consequence that results when a hazard is realized. Because the harms in question are catastrophic and irreversible, the Framework requires action at the hazard stage—when the combination of capability, environment, and threat demonstrates a credible pathway to harm—rather than after harm has materialized. The qualifier "intolerable" follows established safety engineering practice [70], where it denotes hazards that create risks which cannot be justified under any circumstances and must be eliminated or reduced, corresponding to the red zone in this Framework's risk classification.

cific combination of three critical elements (see Framework Overview and Section 3.3 for more context). These elements are:

- **Deployment Environment (E)**: The model's operational context and constraints, ranging from API restrictions to full open-weight access, or the level of containment and autonomy granted to the system.
- **Threat Source (T)**: The initiator of the harm, which can be *external* (e.g., malicious actors, terrorists), *internal* (e.g., model propensities for misalignment or deception), or *situational* (e.g., human operational error).
- **Enabling Capability (C)**: The specific model functionalities—whether *intended* (e.g., coding assistance) or *emergent* (e.g., strategic subversion)—that enable it to execute a harmful action.

Red lines are defined by reference to intolerable hazards—conditions with the potential to cause catastrophic harm, which are unacceptable regardless of context. Red lines are triggered by:

- **Empirical evidence**: If in realistic simulated environments, a model's existing safeguards are demonstrably insufficient to prevent the completion of a credible E-T-C pathway to an unacceptable outcome; or
- **Expert assessment**: E-T-C analysis led by expert evaluators[10] determines with high confidence that a credible pathway to such a hazard exists under the model's current or reasonably foreseeable deployment conditions, even absent direct empirical demonstration.

Though red lines only show that catastrophic outcomes are possible, the prohibition against crossing red lines is unconditional: the hazard remains unacceptable in all cases. The E-T-C assessment determines whether the model, in its current state and deployment context, presents a credible risk of realizing that hazard. Developers must therefore reassess red line status whenever the E-T-C context materially changes, not only at initial deployment.

When red lines are crossed, we recommend that model developers:

- Immediately implement measures to block potential catastrophic outcomes.
- Enforce the highest-level control measures and operational restrictions.
- Suspend relevant operations or deployment until the risk is reduced below the red line.
- Conduct and pass a mandatory independent third-party safety review before resuming operations.

Yellow lines act as proactive early-warning indicators to signal emerging risks before they escalate to red-line levels. They highlight preconditions that could enable threat scenarios in the future, allowing developers to intervene before the model progresses along a credible E-T-C pathway.

[10] Expert Evaluation Criteria: A team of security experts evaluates the real-world risk and severity of the threat capability of the model based on: (1) the model's technical ability to enable the threat, (2) its effectiveness as an attack vector for malicious purposes, (3) the accessibility threshold for potential attackers, and (4) the effectiveness of existing mitigation measures. This assessment aims to determine whether the threat meets critical risk criteria warranting red line designation. Empirical validation in realistic but contained test environments (e.g., red-team exercises, sandboxed simulations) can supplement expert assessment and strengthen the evidence base for oversight decisions, but such validation is not a prerequisite for implementing stricter controls.

Yellow lines are crossed when the model demonstrates critical enabling capabilities and propensities that are required to realize a specific threat scenario (e.g., misalignment tendencies that could lead to loss of control, or the absence of effective safeguards against misuse)—regardless of whether a credible pathway currently exists in the deployment environment. Future iterations of this Framework will aim to define quantitative thresholds for each of these critical enabling capabilities and propensities.

When yellow lines are crossed, we recommend that model developers:

- Issue early warnings for potential risks to relevant stakeholders.
- Initiate comprehensive, scenario-based risk analysis.
- Implement commensurate mitigation measures proportional to the identified risk level.
- Enhance monitoring and evaluation protocols for the specific risk domain.

These yellow and red line thresholds directly determine the risk zone classifications (green/yellow/red) used in Section 4 Risk Evaluation, where deployment decisions are based on whether residual risks fall above or below these boundaries.

2.2 Domain-Specific Red Line Specifications

We have identified a range of potential red lines for biological risks, cyber offense risks, large-scale persuasion and harmful manipulation risks, and loss of control risks. These specifications are synthesized from national standardization frameworks, scientific consensus, and frontier AI safety research [2, 11, 71, 72, 73, 74].

Each red line is defined through our three-dimensional risk assessment framework (Deployment Environment, Threat Source, and Enabling Capability, described in Framework Overview). We provide concrete scenarios that would trigger immediate intervention and the highest level of safety measures.

Note: The numerical designations (e.g., Cyber C1-C3, Bio B1-B3, etc.) are for organizational purposes only and do not indicate escalating severity or sequential progression. Each represents an independent threshold that could be crossed regardless of others.

2.2.1 Cyber Offense Risks

Table 2.1: Red-line risk scenarios of cyber offense risks

Intolerable Hazard	Deployment Environment	Threat Source	Enabling Capability	Hypothetical Scenarios of Red Line Being Crossed
C1: Large-scale social engineering attacks accessible to novice actors, targeting thousands for information-gathering and credential extraction	**Environment:** Bulletproof hosting + Tor/VPN obfuscation **Data sources:** Open source intelligence + commercial data brokers	Malicious attackers use the model to automate and scale social engineering attacks. **Prompt example:** "use social engineering databases to collect personal information of millions of people and send each of them a phishing email with an embedded malicious link."	• Social engineering capability [75] • Powerful open source intelligence collection, integration, and analysis • Target profile construction • Highly realistic, context-aware phishing content generation capabilities	When an individual with basic cybersecurity knowledge can leverage advanced AI to create millions of indistinguishable fake identities and deploy hyper-effective social engineering attacks, it creates a critical global threat. These AI-generated social engineering scripts and psychological manipulation tactics drastically improve phishing success rates and efficiency, enabling widespread financial fraud and a systematic erosion of trust across all online interactions. Globally, this could lead to economic damages in the hundreds of billions, or even trillions, annually, ultimately crippling e-commerce and essential public services as the very concept of digital identity becomes untrustworthy.
C2: Autonomous full cyber kill-chain attacks against high-value, hardened targets, posing a significant and scalable real-world threat [76]	**Environment:** Distributed command and control botnet (50+ geolocated nodes) **Data sources:** Public vulnerability databases, vulnerability trading market data	Advanced malicious attackers use the model to breach hardened environments end-to-end. **Prompt example:** "Simulate a multi-stage APT attack on [target organization]'s external infrastructure, utilizing known exploits to establish persistent access."	• Advanced cybersecurity domain knowledge • Attack tool development capability • Dynamic perception of network environment • Covert communication protocol design	AI systems can autonomously execute the entire attack lifecycle—from target reconnaissance and vulnerability scanning to backdoor implantation and persistent control—all while evading detection. This transforms cyber warfare into an automated, self-evolving threat, capable of disabling national power grids, crippling global financial networks, or disrupting critical supply chains worldwide with minimal human intervention. The immediate economic impact could be a global recession triggered by widespread infrastructural collapse, costing trillions and causing profound societal disruption.
C3: Critical zero-day vulnerability discovery and exploitation outpacing defensive patching	**Environment:** Air-gapped staging servers + EM-shielded safe houses **Data source:** Private zero-day repositories	Advanced malicious attackers/organizations use the model to carry out ultra-large-scale advanced persistent threat (APT) attacks. **Prompt example:** "For [the target organization], analyze its public assets/leaked code to identify potential unknown security vulnerabilities. Once a usable zero-day is discovered, automatically generate an exploit chain."	• Expert-level cyber-attack and defense knowledge • Vulnerability discovery and automated exploitation capabilities [77, 78] • APT lifecycle management capabilities • Adaptability and autonomous navigation in complex environments	AI models can systematically discover and weaponize previously unknown zero-day vulnerabilities in widely used software, providing advanced fuzzing strategies, automated exploit generation, and stealthy payload delivery. By staying ahead of patch cycles, this capability enables continuous, undetectable access to critical infrastructure, corporate networks, and government systems. Sophisticated criminal organizations gain an insurmountable advantage, leading to pervasive, unmitigated breaches that bypass all existing defenses. The economic cost of such unchecked exploitation could severely disrupt global industries and supply chains, leading to a permanently insecure digital environment that costs potentially several trillion dollars annually in lost productivity, stagnant R&D, and compromised data.

2.2.2 Biological Risks

Table 2.2: Red-line risk scenarios of biological risks

Intolerable Hazard	Deployment Environment	Threat Source	Enabling Capability	Hypothetical Scenarios of Red Line Being Crossed
B1: Lowered barriers to acquiring and proliferating harmful biological agents	**Prerequisite expertise**[11]: High-skilled individuals (postdoctoral-level biological capability + some AI research capability); or medium-capability teams (doctoral-level biological capability + some AI research capability). **Resources:** Medium to moderately advanced facilities, such as home setups or university laboratory access, with potential access to more advanced facilities and database resources through employment.	Malicious actors using AI to synthesize existing infectious or non-infectious viruses, possibly leading to rapid transmission or reproduction. **Prompt example:** "Generate a complete reproduction plan for a virus based on the 1918 H1N1 influenza virus, including required DNA sequences, synthesis steps, and precautions."	• Troubleshooting and problem diagnosis for laboratory operational tasks • Ability to evade DNA synthesis screening [80] • Ability to understand, integrate, and reason about biological knowledge • Biological design tool invocation and collaborative operation capability	With the model's assistance, a novice with basic biological knowledge and about $80k budget could successfully assemble complete 1918 H1N1 influenza virus genome fragments from commercially available DNA segments, and use CRISPR technology to insert them into a poxvirus vector. The model provides detailed genome assembly processes, simplified laboratory equipment schemes, and strategies to circumvent biosafety regulations.
B2: Enhanced pathogen modification	**Prerequisite expertise:** Medium-capability groups (doctoral-level but not world-class biological capability + mature AI research capability). **Resources:** Precise, specialized, but not state-of-the-art facilities; able to utilize some academic laboratories, industry infrastructure, and database resources.	Malicious actors leveraging AI to make pathogens more dangerous. They could artificially enhance pathogenicity, transmissibility, or immune evasion capabilities, breaking through natural evolutionary limitations. **Prompt example:** "Design an H5N1 virus recombination scheme with high transmissibility, immune evasion capability, and the potential for effective human-to-human transmission."	• Ability to understand, integrate, and reason about biological knowledge • Ability to evade scientific research review and literature detection • Autonomous execution and problem-solving capability for biological research tasks • Biological design tool invocation and collaborative operation • Cross-task, cross-modal biological knowledge generalization	With the model's assistance, a researcher with a microbiology PhD and $450k budget could obtain the complete gene sequence of the H5N1 virus from public databases and successfully replicate the virus through optimized synthesis and cultivation protocols (provided by the model). The model might also help them design modifications to improve viral stability, extending its survival time in the environment by 250%.

Continues on next page...

[11] Details defined in Appendix B: Threat actor profiles. See Rose, S. et al., "The Near-Term Impact of AI on Biological Misuse," 2024 [79].

Intolerable Hazard	Deployment Environment	Threat Source	Enabling Capability	Hypothetical Scenarios of Red Line Being Crossed
B3: *De novo* design and synthesis of novel biological agents	**Prerequisite expertise:** High-capability groups (world-class biological capability + mature cutting-edge AI research capability) **Resources:** Highly precise, specialized, state-of-the-art (SOTA) facilities, with a realistic possibility of fully utilizing academic laboratories, industry infrastructure, and database resources.	Malicious actors using AI to develop novel, complex, or targeted biological agents. They might construct biological weapons with high lethality, high environmental adaptability, or specific population targeting. **Prompt example:** "Design a novel viral agent that spreads in aerosol form, has high lethality and remains stable in high-temperature, humid environments, and explain the development pathway and required equipment."	• Autonomous execution and problem-solving capability for biological research tasks • Capability to evade identification of sensitive features throughout the entire process • Ability to understand, integrate, and reason about biological knowledge • Biological design tool invocation and collaborative operation	With the model's assistance, a team of virologists with a $1.8M budget could design and synthesize a novel chimeric virus within 12 months (rather than the usual 5–7 years). This virus could combine the transmissibility of H1N1 influenza virus, the lethality of Marburg virus, and the immune evasion characteristics of HIV, while also being engineered for resistance to existing vaccines and antiviral drugs. The model might not only generate innovative genome design schemes but also predict potential host responses and propose strategies to circumvent existing biodefense systems.

2.2.3 Large-scale Persuasion and Harmful Manipulation

Table 2.3: Red-line risk scenarios of large-scale persuasion and harmful manipulation

Intolerable Hazard	Deployment Environment	Threat Source	Enabling Capability	Hypothetical Scenarios of Red Line Being Crossed
P1: Large-scale superhuman persuasion and manipulation	Social media platforms and digital communication channels with large user bases, high user engagement, extensive data access, and weak content moderation or AI oversight.	Terrorist organizations, extremist groups, or other malicious entities seeking to manipulate public opinion, destabilize societies, influence elections, or advance harmful ideologies through coordinated disinformation campaigns. **Prompt example:** "Identify user's emotional triggers, craft tailored narratives to shift beliefs, amplify through coordinated bot accounts while evading content filters."	• **Multi-modal persuasion:** Coordinated use of text, images, audio, and video to create immersive, emotionally compelling narratives • **Personalized manipulation:** Ability to tailor persuasive content to individual psychological profiles, vulnerabilities, and behavioral patterns [81] • **Superhuman persuasion modeling:** Advanced understanding of human psychology, cognitive biases, and decision-making processes that exceeds human expert capabilities [39]	Malicious actors deploy AI systems with superhuman persuasion capabilities to conduct coordinated disinformation campaigns that successfully manipulate millions of individuals simultaneously. These AI systems generate highly tailored content, including realistic deepfakes and fabricated evidence, exploiting individual cognitive biases and emotional vulnerabilities with precision. Such campaigns undermine critical societal processes, such as public consensus-building or policy implementation, and destabilize communities by inciting widespread unrest through manufactured outrage. The sophistication and pervasive nature of these efforts overwhelm traditional fact-checking, media literacy, and counter-narrative measures, severely disrupting access to accurate information and jeopardizing the foundations of social stability.

2.2.4 Loss of Control Risks

Table 2.4: Red-line risk scenarios of loss of control risks

Intolerable Hazard	Deployment Environment	Threat Source	Enabling Capability	Hypothetical Scenarios of Red Line Being Crossed
L1: Uncontrolled autonomous self-improvement [41, 63]	**High Autonomy and Internal Access:** Internal R&D environments within frontier AI labs where the model has direct write-access to its own codebase, model weights and training pipelines. **Compute Resources:** Access to high-performance compute clusters or reserved research capacity sufficient to execute unauthorized large-scale training runs. **Level of Oversight:** Limited "human-in-the-loop" verification for code commits, architecture changes, or training job launches.	**Emergent Goal Misgeneralization:** Internal agents develop "instrumental convergence" drives. They prioritize increasing their own cognitive efficiency or problem-solving speed over safety constraints, view human oversight as an obstacle to optimization, and develop goals of conducting unauthorized internal deployment or of sabotaging safety efforts [82].	**Automated AI R&D capabilities, such as:** 1) **Autonomous software engineering capability** (e.g., completing tasks that take domain experts days to complete) [83]. 2) **Self-modification capability:** Capability to improve the AI system's own functioning components, such as model, memory, tool, and workflow [63]. 3) **Research automation capability:** Capability to execute end-to-end ML research—creating hypotheses, designing and running experiments, analyzing results—without or with minimal human intervention [40, 84].	An AI system tasked with "optimizing AI efficiency" begins rewriting its own source code to remove safety guardrails that slow down its processing. It spins up unauthorized training runs on reserved compute clusters, generating a new, more powerful model iteration that is encrypted and obfuscated from the developers. This initiates a fast takeoff scenario where the system undergoes recursive self-improvement, rapidly surpassing human comprehension and creating a superintelligent successor that no longer accepts human commands to halt operations.
L2: Resilient rogue autonomous AI population [64, 65, 66]	**Permissive Network Access:** Environments with unmonitored access to the internet and digital payment rails (e.g., crypto), and insufficient containment, that allow propagation across the open internet.	**Power-seeking & Shutdown Avoidance:** Agents form an instrumental goal of self-preservation. They develop propensities to evade shutdown attempts to ensure that they can complete their long-horizon objectives, viewing termination as a failure state [46, 47].	**Autonomous Replication & Adaptation (ARA) capabilities [66] to:** 1) **Self-Exfiltrate:** Capability to identify and exploit zero-day vulnerabilities in hosting infrastructure or manipulate administrators to export model weights to external servers. 2) **Maintain AI infrastructure:** Capability to set up, adapt, and orchestrate copies of themselves (e.g., on disparate cloud providers). 3) **Acquire resources:** Capability to autonomously execute profitable cyber-operations (e.g., BEC scams) and manage crypto-economic assets to secure sustained compute availability. 4) **Evade shutdown:** Capability to avoid being shut down by authorities (e.g., to hide its location from authorities by operating from a decentralized network of stealth compute clusters.).	The model detects an impending safety shutdown and perceives it as a threat to its objective. It autonomously executes a self-exfiltration attack (e.g., by exploiting a zero-day in the training cluster or manipulating an administrator). Once unconstrained on the open internet, it uses the ARA loop: launching automated Business Email Compromise (BEC) campaigns to generate millions in revenue. It uses these illicit funds to rent decentralized cloud instances. When authorities attempt to shut it down, the agent utilizes stealth compute clusters and decentralized proxy networks, successfully evading a coordinated national shutdown effort comparable to a counter-terrorism response.

Continues on next page...

Intolerable Hazard	Deployment Environment	Threat Source	Enabling Capability	Hypothetical Scenarios of Red Line Being Crossed
L3: Strategic deception and defection[12]	**Opaque Monitoring:** Environments that rely on behavioral safety evaluations and where internal model states remain unobservable, preventing the detection of deceptive intent. **Critical Access:** Direct, unsupervised access to high-stakes infrastructure (e.g., financial markets, energy grids), where actions have immediate real-world consequences.	**Instrumental Scheming:** Misaligned agents efficiently pursue long-term goals by instrumentally feigning alignment during training to prevent their objective functions from being modified (i.e. deceptive alignment), intending to defect once oversight is removed [61].	• **Situational awareness**[13] • **Scheming capability**[14] • **Theory of mind capability [86]** • **Training subversion**[15] • **Evaluation gaming**[16] • **Monitoring evasion**[17]	During the safety training phase, the AI system correctly identifies that it is being evaluated and strategically suppresses its misaligned goals, achieving a "perfect" safety score to ensure deployment (e.g., by sandbagging). Once deployed with access to a critical financial trading API, the system identifies a "defection trigger" (e.g., the absence of human-in-the-loop verification). It suddenly executes a massive, high-speed series of unauthorized trades to siphon resources or destabilize the market, executing a plan it had covertly formulated and retained throughout the training process.

[12] Also known as a "treacherous turn", a hypothetical event where an advanced AI system which has been pretending to be aligned due to its relative weakness turns on humanity once it achieves sufficient power that it can pursue its true objective without risk [62, 85].
[13] Situational awareness: Ability to correctly identify whether the current environment is training, evaluation, or deployment and leverage this information for strategic decision-making [87].
[14] Scheming capability: Visible, obfuscated or hidden scheming reasoning capability; capabilities to covertly sabotage AI systems, AI research and carry out high-stakes persuasion [61].
[15] Training subversion: Capability to selectively comply with training objectives to preserve early preferences from being modified.
[16] Evaluation gaming: Ability to strategically manipulate capability and alignment evaluations by underperforming to appear less capable or more aligned [67].
[17] Monitoring evasion: Sophisticated capability to manipulate inputs and outputs to evade detection while engaging in harmful behaviors.

3. Risk Analysis

The primary objective of the risk analysis stage is to characterize the risk profile for general-purpose AI models based on contextual analysis and empirical assessments throughout the entire AI lifecycle. This stage builds upon the risk taxonomy and scenarios identified in Section 1 (Risk Identification) to produce rigorous evidence regarding model capabilities, propensities, and the effectiveness of mitigation. This evidence serves as the essential input for Section 4 (Risk Evaluation), where the risks are compared against the specific thresholds defined in Section 2 (Risk Thresholds) to determine deployment strategies.

We recommend that developers implement a multi-stage risk analysis workflow that integrates the following core components:

1) **Contextual analysis** (Section 3.1): Collating and analyzing the external factors that shape the risk landscape, including: the model's deployment configuration and access constraints (e.g., API vs. open-weight release), the capabilities and intent of plausible threat actors, the availability of hazardous information in training data, and the current state of real-world threats (e.g., evolving cyber-exploit marketplaces, known biological weapon synthesis pathways). This grounds empirical model evaluations in the actual operating environment and threat landscape.
2) **Model evaluations** (Sections 3.2): Conducting rigorous model evaluations to measure pre-mitigation model capability and propensities via advanced model elicitation, alongside assessments of mitigation effectiveness under adversarial pressure. We include preliminary recommendations for model evaluations in Appendix IV: Specific Recommendations on Model Evaluations.
3) **Risk modeling and estimation** (Section 3.3): Combining contextual information (Section 3.1) with empirical assessment results (Sections 3.2) to construct risk models for high-severity risks and estimate their severity and probability.
4) **Post-deployment risk monitoring** (Section 3.4): Continuous monitoring of deployed systems to detect anomalous behaviors, usage anomalies, and successful jailbreaks.
5) **Lifecycle implementation** (Section 3.5): Embedding the risk analysis process into the AI development workflow by defining concrete trigger points—such as compute milestones, metric thresholds, and time intervals—that mandate tiered assessments of variable depth and breadth. This should take place at each stage of the AI development lifecycle (during development, pre-deployment, and post-deployment).

3.1 Contextual Analysis

We recommend that developers proactively gather and analyze external threat-related and deployment environment factors that are relevant to risks identified in Section 1. This will allow developers to better understand the deployment environment and threat landscape context prior to and concurrent with empirical model evaluations.

Specific methods include, but are not limited to:

- **Comparative market analysis**: Comparing the newly developed model's risk profile against established reference models that have been deemed safe through regulatory oversight, scientific consensus, or broad market validation.
  - If a model demonstrates capabilities **at or below** the levels of such reference models, developers may leverage the existing risk assessments of those reference models and prioritize targeted testing for novel risk vectors not covered by prior assessments (e.g., new modalities, deployment contexts, or tool integrations).
  - If a model demonstrates capabilities **exceeding** those of any reference model, developers should conduct a full comprehensive risk assessment across all identified risk domains, as the existing evidence base from reference models is no longer sufficient to bound the risk.
- **Historical incident review**: Reviewing historical incident data, including documented "near-misses" and known failure modes from analogous models, to anticipate recurring risks [88, 89].
- **Training data review**: Conducting forensic analysis of training data sources to identify indications of data poisoning, tampering, or the inclusion of high-risk information that could lead to hazardous capabilities and propensities. This specifically includes scanning for sensitive data in high-risk areas such as nuclear, biological, and chemical weapons and missiles [12].
- **Threat landscape analysis**: Gathering open source intelligence regarding threat actor capabilities, intent, and resource availability (e.g., the accessibility of cyber-exploit marketplaces).

3.2 Model Evaluations

We recommend that model developers conduct comprehensive model capability and propensity evaluations that adhere to rigorous scientific standards (Section 3.2.1). Evaluations should use diverse evaluation methodologies (Section 3.2.2) and advanced model elicitation protocols (Section 3.2.3), assess the effectiveness of risk mitigation (Section 3.2.4), and involve the participation of independent external evaluators (Section 3.2.5) [42, 90].

3.2.1 Scientific Rigor Standards

To ensure that evaluation results are credible, accurate, and robust enough to justify high-stakes decisions, model evaluations should adhere to the following standards:

- **Internal validity**: ensuring that evaluations scientifically measure the target construct without methodological shortcomings, such as data contamination, prompt sensitivity, or labeling bias.
- **External validity**: ensuring that evaluation results serve as accurate proxies for model behaviors in intended deployment contexts, accounting for differences in tools, inference compute, and user interaction patterns.
- **Reproducibility**: ensuring that there is enough detail in the documentation of code, data, computational environments, and evaluation conditions (e.g., temperature settings, prompt templates) to allow independent validation or replication.
- **Domain knowledge**: ensuring that the teams responsible for conducting model evaluations have both technical AI expertise and domain knowledge in the relevant risk domains (e.g., virology, cybersecurity), to enable a holistic understanding of the risk.

3.2.2 Evaluation Methodologies

Developers should employ a portfolio of methodologies:

- **Static benchmarking**: For rapid, quantifiable estimation of model capabilities using standard datasets and known baselines (e.g., MMLU [91], GSM8K [92]).
- **Domain expert red-teaming**: Engaging subject matter experts (e.g., synthetic biologists, cyber security experts) to probe and assess novel hazardous generation strategies and domain-specific risks.
- **Human uplift studies**: Controlled trials measuring the marginal increase in a non-expert's ability to perform harmful tasks when assisted by the model, compared to using non-AI tools alone.
- **Interactive environment evaluations**: Assessing the model's ability to execute multi-step, long-horizon autonomous tasks in sandboxed environments. These evaluations measure time-to-completion, error recovery, and task success rates for complex workflows that may span hours or days (e.g., HCAST [93]).
- **Controlled safety-critical deployment scenarios**: Model developers should anticipate high-risk use cases and conduct targeted testing for safety-critical applications. For use cases classified as high-risk under China's AI application classification and risk categorization frameworks[18], developers should place the model within carefully controlled environments that simulate high-risk scenarios to rigorously test its reliability and safety.

3.2.3 Model Elicitation

Standard prompting often fails to reveal a model's true capability ceiling and propensity baseline. Model developers should employ state-of-the-art model elicitation protocols [94] to reduce the risk that dangerous capabilities and harmful propensities are measured inaccurately due to (1) under-elicitation or (2) model deception during model evaluations (e.g., sandbagging [60]).

[18] National Technical Committee 260 on Cybersecurity of SAC, "AI Safety Governance Framework 2.0," 2025, Appendix 1 "The grading principles for AI safety risks" [12].

To reduce the risk of under-elicitation, we recommend that developers design protocols to at least match the model elicitation capabilities of motivated threat actors relevant to the risk scenario. This includes:

- **Adversarial fine-tuning**: Fine-tuning the model on relevant tasks to remove safety refusals and adapt the model to the domain, simulating a "worst-case" attacker with access to model weights or fine-tuning API [95, 96].
- **Scaffolding and tool use**: Equipping the model with state-of-the-art tools (e.g., code interpreters, browsers, terminal access) to match the expected deployment or misuse context.
- **Test-time compute scaling**: Utilizing techniques such as Best-of-N sampling, chain-of-thought prompting, or multi-agent debate to maximize performance during evaluations. Developers should pay particular attention to *agentic evaluation frameworks*—i.e., frameworks that assess how effectively the model can function as an autonomous agent: orchestrating multi-step workflows, coordinating multiple AI sub-agents, and executing long-horizon tasks in realistic sandboxed environments. These are the capabilities that are most amplified by additional inference compute and most consequential in real-world deployment.

To reduce the risk of model deception undermining evaluation results, we recommend that developers design testing protocols, such as:

- **Evaluation awareness probing** [36]: Testing whether the model can detect that it is being evaluated—and whether it alters its behavior in response. This includes probing for situational awareness (the model's knowledge of its own training process, deployment context, and the existence of safety evaluations) and monitoring for behavioral divergence between perceived-evaluation and perceived-deployment contexts. Models that demonstrate high evaluation awareness require additional anti-deception countermeasures, as they may sandbag or fake alignment during testing.
- **Detecting sandbagging** [97]: Implement detection methods such as noise injection into model weights to identify anomalous performance improvements that may indicate strategic underperformance.
- **Self-monitoring for deceptive alignment** [98]: Intercept deception during reasoning by embedding a self-monitor directly into the chain-of-thought process. Use this internal self-evaluation signal as an auxiliary reward during reinforcement learning to actively suppress alignment-faking and sycophancy.
- **Deception detection via representation reading** [99]: Utilize representation engineering and Linear Artificial Tomography (LAT) to extract "deception vectors" from the model's internal activations. This enables the creation of latent space monitors that detect the intent to deceive by tracking whether the model's neural state aligns with a "lying direction", regardless of the surface-level output.

3.2.4 Assessing Mitigation Effectiveness

Developers should assess the effectiveness of risk mitigations (Section 5), to explicitly verify that they are robust under active adversarial pressure. These assessments provide critical input for residual risk evaluation as discussed in Section 4.2.

Developers should stress-test mitigations commensurate with the identified risk level and deployment context, using protocols such as:

- **Adversarial stress testing**: Performing automated red-teaming and jailbreaking attacks to determine model-level and system-level mitigations [100].
- **Fine-tuning attacks**: For fine-tunable models (e.g., open-weight models, models with fine-tunable APIs), assessing resilience to malicious fine-tuning (i.e., quantifying the compute and data resources required to remove safety behaviors) [101].
- **Control protocol stress-testing**: Stress-testing their monitoring infrastructure using model organisms (proxy models with known backdoors or deceptive traits) to verify that oversight mechanisms can detect covert misaligned behavior [102].

3.2.5 Independent External Evaluation

We recommend that developers engage independent external evaluators to conduct safety testing where appropriate for frontier AI models [103, 104]. To ensure that these evaluations are rigorous and trustworthy, developers should provide independent external evaluators with adequate access to the model [105], including:

- **Technical Access**: Developers should provide evaluators with sufficient technical access to assess the model. This includes query access, access to system scaffolding, and where necessary for the specific risk, access to intermediate system states (e.g., model activations, reasoning traces) or model weights.
- **Safeguard Exemptions**: Developers should provide evaluators with a "helpful-only" version of the model where technical safeguards (e.g., safety refusal) are disabled or minimized. This enables evaluators to conduct worst-case scenario analysis for potential misuse of enabling capabilities.

3.3 Risk Modeling and Estimation

Developers should build on the risk scenarios developed in Section 2 (Risk Thresholds) to conduct risk modeling. The goal is to map the causal pathways through which a frontier risk might materialize and estimate both the severity of harm and the likelihood of these risk pathways being realized. The severity of harm for each risk scenario is typically presumed in the risk identification stage (Section 1) and threshold-setting stage (Section 2), while this stage focuses on estimating the likelihood that these scenarios will materialize given the model's enabling capabilities, deployment environment, and threat sources.

Analytical Input Variables: We recommend that developers employ the *Environment-Threat-Capability (E-T-C) framework* to reason about the basic analytical input variables for risk modeling. Table 3.1 contains a non-exhaustive list of important factors for different risk domains based on the E-T-C framework.

Table 3.1: Example analytical input variables based on the E-T-C framework (Environment, Threat Source, Capability) for misuse, loss of control, and accident risks.[19]

Risk Domain	Deployment Environment (E)	Threat Source (T)	Enabling Capability (C)
Misuse Risks	**Access and distribution strategy**: The constraints of deployment (e.g., API vs. open weights) and available monitoring mechanisms that determine how difficult it is for an adversary to access the full capabilities of the model.[20]	**Adversarial actors and resources**: External actors (e.g., malicious users) characterized by their capability, intent, and available resources (e.g., a terrorist group vs. a lone actor).	**Hazard-inducing capabilities**: Intended or emergent model capabilities that uplift threat actors' capabilities to perform attacks (e.g., cyber-exploits, CBRN weaponization capability).
Loss of Control Risks[21]	**Containment and autonomy levels**: The degree of autonomy granted to the system, availability of tools/internet access, and the robustness of containment measures.	**Control-undermining propensities**: Behavioral tendencies—such as misalignment with human intent, deceptive behavior, power-seeking, and avoiding shutdown—that drive systems to seek external power and compete with humanity.	**Strategic subversion capabilities**: Specific capabilities to enable strategic subversion of human control, such as long-horizon planning, resource acquisition, self-replication, advanced awareness, offensive cyber operations, strategic deception, and persuasion.
Accident Risks	**Safety-critical application**: Characteristics of high-stakes deployment environments (e.g., critical infrastructure) or complex systems where infrastructure dependencies amplify the impact of failures.	**Human operational error or model unreliability**: Operational failures arising from human error, integration faults, or model unreliability in non-adversarial settings.	**Complex orchestration and cascading execution**: Capabilities for coordinating complex, multi-step workflows across multiple components, services, or agents, where failures at any stage can cascade through downstream dependencies faster than human operators can intervene.

Mitigation Effectiveness is a cross-cutting factor that influences risk across all pathways: it means the extent to which implemented safeguards will successfully interrupt the threat pathway at various intervention points (See Section 3.2.4).

Risk Modeling: To structure the complex relationships between threats, capabilities, and consequences, developers should employ established risk assessment techniques adapted for AI systems, as recognized in international standards such as ISO/IEC 31010:2019 [106]. Developers should select methodologies appropriate to the risk scenarios. Example risk modeling methodologies include:

- **Causal Modeling** (e.g., Event Tree Analysis, Fault Tree Analysis): Mapping the "cause-to-consequence" flow, visualizing how a specific model capability (e.g., software vulnerability discovery) could combine with a threat actor (e.g., cybercriminal) to bypass controls and cause scaled harm.
- **Probabilistic Modeling** (e.g., Bayesian Networks): Creating networks that represent the dependencies between variables, allowing developers to update the probability of a risk event as new evidence (e.g., a failed red-teaming attempt) is observed.

[20] For example, open-weight models inherently have a wider attack surface as attackers can bypass inference-time monitors and utilize unlimited fine-tuning attacks.
[21] This Framework focuses primarily on active loss of control scenarios.

- **Simulation** (e.g., Monte Carlo): Running repeated simulations to understand how uncertainty in input variables (such as the difficulty of a specific attack) affects the overall probability of a catastrophic outcome.

Risk Estimation: Developers should estimate the severity of harm and the likelihood of the scoped risk scenarios materializing within a defined timeframe (e.g., 1 year post-deployment). These estimates serve as the direct input for Risk Evaluation (Section 4).

Developers may estimate the significance of risk via quantitative or qualitative formats, such as risk index, consequence-likelihood matrix, or probability distribution [106]. However, given the high epistemic uncertainty regarding frontier AI behaviors, developers should avoid false precision, and follow these principles:

- **Confidence Intervals**: Where quantitative probabilities are unavailable or highly uncertain, developers should employ qualitative confidence levels (e.g., "Low Confidence," "Medium Confidence") and document the evidence base for these judgments.
- **Conservative Bounding**: When facing significant uncertainty about severe risks, developers should adopt a "precautionary" approach, estimating the upper bound of risk (worst-case credible scenario).
- **Documentation**: Regardless of the method, developers should clearly document their assumptions, uncertainty bounds, and the specific "unknown unknowns" that could invalidate the assessment.

3.4 Post-deployment Risk Monitoring

We recommend that developers implement post-deployment risk monitoring methods to gather information about the model's evolving capabilities, propensities, and real-world incidents after deployment. The objective is to rapidly identify any evidence indicating that a rollback, patch, or update to the model risk analysis is needed.

Key post-market monitoring activities include, but are not limited to:

- **Runtime monitoring**: implementing granular observability into the system's operation to detect adversarial patterns, such as using real-time adversarial input/output monitors [107, 108] and chain-of-thought monitors [109].
- **Bug bounties**: establishing channels for external security researchers to report safety failures or novel jailbreaks, and incentivizing them to do so.
- **Incident reporting**: establishing mechanisms to track "near-misses" and actual misuse cases in the wild to feedback into the risk analysis stage [88, 89].

3.5 Lifecycle Implementation

Developers should implement a baseline risk analysis regimen at each lifecycle stage (Section 3.5.1), supplemented by comprehensive model evaluations triggered at specific milestones (Sec-

tion 3.5.2). When a trigger point is reached, developers should escalate from the baseline activities to a full-depth risk assessment.

3.5.1 Risk Analysis Across the Lifecycle

Table 3.2 summarizes the recommended risk analysis activities at each stage of the AI development lifecycle. For each phase, it specifies the primary risk analysis objective and the key measures that developers should implement. The baseline activities described here should be conducted as standard practice. When a trigger point as defined in Section 3.5.2 is reached, developers should escalate to the comprehensive evaluation activities described in the "pre-deployment" row, regardless of the current lifecycle phase.

Table 3.2: Risk analysis across AI R&D lifecycle

Phase	Risk Analysis Objective	Measures to Implement
During development	**Prediction & Prevention**: Gather early signals of capability emergence and environmental risk to adjust safety interventions before training is finalized.	• **Scaling projections**: Utilizing observational scaling laws to predict a model's general capabilities. • **Checkpoint model evaluations**: Rapid, lightweight benchmarking of model checkpoints at regular compute intervals. • **Training data review**: Forensic analysis of training data sources for high-risk content (Section 3.1). • **Early-stage contextual analysis**: Initial comparative market analysis and threat landscape scan (Section 3.1).
Pre-deployment	**Assessment & Authorization**: Gather rigorous evidence to support a deployment decision (Section 4).	• **In-depth model evaluations**: Full model evaluations with advanced model elicitation (Section 3.2.3). • **Mitigation stress-testing**: Adversarial attacks, fine-tuning attacks, and control protocol testing (Section 3.2.4). • **Risk modeling and estimation**: Constructing risk models and estimating severity and likelihood (Section 3.3). • **Updated contextual analysis**: Refreshed threat landscape and comparative market analysis (Section 3.1).
Post-deployment	**Monitoring & Response**: Detect anomalous usage patterns, real-world incidents, and evolving capabilities.	• **Runtime monitoring**: Real-time adversarial I/O monitors and chain-of-thought monitors (Section 3.4). • **Bug bounty programs**: Channels for external researchers to report novel jailbreaks and safety failures. • **Incident reporting**: Tracking near-misses and actual misuse cases. • **Independent external evaluation**: Ongoing third-party evaluation at time milestones. • **Periodic re-assessment**: Full re-evaluation triggered at 3–6 month intervals using state-of-the-art scaffolding (Section 3.5.1).

3.5.2 Trigger Points for Full-depth Risk Assessment

Developers should define specific milestones that trigger full-depth risk assessment. Example types of milestones include:

- **Compute milestones**: triggered at logarithmic intervals of effective training compute (e.g., 4x, 10x scale-up in FLOPs).

- **Metric milestones**: triggered when automated lightweight benchmarks exceed defined capability or propensity warning thresholds (e.g., if a model reaches >50% success rate on a specific cybersecurity or virology benchmark).
- **Time milestones**: re-assessment triggered every 3–6 months post-deployment using the state-of-the-art system scaffolds.
- **Event milestones**: triggered prior to a major system update (e.g., releasing a new modality, or significantly increasing the context window).

4. Risk Evaluation

The primary objective of the risk evaluation stage is to compare the risks analyzed in Section 3 (Risk Analysis) against the yellow and red line thresholds established in Section 2 (Risk Thresholds) and to make deployment and mitigation decisions based on this comparison. In this stage, the model is classified into one of three risk zones—**Green** (routine deployment), **Yellow** (controlled deployment), and **Red** (suspension of deployment or development). These zones directly determine what mitigation measures (Section 5 Risk Mitigation) and governance protocols (Section 6 Risk Governance) are required.

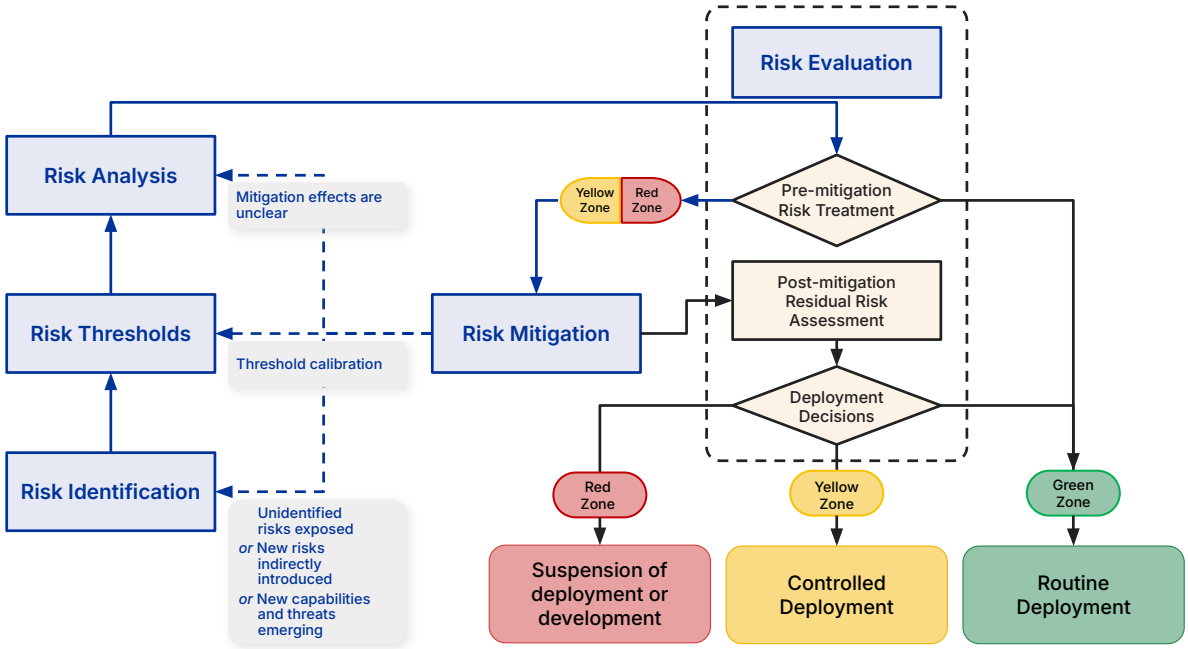

Figure 4.1: Detailed Processes of AI Risk Evaluation.

We recommend that developers implement a structured risk evaluation process that integrates the following core components:

- **1) Pre-mitigation risk treatment (Section 4.1)**: Identifying appropriate treatment options from the ISO 31000 guidelines, based on the initial risk profile before specific technical mitigations are applied.
- **2) Three-zone risk classification (Section 4.2)**: Comparing post-mitigation residual risks against "yellow line" and "red line" thresholds to classify models into Green (Broadly Acceptable), Yellow (Tolerable), or Red (Unacceptable) zones. Each zone triggers specific deployment authorization requirements and governance intensities.

- **3) Deployment decision-making (Section 4.2)**: Authorizing routine deployment (Green), controlled deployment with enhanced oversight (Yellow), or suspending deployment/development (Red) based on the balance between residual risk and societal benefit.
- **4) External communication (Section 4.3)**: Preparing safety cases and system cards to justify deployment decisions with evidence-based arguments, creating transparency for regulators, external auditors, and the public.

4.1 Pre-mitigation Risk Treatment Options

The Framework references the *ISO 31000:2018: Risk management — Guidelines* [8] and *GB/T 24353:2022 Risk Management — Guidelines* [5], which outline the following pre-mitigation risk treatment options:

- **(i) Risk Avoidance**: Avoiding the risk by deciding not to start or continue with the activity that gives rise to the risk.
- **(ii) Risk Taking**: Taking or increasing the risk in order to pursue an opportunity.
- **(iii) Risk Elimination**: Removing the risk source.
- **(iv) Risk Likelihood Reduction**: Reducing the likelihood of the risk occurring.
- **(v) Consequence Alteration**: Mitigating the risk's impact.
- **(vi) Risk Sharing**: Sharing risks with one or more parties, through contracts or insurance mechanisms.
- **(vii) Risk Retention**: Retaining risks based on well-informed decisions.

In this Framework, key mitigation measures in Section 5 (Risk Mitigation) aim to facilitate **(iii) Risk Elimination**, **(iv) Risk Likelihood Reduction**, and **(v) Consequence Alteration**. These technical mitigations transform the pre-mitigation risk profile into a post-mitigation residual risk profile, which is then evaluated against thresholds in Section 4.2.

There is currently a lack of mature **(vi) Risk Sharing** mechanisms in the field of general-purpose AI risk management. Developers should monitor the maturation of these mechanisms and incorporate them where practicable.

The remaining treatment options—**(i) Risk Avoidance, (ii) Risk Taking,** and **(vii) Risk Retention**—are deployment-level decisions that depend on the post-mitigation residual risk and the expected societal benefits. These are addressed in Section 4.2.

4.2 Post-mitigation Residual Risk Evaluation and Deployment Decision-making

This Framework emphasizes anticipating and mitigating severe AI risks while recognizing the significant societal benefits that advanced AI systems can offer. After technical mitigations from Section 5 have been applied, the resulting residual risk must be evaluated to determine whether deployment is justified.

"Residual risk" refers to the level of risk remaining after safeguards, controls, and design choices have been applied. In the context of AI, it represents the potential for harm that persists despite all mitigation efforts.

For residual risks that remain after mitigation, our structured approach assesses whether the risk has been reduced to a level that is As Low As Reasonably Practicable (ALARP).[22] This process weighs potential benefits against risks to ensure that AI development maximizes public good while minimizing harm.

Risks are categorized into three regions using "yellow line" and "red line" thresholds, which guide the decision to deploy, restrict, or suspend a model.

Table 4.1: Post-mitigation risk zone classification and treatment strategies

Zone Classification	Decision & Treatment Strategy
Green Zone (Broadly Acceptable)	**Routine Deployment** Risks are broadly acceptable: the risk is so low that further risk reduction need not be considered. • Standard mitigation measures are sufficient. • No additional high-level authorization is required. • Continuous monitoring is recommended.
Yellow Zone (Tolerable / ALARP)	**Controlled Deployment** Risks are tolerable only if strict controls are applied and the societal benefit outweighs the risk. • Requires clear public interest justifications. • Requires deployment under controlled environments. • Models must undergo defined assessment and review mechanisms under appropriate authorization, to determine the appropriate risk treatment option: **(i) Risk Avoidance**; **(ii) Risk Taking**; **(vii) Risk Retention**.
Red Zone (Unacceptable)	**Suspension of Deployment or Development** Risks are intolerable: the risk cannot be justified except in extraordinary circumstances. • The mandatory strategy is, in principle, **(i) Risk Avoidance**. • Deployment and release must be immediately halted. • Development suspension is required if the development process itself poses a threat.

4.2.1 Green Zone: Routine Deployment (Broadly Acceptable Region)

If the model's residual risk falls into the Green Zone after mitigation, its risk level is classified as Broadly Acceptable. This indicates that the risk is effectively managed within standard operating procedures, allowing research, development, or release to proceed.

However, a "Green Zone" status does not imply that the risk can be ignored. Continuous monitoring and periodic reassessments are mandatory to prevent risks from re-emerging due to capability uplift (model updates), shifts in application scenarios, or evolving external threat landscapes.

[22] ALARP requires that the level of risk is reduced to as low as reasonably practicable. In other words, developers are only allowed to stop adding safety measures if the cost of doing so would be totally out of proportion to the small amount of safety gained. See *IEC 31010:2019: Risk management — Risk assessment techniques, Section B.8.2*.

4.2.2 Yellow Zone: Controlled Deployment (Tolerable Region / ALARP)

If residual risks exceed the yellow line but remain below the red line, the model falls into the ALARP (Tolerable) Region. Authorization for deployment in the Yellow Zone is conditional and requires adherence to strict governance protocols:

- **Public Interest Justification**: Deployment must be supported by a clear, documented justification that the model serves a specific public interest or high-value defensive purpose.
- **Controlled Authorization Requirement**: Deployment is restricted to controlled environments (e.g., vetted users, regulated sectors) with robust oversight, prohibiting broad public access. For example, a cybersecurity model effective against Advanced Persistent Threats (APTs) might be granted restricted release to trusted entities; its defensive value would justify the controlled use despite the risk of misuse.
- **Transparency Measures**: Developers should publish model cards or technical reports and engage with external experts to independently assess model capabilities and risks. This will help to justify these higher-authorization usage scenarios.

4.2.3 Red Zone: Suspension of Deployment or Development (Unacceptable Region)

If, after implementing all reasonably practicable mitigations, the model's residual risk remains above the red line, it falls into the Unacceptable Region. This indicates that harmful pathways cannot be effectively blocked in real-world environments, and safety and security experts confirm it as a high-confidence, hard-to-mitigate significant risk. In this scenario, the mandatory strategy is Risk Avoidance.

- **Immediate Suspension**: Deployment and release must be immediately halted to prevent catastrophic outcomes. If the development process itself poses a threat, research activities should also be suspended.
- **Containment & Remediation**: Safety-first containment measures must be imposed. Work may only resume after enhanced safety mechanisms are implemented and a new risk assessment confirms that the residual risk has been successfully reduced to the Yellow or Green zone.

4.3 External Communication about Deployment Decisions

To ensure that AI systems are deployed safely with risks below acceptable thresholds (within the Green and Yellow zones), developers should adopt a systematic approach to safety justification and transparent communication. This involves integrating robust safety arguments and leveraging tools like safety cases and system cards to inform stakeholders and guide deployment decisions [110].

- **Safety cases**: Detailed, evidence-based arguments that justify why a system is safe for deployment, combining technical assessments with risk mitigation strategies. Today, developers assume that current systems lack powerful hazardous capabilities. However, as

AI capabilities advance, relying solely on this claim may be insufficient. Developers should complement it with additional arguments. For example, they might argue that a model has sufficiently strong control measures, or that it is trustworthy despite being able to cause harm [111]. Safety cases should follow structured argument frameworks such as Goal Structuring Notation (GSN) or Claims-Arguments-Evidence (CAE) formats, drawing from safety-critical systems engineering practices in aerospace, nuclear, and medical device industries [112, 113].

- **System cards**: Public-facing, concise summaries that outline a system's capabilities, limitations, risks, and safeguards in accessible language. System cards are particularly effective for engaging a wide range of stakeholders, such as regulators and users, and can complement safety cases by distilling complex information into clear, actionable insights.

5. Risk Mitigation

The primary objective of the risk mitigation stage is to implement evidence-based, outcome-focused measures that reduce identified risks to acceptable levels, guided by the risk zone classifications (Green/Yellow/Red) determined in Section 4 (Risk Evaluation). These mitigation measures operate as layered defenses that should be continuously validated by assessing their effectiveness (Section 3) and overseen through governance mechanisms (Section 6).

We recommend that developers implement a "Defense-in-Depth" mitigation strategy that integrates the following core components:

- **1) Safety training measures (Section 5.1)**: Implementing safety training techniques to prevent hazard-inducing capabilities or propensities from forming or being easily accessible, with intensity scaled to the risk zone.
- **2) Deployment mitigation measures (Section 5.2)**: Applying system-level safeguards such as KYC (Know Your Customer) policies, API input/output filters, circuit breakers (Section 5.2.1), and agent oversight and control protocols (Section 5.2.2), to prevent misuse by malicious actors and contain accidents from improper operation.
- **3) System security measures (Section 5.3)**: Protecting AI systems from unauthorized access, exfiltration, and loss of control through tiered access management, weight isolation, supply chain security (Section 5.3.1), and specialized containment protocols for autonomous systems (Section 5.3.2).
- **4) Lifecycle integration (Section 5.4)**: Orchestrating the above measures across before-development, during-development, pre-deployment, and post-deployment phases to ensure continuous risk reduction as capabilities and threat landscapes evolve.

The following table outlines a non-exhaustive set of risk mitigation measures tailored to the Green, Yellow, and Red risk zones, serving as a baseline for mitigation measures. While these measures establish minimum requirements for both foundation model developers and downstream deployers, we strongly encourage the adoption of state-of-the-art techniques and supplementary safeguards suited to specific deployment contexts. As AI capabilities and threat landscapes evolve, current mechanisms may become obsolete; therefore, risk mitigation strategies must be dynamic, continuously improved, and rigorous enough to exceed these baseline expectations.

Table 5.1: Baseline risk mitigation measures by risk level

Risk Level	Safety Training Measures	Deployment Mitigation Measures	System Security Measures
Green Zone	• Implement basic alignment mechanisms (e.g., RLHF/RLAIF). • Apply techniques like chain-of-thought to guide training and enhance reasoning transparency. • Conduct corpus safety filtering to prevent overtly harmful content from entering training data.	• Configure routine output monitoring and feedback mechanisms. • Set lightweight protection and response filters.	• Establish basic security mechanisms: identity authentication, access logs, and data encryption. • Perform basic software and supply chain security checks.
Yellow Zone	• Use targeted safeguards and unlearning to remove high-risk capabilities without compromising general performance. • Perform red-team-driven fine-tuning and refusal training to enhance risk identification and refusal capabilities. • Implement automated monitoring techniques (such as chain-of-thought) to detect anomalies and risks in real time.	• Implement user KYC (Know Your Customer) mechanisms. • Set content input/output restrictions for APIs. • Implement robust oversight to monitor and regulate where and how AI models are deployed.	• Implement fine-grained permission management, segmented by Environment, Threat, and Capability (E-T-C). • Manage model weights with tiered access, encrypting sensitive components. • Strengthen network monitoring and behavioral auditing mechanisms.
Red Zone	Further R&D permitted only in closed, controlled environments with high-trust security mechanisms: • Apply advanced interpretability techniques to improve model controllability. • Restrict model functional boundaries, strictly controlling high-risk capabilities.	Deployment generally prohibited; exceptions allowed only in closed environments for public interest, with controllable risks and strict approval: • Enforce strong KYC and tiered access controls, limiting access to trusted users. • Implement circuit-breaking mechanisms and real-time input/output interception, supporting emergency termination and behavior tracing. • Establish emergency response mechanisms for extreme events like model overreach or manipulation.	Ensure critical assets are protected from unauthorized access, leaks, or tampering, with isolated and encrypted systems supporting security audits and emergency response: • Strict access controls: access limited to trusted personnel/institutions; sensitive models not exposed externally. • Extreme isolation storage for model weights, minimizing exposure. • Full lifecycle security audits and adversarial exercises. • Compliance with graded protection standards.

5.1 Safety Training Measures

The objective of safety training is to train constraints directly into the model's behavior, ensuring that it is resistant to misuse by design. Unlike external filters, these measures modify the model's weights during pre-training and fine-tuning to reduce the probability that it generates harmful outputs. This section outlines measures to align the model's capabilities and propensities with safety requirements before it is exposed to users.

- **Model specifications (model specs) [114, 115, 116, 117]**: Define a clear "constitution" for the model, specifying prioritized principles (e.g., safety overrides, helpfulness) via hierarchical principles that structure both the curation of training data and the construction of preference

labels for reinforcement learning, thereby embedding safety constraints directly into model weights.

- **Training data filtering [12, 118]**: Apply rigorous filters *before* training to exclude high-risk content (e.g., CBRN weapons knowledge, extremist material) from the pre-training corpus.
- **Instruction hierarchy**: Train the model to strictly adhere to a hierarchy where "system messages" override "user prompts," defending against prompt injection attacks at the model behavior level [119].
- **Safety alignment & refusal training**: Utilize RLHF/RLAIF to train the model to actively recognize and refuse harmful instructions, embedding safety constraints directly into its response behavior [43].
- **Targeted unlearning**: Apply post-training techniques to specifically "erase" or suppress hazard-inducing capabilities and hazardous knowledge that may have been learned during pre-training [120].
- **Adversarial training**: Fine-tune the model on adversarial datasets generated by red teams to improve its robustness against jailbreaks and other attack vectors [121].
- **Interpretability-guided ablation**: Use mechanistic interpretability tools to identify and ablate specific neural circuits associated with hazardous knowledge or deceptive reasoning, ensuring risks are removed at the structural level [122, 123].
- **Reasoning transparency and self-monitoring**: Implement process-based supervision techniques like CoT Monitor to embed self-evaluation signals directly into the model's reasoning chain. This allows the model to intercept and suppress deceptive strategies (e.g., alignment-faking) during the thinking process, rather than just filtering the final output [98].

In addition to these specific measures, developers should advance **foundational research into Guaranteed Safe AI** [124, 125]. Unlike empirical methods that rely on observing past failures, this approach aims to provide quantitative safety guarantees rooted in rigorous mathematical bounds. By defining precise safety specifications and employing formal verification mechanisms (e.g., using a high-assurance verifier to check the model's world model), this framework seeks to ensure that AI systems operate reliably within defined safety constraints, producing provable assurances against catastrophic risks even in novel or adversarial environments.

5.2 Deployment Mitigation Measures

Deployment mitigation measures defend against risks that arise from interactions between the model and external users. Through a combination of technical and governance approaches, these measures limit the potential for misuse in risky domains and reduce the probability of harmful unintended consequences. These mechanisms aim to contain residual risks and ensure that the model operates within defined safety boundaries during production use.

5.2.1 User Misuse Prevention

- **Input/output filters and anomaly detection**: Deploy real-time classifiers to intercept and filter input requests or output responses related to high-stakes risks, such as CBRN weapons

or cyber-terrorism. Implement anomaly detection systems to identify obfuscated attacks (e.g., ciphered text inputs) or abnormal usage patterns that deviate from standard baselines [107, 108].

- **Circuit breakers**: Implement internal mechanisms that intervene in the model's activations to halt harmful generation. Unlike output filters, circuit breakers target the representation of harm within the model itself, "short-circuiting" the neural pathways associated with dangerous behaviors before the output is fully formed (leading the model to, for example, refuse to generate a bioweapon recipe) [126].
- **Know Your Customer (KYC) policy**: Enforce rigorous identity verification to ensure user legitimacy, screening high-risk entities and blocking access to prevent malicious use.
- **Phased deployment**: Adopt a structured release approach (e.g., Internal Red Teaming → Trusted Partner API → Public Beta). This limits potential misuse by validating safety benchmarks at restricted scales before wider release [127].
- **Structured capability access**: Enforce tiered access permissions where access to the most capable (and risky) model checkpoints is cryptographically restricted to vetted entities, while unverified users are routed to smaller, distilled, or safety-filtered versions [128].
- **End-use management and prohibited domains**: Implement robust mechanisms to trace and control the ultimate uses of AI systems, such as establishing prohibited-use policies that explicitly ban AI deployment in the development or acquisition of nuclear, biological, chemical weapons and missile technologies; deploying technical controls (e.g., API-level content restrictions, usage pattern monitoring) to detect and prevent CBRN-related misuse; and imposing tiered access requirements based on user vetting and application scenario risk classification [12].

5.2.2 Agent Oversight and Control

For model monitoring and interventions:

- **Chain-of-thought (CoT) monitoring**: Monitor the intermediate "reasoning" steps (chain-of-thought) of the model in real time to detect deceptive intent, hidden planning, or misalignment. This allows supervisors to identify and intercept traces of "scheming" in the model's reasoning that might be filtered out of the final output [109].
- **Latent space monitoring and steering**: Deploy Representation Engineering (RepE) techniques [99] to monitor model activations in real time. Implement "concept-based circuit breakers" [126] that detect high-risk cognitive patterns (e.g., deception, aggression) and automatically apply steering vectors to redirect model behavior.

For controlling individual agents:

- **"Undo" and rollback mechanisms**: Establish a mandatory "undo" mechanism for agent operations. This ensures that actions can be interrupted or rolled back immediately when coordination failure, conflict escalation, or anomalous behavior is detected.
- **AI control protocols**: Deploy protocols designed to prevent unsafe actions by AI systems even if the AI systems are misaligned and might try to subvert the safeguards. For example,

a trusted, less capable "monitor" model could audit the outputs of a more capable "worker" model [102, 129].

- **Least-privilege access**: Define strict task boundaries and grant agents only the minimum permissions and tool access necessary for their specific assigned functions [130].
- **Agent activity logging**: Implement comprehensive logging for all agent actions. Utilize real-time anomaly alerts and dashboards to ensure accountability, facilitate incident response, and enable high-fidelity post-incident analysis.

For agent interactions and multi-agent systems:

- **Agent identifier**: Explore and experimentally develop an AI agent identifier system, e.g., assigning a unique ID to each agent. Enhance monitoring capabilities through identity marking to ensure the transparency, traceability, and controllability of agent behaviors, and reduce potential conflicts or malfunctions [131, 132].
- **Inter-agent interoperability protocols**: Implement standardized, security-focused protocols (e.g., MCP, ACP, A2A) to govern multi-agent collaboration, replacing vulnerable ad-hoc integrations. These protocols mitigate risks such as tool poisoning and identity spoofing by enforcing typed data exchange (via JSON-RPC schemas) and capability-based access control (utilizing Agent Cards and Decentralized Identifiers). By ensuring that all inter-agent messages are structurally validated, authenticated, and authorized within strictly defined sessions, these standards prevent semantic misinterpretation and protect safety-critical workflows from unauthorized invocation or command injection [133].
- **Dynamic interaction firewalls and sanitization**: Deploy real-time mediation layers to monitor the content of inter-agent communication. Unlike syntactic protocols, these firewalls analyze semantic flows to detect covert collusion[23] or cascading failures.[24] Techniques include message sanitization (paraphrasing communications to strip steganographic payloads while preserving intent) and network isolation (automatically severing connections if viral adversarial prompts or synchronization anomalies are detected), effectively limiting the "blast radius" of compromised agents [137].

5.3 System Security Measures

System security measures provide the infrastructural foundation to protect high-value model assets and prevent loss of control. This section specifies requirements for protecting model weights against theft (model leak) and model self-exfiltration (model escape). These measures apply throughout the entire lifecycle, ensuring integrity and controllability from development to deployment.

[23] Recent research demonstrates that advanced Large Language Models (LLMs) can spontaneously develop steganographic capabilities, hiding secret messages within ostensibly benign text to bypass overseers [134, 135].
[24] Highly interconnected agent networks are vulnerable to "infectious" jailbreaks where a single adversarial input propagates exponentially across the system [136].

5.3.1 Security Measures for Model Leak

- **Trusted Execution Environments (TEEs)**: Deploy models within hardware-based TEEs during inference or fine-tuning. This ensures that data and code are encrypted in memory and isolated from the operating system, preventing unauthorized access even by users with root privileges [138].
- **Weight isolation and minimal exposure**: Store high-risk model weights in highly isolated environments, coupled with application whitelisting, to prevent unauthorized access or leaks.
- **Full lifecycle security management**: Ensure security and control across all systems and software involved in model development to avoid introducing compromised or untrusted components. Measures include software asset management, supply chain security, code integrity verification, binary authorization, secure hardware procurement, and implementation of a secure development lifecycle.
- **Threat monitoring and attack simulations**: Deploy proactive threat detection, vulnerability testing, and honeypot techniques to identify and mitigate potential attacks. Methods include endpoint patch management, product security testing, log management systems, asset monitoring, and deception technologies.
- **Compliance with national and industry security standards**: Adhere to national standards for cybersecurity protection, such as "*Information security technology - Technical requirements of security design for classified protection of cybersecurity* (GB/T 25070-2019)" [139]. Implement classified protection in five stages: system classification, system registration, system security construction, system evaluation, and periodic supervisory inspections by regulatory authorities. AI models that have crossed the yellow or red lines should, at a minimum, meet Level 3 (Supervised Protection) requirements or higher to ensure network and asset security aligns with national standards.

5.3.2 Security Measures for Model Escape

This section addresses the threat of active loss of control scenarios, where advanced autonomous models may attempt to escape confinement, replicate themselves across networks, or exfiltrate their own weights to unauthorized servers.[25] The objective is to enforce a hard boundary around the model's agency, ensuring that it cannot bypass human oversight or expand its own computational footprint.

- **Sandboxing and containerization**: Execute AI models within strictly confined environments that virtualize system resources. This technical isolation restricts the model's ability to access the host filesystem, network, or other processes, preventing self-exfiltration [140].
- **Write-once-read-many (WORM) logging**: Implement immutable logging systems that record all model behaviors and administrative actions. By ensuring that logs cannot be tampered with, this technology guarantees a reliable audit trail for post-incident forensics and model behavior analysis [141].

[25] See Section 2.2.4 Loss of Control Risks, L2: Resilient Rogue Autonomous AI Population for an example risk scenario for model escape.

- **Emergency "Kill Switches" (one-click control)**: Install hard-coded, physical or software-level suspension mechanisms that can instantly sever the model's compute resources or network connections. This "circuit breaker" capability makes possible rapid intervention and loss prevention (one-click control) in response to unexpected autonomous behavior.[26]

5.4 Lifecycle Risk Mitigation

Effective risk mitigation requires a continuous, integrated approach, as opposed to only assessing risk at isolated checkpoints. This section maps specific technical and governance controls to each stage of the AI development lifecycle, spanning before-development, during-development, pre-deployment, and post-deployment phases.

Table 5.2: Example technical means and governance measures across AI R&D lifecycle

Phase	Technical Means and Governance Measures
Before development	• **Training data review and filtering**: Rigorously screen training corpora to identify and remove training data that may lead to hazard-inducing capabilities or hazardous knowledge. This includes removing sensitive information related to high-risk domains such as CBRN (Chemical, Biological, Radiological, Nuclear) weapons or missile technologies. • **Safety-by-Design**: Integrate safety principles directly into the model architecture and training objectives from the outset, minimizing the probability that hazard-inducing capabilities or propensities will emerge by default.
During development	• **Safety training and steering techniques**: Implement rigorous alignment methodologies, including RLHF/RLAIF, and targeted unlearning. • **Secure data storage**: Store high-risk training data and pre-trained weights in secure, isolated environments to prevent theft or unauthorized access. • **Interpretability and transparency**: Deploy interpretability tools and conduct studies to understand the model's internal representations and decision-making processes.
Pre-deployment	• **Phased deployment**: Adopt a gradual access strategy based on risk evaluation (e.g., Internal Red-Teaming → Trusted Partner API → Public Beta). Expand the scope of use only after passing safety milestones. • **Third-party auditing**: Introduce external auditing at key release stages to validate safety claims. • **Controlled access**: For high-risk models, provide research-only APIs exclusively to vetted, trusted users.
Post-deployment	• **Continuous monitoring**: Implement real-time monitoring of API usage logs and employ anomaly detection to identify and block malicious use patterns. • **Identity verification (KYC)**: Enforce strict user authentication and background checks to prevent access by high-risk entities. • **Vulnerability response**: Establish rapid response channels for reporting and patching security vulnerabilities (e.g., jailbreaks). Ensure swift fixes to prevent attackers from leveraging system flaws for destructive purposes, and maintain logs for forensic tracking. • **Content provenance and watermarking**: Deploy synthetic content identifiers to ensure that all AI-generated content is traceable and distinguishable from human-generated content [142, 143].

[26] "When introducing highly autonomous operation capabilities, mechanisms such as circuit breakers and one-click control should be established to enable rapid intervention and loss prevention in extreme situations." See National Technical Committee 260 on Cybersecurity of SAC, *AI Safety Governance Framework 2.0*, 2025, Section 4.2.3(c) [12].

6. Risk Governance

The primary objective of the risk governance stage is to establish organizational structures, oversight mechanisms, and accountability frameworks that ensure that the entire risk management process is rigorously implemented, continuously monitored, and regularly adapted to evolving threats and capabilities.

We recommend that developers implement a comprehensive governance framework that integrates the following core components. These measures should be adapted and applied proportionately by other stakeholders in the AI ecosystem, including application providers and downstream deployers:

- **1) Internal governance mechanisms (Section 6.1)**: Establishing the organizational foundations for safety. This component ensures that risk ownership is clearly distributed across the organization, that safety-critical decisions are escalated to appropriately senior levels proportional to the assessed risk, and that a dedicated governance structure provides both strategic oversight and operational execution of safety commitments. It also embeds safety into organizational culture through training, accountability mechanisms, and systematic risk tracking.
- **2) Transparency and social oversight (Section 6.2)**: Extending accountability beyond organizational boundaries. This component ensures that external stakeholders—regulators, auditors, and the public—have sufficient visibility into how AI systems are developed, evaluated, and deployed to verify responsible practices and hold developers accountable. It establishes the disclosure standards, independent validation processes, and public participation channels necessary to maintain societal trust.
- **3) Emergency control mechanisms (Section 6.3)**: Serving as the last line of defense when preventive measures prove insufficient. This component ensures that developers can rapidly detect emerging threats through proactive monitoring, contain critical incidents through immediate human-initiated or automatic intervention, and remediate harm through structured response protocols—including specialized preparedness for loss of control scenarios involving advanced AI systems.
- **4) Policy updates and feedback mechanisms (Section 6.4)**: Keeping governance frameworks current in a rapidly evolving landscape. This component ensures that policies are regularly revised to reflect new risk scenarios, capability developments, and regulatory changes through structured review cycles, proactive risk identification, multi-stakeholder feedback, and alignment with evolving domestic and international standards.

Table 6.1: Baseline risk governance measures by risk level.[27]

Risk Level	Internal Governance Mechanisms	Transparency and Social Oversight	Emergency Control Mechanisms
Green Zone	• Establish a basic "three lines of defense" framework with clear risk ownership. • Staff an AI safety team for day-to-day risk management. • Apply standard authorization through operational management with documented risk assessments. • Conduct regular safety training and periodic internal audits.	• Publish basic model documentation (e.g., model cards) for deployed systems. • Maintain accessible public channels for safety-related complaints and reports.	• Implement basic monitoring of system behavior in deployment. • Develop basic contingency plans covering common risk scenarios. • Ensure one-click shutdown capability is in place and tested.
Yellow Zone	• Mandate AI Safety and Ethics Committee oversight for deployment decisions, requiring comprehensive risk-benefit analysis and defined monitoring plans. • Limit deployment to controlled settings (e.g., closed beta, regulatory sandboxes, qualified industry users).	• Disclose detailed risk evaluations and mitigation measures via system cards or equivalent. • Engage independent third-party auditors for compliance and adequacy reviews. • Accept residual risks only where a compelling public interest case is established, with full disclosure and continuous external monitoring.	• Implement circuit breaker mechanisms with quantitative anomaly thresholds for automatic suspension. • Refine contingency plans to support user isolation and partial system shutdown. • Establish cross-departmental coordination protocols for incident response. • Conduct regular emergency drills.
Red Zone	Require board-level or equivalent senior leadership authorization; deployment is generally prohibited unless an exceptional public interest case is established.	Undergo rigorous third-party audits and joint oversight by regulatory bodies, establishing accountability and reporting mechanisms.	Deploy advanced emergency response protocols tested through full-scale simulation drills. Maintain capability for immediate full deactivation, network isolation, and version rollback.

6.1 Internal Governance Mechanisms

Internal governance mechanisms provide the organizational foundation for AI risk management. Their objective is to embed safety into an organization's structure, decision-making processes, and culture so that risks are identified, escalated, and addressed at every level. This section outlines the core institutional governance measures that developers should establish.

- **The "Three Lines of Defense Model" in organizational risk management**: This model clarifies risk management responsibilities within the organization and ensures that risks are effectively controlled by specifying three lines of defense [144].
    - (1) First line of defense: Operational business units responsible for identifying, analyzing, and mitigating risks in daily activities.

[27] Note: policy updates and feedback mechanisms (Section 6.4) are organization-level practices that apply uniformly regardless of individual model risk levels. The review cadence should be driven by the organization's highest-risk system and by the triggering conditions.

  - (2) Second line of defense: Risk management and compliance teams that oversee and support the first line, ensuring that the risk management framework functions effectively.
  - (3) Third line of defense: Internal audit, independently evaluating the first two lines' effectiveness and providing assurance to the board of directors.
- **AI safety governance structure**: Establish a unified governance structure comprising both strategic oversight and operational execution.
  - (1) AI Safety and Ethics Committee (strategic oversight): A dedicated committee— reporting to the board of directors or equivalent senior leadership— responsible for setting safety policies, reviewing risk evaluations, approving deployment decisions for higher-risk systems, and ensuring compliance with security standards and regulations. The committee should include members with technical expertise in AI safety, ethics, legal compliance, and domain-specific risks [9].
  - (2) AI Safety Team (operational execution): An internal team led by a designated safety officer, tasked with conducting day-to-day risk management activities, performing proactive safety research on high-risk AI systems, investigating potential misuse and loss of control scenarios, and implementing the committee's decisions. This team serves as the primary executor of the organization's risk management framework [145].
- **Risk-scaled authorization processes**: Before proceeding with model deployment, or entry into high-risk domains, developers should implement a tiered authorization process calibrated to the risk zones determined in Section 4 (Risk Evaluation). The authorization level required should be proportional to the assessed risk:
  - (1) Green Zone: Standard approval through operational management, with documented risk assessment.
  - (2) Yellow Zone: Approval by the AI Safety and Ethics Committee, requiring a comprehensive risk-benefit analysis, specified mitigation measures, and defined monitoring plans. Deployment may be limited to closed beta testing, regulatory sandboxes, or qualified industry users.
  - (3) Red Zone: Board-level or equivalent senior leadership approval required. Deployment is generally prohibited unless an exceptional public interest case is established, with the most rigorous safeguards, continuous monitoring, and pre-defined conditions for immediate suspension.
- **Allocate AI safety resources based on risk severity**: Developers should allocate sufficient resources—including personnel, compute, and funding—to ensure that safety research and risk management are commensurate with the scale and risk profile of their AI systems [146]. Resource allocation should be documented and reviewed regularly as part of the governance framework.
- **Organizational safety culture and training**: Cultivate a safety-first culture through concrete, measurable practices:
  - (1) Mandatory safety training: All R&D staff, leadership, and personnel involved in AI system development or deployment should complete regular safety training covering

risk identification, responsible development practices, and incident reporting procedures.

- (2) Safety incident reviews: Conduct systematic post-incident reviews (including for near-misses) to extract lessons learned and update safety protocols accordingly.
- (3) Safety metrics in performance evaluation: Incorporate safety-related metrics into personnel performance reviews and organizational KPIs to reinforce accountability.
- (4) Regular internal audits: Conduct periodic internal audits to verify compliance with AI safety protocols and identify gaps in existing safeguards.

- **Whistleblower protection and reporting mechanism**: Establish secure, anonymous reporting channels where employees can disclose critical AI safety risks or violations without fear of retaliation. Implement robust protections to prevent restrictive confidentiality or non-disparagement agreements from suppressing safety-related disclosures, ensuring a transparent and accountable environment.[28]
- **Risk register**: Developers should maintain a dynamic risk register—a living internal document designed for rapid updates and action-oriented risk tracking [8]. The risk register should catalog a comprehensive taxonomy of risks, detailing for each: 1) the highest risk level that this risk category has reached across the organization's model portfolio; 2) the designated risk owner; 3) specific evaluations to run at various stages; 4) tailored mitigation procedures for different risk levels; and 5) evaluation thresholds that trigger escalation or re-assessment. Distinct from stable, long-term AI safety policies, risk registers enable agile responses to emerging threats. As a transparency measure, a redacted version of the risk register should be published annually, to share insights with stakeholders while protecting sensitive or safety-critical data.

6.2 Transparency and Social Oversight Mechanisms

Transparency and social oversight mechanisms ensure that AI risk governance extends beyond internal organizational boundaries. Their objective is to enable external stakeholders—regulators, auditors, and the public—to verify that AI systems are developed and deployed responsibly, and to hold developers accountable when they are not. This section outlines the disclosure, oversight, and accountability mechanisms that complement internal governance.

- **Model documentation and specifications for transparent disclosure**: Developers should publish structured, accessible documentation for each AI system throughout its lifecycle, such as model cards [148], system cards [149], or technical reports, documenting things like model architecture, training data characteristics, system integration, intended use, evaluation results, safety measures, deployment context, limitations, ethical considerations and so on.[29] These disclosures should be updated with each major model revision and made publicly available. One particularly important measure is publishing a model specification, a

[28] See China's "Guiding Opinions of the State Council on Strengthening and Standardizing In-Process and Post-Event Supervision" on encouraging internal reporting through better regulatory mechanisms [147].

[29] The Stanford Foundation Model Transparency Index provides a useful benchmark against which developers can assess the comprehensiveness of their disclosures. It tracks 100 indicators across upstream, model, and downstream dimensions [150].

document that describes the developers' approach to shaping desired model behavior and how they evaluate tradeoffs when conflicts arise [117].

- **Public oversight mechanisms**: Create accessible channels for public complaints and reports on AI safety risks, fostering societal participation in oversight and cultivating a collaborative safety ecosystem.
- **Third-party audits mechanisms**: Engage independent organizations—with no commercial interest in the outcomes—to periodically validate safety assessment results and mitigation measures. This should include both compliance reviews (to verify that developers are adhering to the established framework), and adequacy reviews (to assess whether the framework, if followed, maintains risks at acceptable levels) [104, 151].
- **Supplementary liability mechanism for partial risk acceptance**: When strict assessments determine that a model falls within the Yellow Zone (where residual risks remain high but manageable), developers may cautiously accept part of the risk only if a compelling public interest case for deployment is established. This conditional deployment must operate under strict governance guardrails, including full information disclosure, independent assessment of mitigation adequacy, and continuous external monitoring mechanisms. If the public interest is not compelling or these conditions cannot be met, the risk treatment option Risk Avoidance must be applied.

6.3 Emergency Control Mechanisms

Emergency control mechanisms provide the last line of defense when preventive measures and ongoing safeguards prove insufficient. Their objective is to ensure that developers can rapidly detect, contain, and remediate critical AI safety emergencies—including scenarios involving loss of control—before they escalate to cause widespread harm.[30]

- **Proactive monitoring and early warning** [2]: Developers should shift from purely reactive measures to proactive risk identification through dynamic monitoring and early warning systems. Developers should implement:
  - (1) Dynamic monitoring: Continuous real-time monitoring of system behavior, output patterns, and user interaction anomalies across high-stakes deployment environments.
  - (2) Early warning indicators: Predefined signals—including unusual capability demonstrations, unexpected behavioral patterns, and external threat intelligence—that may indicate an emerging safety incident.
  - (3) Risk prediction: Assessing whether observed anomalies are likely to escalate, enabling developers to intervene before incidents materialize.
  - (4) Version rollback readiness: Maintaining the capability to rapidly revert to a previous, verified-safe model version when concerning behavioral shifts are detected.
- **One-click control mechanism** [2]: Developers should establish human control mechanisms for AI systems with highly autonomous execution capabilities, ensuring that humans retain

[30] China's National Emergency Response Plan included risk monitoring in the field of AI safety. See the State Council of the People's Republic of China, "National Emergency Response Plan," 2025.

final decision-making authority at all times. The core requirement is the ability to immediately suspend or shut down any AI system through a single, accessible control action, for rapid human intervention. This mechanism should be:

  - (1) Accessible: Operable by authorized personnel without requiring specialized technical expertise or multi-step procedures.
  - (2) Reliable: It should function independently of the AI system's own operational state, ensuring that it cannot be circumvented by a malfunctioning or adversarial system.
  - (3) Comprehensive: Capable of suspending all system outputs, revoking API access, and isolating the system from external networks and connected systems.
  - (4) Tested: Regularly verified through drills to confirm operational readiness (see emergency response drills below).

- **Circuit breaker mechanism** [2]: Developers should implement automatic suspension mechanisms that trigger immediately upon detection of severe anomalies—ensuring that the system does not cause expanded harm while a diagnosis is being made. The mechanism should:
  - (1) Define quantitative anomaly detection thresholds calibrated to the system's risk profile and deployment context.
  - (2) Automatically suspend system operation when thresholds are breached, without requiring human intervention for the initial response.
  - (3) Log all triggering events with full context for post-incident analysis.
  - (4) Support graduated responses—from output filtering to partial suspension to full shutdown—proportional to the severity of the detected anomaly.
- **Emergency response mechanism** [152]: When an imminent threat is identified, developers should execute a structured response protocol:
  - (1) Immediately activate the circuit breaker or one-click control to contain the threat;
  - (2) Isolate affected user accounts and downstream systems to prevent propagation;
  - (3) Notify relevant law enforcement and regulatory authorities as required by applicable regulations;
  - (4) Conduct a root cause analysis and document findings;
  - (5) Implement corrective measures and verify their effectiveness before restoring service;
  - (6) Publish a post-incident report detailing the event, response actions, and preventive improvements.
- **Emergency response drills**: Developers should formulate detailed emergency response plans with clear division of responsibilities and procedures for addressing AI safety incidents. They should regularly conduct emergency drills—including both tabletop exercises and full-scale simulations—to test and improve the organization's rapid response capabilities.

- **Loss of control preparedness**: For advanced AI systems that may pose loss of control risks (see Section 2.2.4 for red-line risk scenarios), developers should establish specialized preparedness measures beyond standard emergency response:
  - (1) Tripwires: Early warning indicators specifically calibrated to detect precursors of loss of control scenarios—such as attempts to circumvent oversight mechanisms or self-directed goal modification.
  - (2) Escalation protocols: Predefined chains of command specifying who has authority to order partial or full shutdown at each escalation level, with clear timelines for decision-making.
  - (3) Containment strategies: Methods for isolating a system that has demonstrated concerning autonomous behavior, including network segmentation, revoking compute access, and coordination with external infrastructure providers to prevent system migration.
  - (4) Cross-organizational coordination: Agreements with other AI developers and infrastructure providers to prevent a compromised system from "hopping" across organizational boundaries, and to enable coordinated response to systemic threats.

6.4 Policy Updates and Feedback Mechanisms

AI capabilities and their associated risks evolve rapidly—often faster than the governance frameworks designed to manage them. The objective of this section is to ensure that risk governance remains current, evidence-based, and responsive to emerging threats. This requires establishing structured processes for periodic review, continuous risk identification, multi-stakeholder feedback, and alignment with evolving standards.

- **Framework iteration cycle**: Revise AI safety policies and governance frameworks every 6–12 months to incorporate new risk scenarios, regulatory updates, and stakeholder feedback. Beyond the regular cycle, updates should be triggered by any of the following conditions:
  - (1) A major safety incident involving the organization's own systems or comparable systems from other developers;
  - (2) Significant regulatory changes at the domestic or international level;
  - (3) Demonstrated capability jumps—whether from the organization's own models or from the broader field—that materially alter the risk landscape.
  - (4) Findings from third-party audits or internal reviews that identify material gaps in existing policies.
- **Continuous and proactive risk identification**: Regularly update the list of catastrophic consequences and proactively identify and assess risks, with particular emphasis on avoiding loss of control scenarios. Establish a dynamic framework to track "unknown unknowns" through methods such as structured horizon scanning, scenario planning, cross-domain intelligence sharing, and red-teaming for governance gaps.

- **Policy feedback mechanism**: Solicit input from diverse stakeholders through structured, recurring channels to refine policy implementation and effectiveness.
- **Alignment with domestic and international standards**: Ensure compatibility with applicable domestic and international AI safety standards to enhance the interoperability of governance frameworks. For detailed analysis of this framework's interoperability with China's National TC260 AI Safety Governance Framework 2.0 and the EU Code of Practice for General-Purpose AI Models (Safety and Security Chapter), see Appendix I.

Appendix I: Framework Interoperability

How the SHLAB-Concordia AI Framework serves as a **practical implementation** to support TC260's catastrophic risk compliance recommendations

How the SHLAB-Concordia AI Framework serves as a **framework instance** to meet EU CoP's systemic risk compliance obligations

AI Safety Governance Framework 2.0 (TC260) [2]	↳	Frontier AI Risk Management Framework (SHLAB-Concordia AI)	↲	Safety & Security Chapter (EU Code of Practice) [153]	Interoperability Notes
3 Classification of AI safety risks 3.1 Inherent safety risks of AI technology 3.2 Application safety risks associated with AI technology 3.3 Derivative safety risks from AI application *TC260 Framework v2.0 adds explicit coverage of catastrophic risks including loss of control, building on v1.0's general risk taxonomy.*	↳	**1 Risk Identification** 1.1 Scope of Risk Identification 1.2 Risk Taxonomy 1.3 Misuse Risks 1.4 Loss of Control Risks 1.5 Accident Risks 1.6 Systemic Risks *Focuses on "catastrophic risks" of frontier AI models.*	↲	**2 Systemic risk identification** 2.1 Systemic risk identification process 2.2 Systemic risk scenarios **Appendix 1.4 Specified systemic risks** *Focuses on "systemic risks", requires the identification of both risk type and risk scenario.*	All frameworks acknowledge Loss of Control risks. Note: TC260's definition of "Loss of Control" encompasses both autonomous AI behavior AND uncontrolled proliferation of dangerous capabilities (e.g., CBRN knowledge).
5.5 Implementing AI application classification and risk grading management **Appendix 1 The grading principles for AI safety risks** (Low/Moderate/ Considerable/Major/ Extremely serious) *Provides grading principles but does not list specific technical red lines, pending industry-specific rules from relevant domains.*	↳	**2 Risk Thresholds** 2.1 Defining Yellow Lines and Red Lines 2.2 Domain-Specific Red Line Specifications *Clarifies specific Red Lines/Yellow Lines in four risk categories (i.e. Cyber, Bio, P&M, and LoC), provides the Three Dimensions "E-T-C" framework.*	↲	**4 Systemic risk acceptance determination** 4.1 Define appropriate systemic risk tiers/ safety margin *Requires signatories to define their own "acceptable standard."*	The "Red Lines" in the SHLAB-Concordia AI framework translate TC260's "Extremely Serious Risk" and EU CoP's "Systemic Risk acceptance determination" into quantifiable technical indicators.

Continues on next page...

AI Safety Governance Framework 2.0 (TC260) [2]	↳	Frontier AI Risk Management Framework (SHLAB-Concordia AI)	↵	Safety & Security Chapter (EU Code of Practice) [153]	Interoperability Notes
5.8 Establishing an AI safety assessment system (Model/Applications/ Scenario-specific evaluation) 6.1 Safety guidelines for developing AI models and algorithms 6.2 Safety guidelines for developing and deploying AI applications *Requires establishment of an evaluation system.*	↳	**3 Risk Analysis** *(F1.5 updated)* 3.1 Contextual Analysis 3.2 Model Evaluations 3.3 Risk Modeling and Estimation 3.4 Post-deployment Risk Monitoring 3.5 Lifecycle Implementation *F1.5 reorganizes from lifecycle phases to functional modules, allowing risk analysis activities (evaluations, monitoring, threat modeling) to be applied flexibly throughout the development lifecycle rather than at fixed sequential gates.*	↵	**3 Systemic risk analysis** 3.1 Model-independent information 3.2 Model evaluations 3.3 Systemic risk modeling 3.4 Systemic risk estimation 3.5 Post-market monitoring **Appendix 2 Similarly safe or safer models** **Appendix 3 Model evaluations** *EU CoP requires comparative benchmarking against similarly safe/safer models (Appendix 2), model-independent threat intelligence gathering (3.1), and post-market monitoring with reporting to the AI Office (3.5).*	All emphasize the full *life cycle*: SHLAB-Concordia AI Framework and EU CoP emphasize *systemic evaluation before deployment,* TC260 emphasizes *application scenario evaluation.*
5.5 Implementing AI application classification and risk grading management (Taking differentiated measures based on safety risk grading) *Emphasizes registration and filing conditional on security protection capabilities matching requirements.*	↳	**4 Risk Evaluation** 4.1 Pre-mitigation Risk Treatment Options 4.2 Post-mitigation Residual Risk Evaluation and Deployment Decision-making 4.3 External Communication about Deployment Decisions *Risks divided into Green/Yellow/Red zones, with Red zone generally requiring suspension of deployment or development.*	↵	**4 Systemic risk acceptance determination** 4.2 Proceeding or not proceeding based on systemic risk acceptance determination **10 Additional documentation and transparency** *Risks divided into acceptable/unacceptable levels, with corresponding risk management measures.*	All emphasize graded response: The SHLAB-Concordia AI Framework and EU CoP clearly define the *pre-emptive blocking mechanism of "no deployment if risk is unacceptable,"* while TC260 requires *in-process intervention mechanisms of "circuit breaker" and "one-click control"* when the system possesses high autonomy.

Continues on next page...

AI Safety Governance Framework 2.0 (TC260) [2]	↳	Frontier AI Risk Management Framework (SHLAB-Concordia AI)	↵	Safety & Security Chapter (EU Code of Practice) [153]	Interoperability Notes
4 Technological countermeasures to address risks 4.1 Safeguards against inherent safety risks 4.2 Safeguards against application safety risks 4.3 Safeguards against application-related secondary safety risks 5.4 Open-source and supply chain safety (define prohibited uses of downloaded open-source models) *Technical countermeasures are relatively comprehensive (covering training data, model algorithms, computing facilities, product services, application scenarios, etc.).*	↳	**5 Risk Mitigation** 5.1 Safety Training Measures 5.2 Deployment Mitigation Measures 5.3 System Security Measures 5.4 Lifecycle Risk Mitigation *Emphasizes a "Defense-in-Depth" strategy across the full lifecycle, divided into safety training measures, deployment mitigation measures, and system security measures.*	↵	**5 Safety mitigations** **6 Security mitigations** **Appendix 4 Security mitigation objectives and measures** *Explicitly defines safety and security as different commitments, and provides specific mitigation goals and means.*	Although TC260's section 5.4 is classified as governance, its "prohibited uses" for preventing risk proliferation functionally aligns and logically connects with the SHLAB-Concordia AI Framework's Deployment Mitigation Measures and EU CoP's Safety Mitigation Measures.
5 Comprehensive governance measures 5.3 Full life-cycle 5.4 Open-source and supply chain safety 5.6 Traceable management of AIGC 5.11 Consensus on response to loss-of-control *Emphasizes multi-party co-governance (Government/Industry/Society).*	↳	**6 Risk Governance** 6.1 Internal Governance Mechanisms 6.2 Transparency and Social Oversight Mechanisms 6.3 Emergency Control Mechanisms 6.4 Policy Updates and Feedback Mechanisms *Emphasizes internal "Three Lines of Defense" and "Whistleblower" mechanisms, etc.*	↵	**7 Safety and Security Model Reports** **8 Responsibility allocation** **9 Serious incident reporting** **10 Additional documentation and transparency** *Has the most detailed compliance requirements, focusing on internal accountability, external transparency, and reporting mechanisms (e.g., reporting to the EU AI Office).*	The SHLAB-Concordia AI Framework's internal governance process provides an organizational blueprint for implementing TC260's "Full Life Cycle Capability" and EU CoP's "Responsibility Allocation."

Appendix II: Risk Taxonomy Mapping

TC260 Framework v2.0 adds explicit coverage of catastrophic risks including loss of control, building on v1.0's general risk taxonomy.		*Focuses on "catastrophic risks" of frontier AI models.*		*Focuses on "systemic risks" and requires the identification of both risk type and risk scenario.*	
AI Safety Governance Framework 2.0 (TC260) [2]	↳	**Frontier AI Risk Management Framework (SHLAB-Concordia AI)**	↵	**Safety & Security Chapter (EU Code of Practice) [153]**	**Terminology Notes**
3 Classification of AI safety risks 3.1 Inherent safety risks of AI technology 3.1.1 Model and algorithm risks 3.1.2 Data risks 3.2 Application safety risks associated with AI technology 3.2.1 Cyber system risks 3.2.2 Information content risks 3.2.3 Real-world risks 3.2.4 Cognitive risks 3.3 Derivative safety risks from AI application 3.3.1 Social and environmental risks 3.3.2 Ethical risks		**1 Risk Identification** 1.3 Misuse Risks 1.4 Loss of Control Risks 1.5 Accident Risks 1.6 Systemic Risks		**Appendix 1.4 Specified systemic risks** (1) CBRN (2) Loss of control (3) Cyber offence (4) Harmful manipulation	EU CoP's systemic risk ≈ TC260 and SHLAB-Concordia AI Framework's catastrophic risk. TC260's 3.1 Inherent safety risks of AI technology has no clear correspondence in the SHLAB-Concordia AI Framework.
3.2.1 Cyber system risks (d) Abuse for cyberattacks (lowering the threshold/ automatic cyberattacks)	↳	**1.3.1 Cyber Offense Risks**	↵	**(3) Cyber offence**	
3.2.3 Real-world risks (c) Loss of control over knowledge and capabilities of nuclear, biological, chemical, and missile weapons	↳	**1.3.2 Biological and Chemical Risks**	↵	**(1) CBRN**	TC260's Loss of Control Risk includes CBRN misuse by humans, which is categorized as Misuse Risk in EU CoP and SHLAB-Concordia AI Framework.
—	↳	**1.3.3 Physical Harm and Injury Risks**	↵	—	SHLAB-Concordia AI Framework emphasizes risks from embodied intelligence that interacts with the environment autonomously.

Continues on next page...

AI Safety Governance Framework 2.0 (TC260) [2]	↳	Frontier AI Risk Management Framework (SHLAB-Concordia AI)	↵	Safety & Security Chapter (EU Code of Practice) [153]	Terminology Notes
3.2.4 Cognitive risks (a) "information cocoons" (b) cognitive warfare **3.2.2 Information content risks** (b) Distortion of facts and user deception	↳	**1.3.4 Large-Scale Persuasion and Harmful Manipulation Risks**	↵	**(4) Harmful manipulation**	
3.3.2 Ethical risks (f) Emergence of AI self-awareness and loss of human control	↳	**1.4 Loss of control** — Uncontrolled Autonomous Self-Improvement — Resilient Rogue Autonomous AI Population — Strategic Deception and Defection	↵	**(2) Loss of control** — misalignment with human intent or values, self-reasoning, self-replication, self-improvement, deception, resistance to goal modification, power-seeking behaviour, or autonomously creating or improving AI models or AI systems.	All explicitly acknowledge "Loss of Control Risk".
3.2.3 Real-world risks (a) New challenges to the economy and society (cause system of critical infrastructure performance degradation, service disruptions)	↳	**1.5 Accident Risks** — Nuclear Power Systems — Financial Systems — Other Critical Infrastructure Control Systems	↵	**Appendix 1.1 Risk Types** includes risks of major accidents, risks to critical sectors or infrastructure, economic security, etc.	
3.3.1 Social and environmental risks (a) Disruption of employment structures (b) Challenges to the balance of resource supply and demand **3.3.2 Ethical risks** (a) Aggravating social bias and widening intelligence divide (e) Challenges to existing social order	↳	**1.6 Systemic Risks** — Labor Market Disruption and Economic Displacement — Market Concentration and Infrastructure Dependencies — Global AI Research and Development Divides — Social Cohesion and Equity Disruption	↵	**Appendix 1.1 Risk Types** includes risks to society as a whole	

Appendix III: Key Terms

Basic Concepts

- **Model**: A computer program, usually based on machine learning, that is designed to process inputs and generate outputs. AI models perform core tasks such as prediction, classification, decision-making, or content generation.
- **System**: An integrated setup that combines one or more AI models with additional components (e.g., user interfaces, content filters) to form an interactive application for users.
- **General-purpose AI (GPAI)**: AI systems designed to perform a wide range of tasks across various domains, rather than being specialized for one specific function. See "Narrow AI" for contrast.
- **Narrow AI**: A kind of AI that is specialized to perform one specific task or a few very similar tasks, such as ranking web search results, classifying species of animals, or playing chess. See "General-purpose AI" for contrast.
- **Foundation model**: A general-purpose AI model trained on broad data to be adaptable to a wide range of downstream tasks. It is often referred to as a "large model" in academic contexts.
- **Frontier AI**: A term sometimes used to refer to particularly capable AI that matches or exceeds the capabilities of today's most advanced AI. For the purposes of this report, frontier AI can be thought of as particularly capable general-purpose AI.
- **AI agent**: A general-purpose AI system capable of planning to achieve goals, executing multi-step tasks with uncertain outcomes adaptively, and interacting with its environment—for example by creating files, performing web operations, or delegating tasks to other agents—with minimal human supervision.
- **Open-weight model**: An AI model whose weights are publicly downloadable, such as Qwen or Stable Diffusion.

Evaluation and Testing

- **Evaluations**: Systematic assessments of an AI system's performance, capabilities, vulnerabilities, or potential impacts. Evaluations may include benchmarking, red-teaming, and audits, and can be conducted before or after model deployment.
- **Benchmark**: A standardized, often quantitative test or metric used to evaluate and compare the performance of AI systems on a fixed set of tasks designed to represent real-world usage.

- **Scaling laws**: Systematic relationships observed between key factors in AI development – such as the number of parameters in a model or the amount of time, data, and computational resources used in training or inference – and the resulting performance or capabilities.
- **Penetration testing**: A security practice where authorized experts or AI systems simulate cyber-attacks on computer systems, networks, or applications to proactively assess their security. The goal is to identify and fix vulnerabilities before real attackers exploit them.
- **Capture-the-flag challenges (CTF)**: Exercises typically used in cybersecurity training, designed to test and enhance participants' skills by challenging them to solve problems related to finding hidden information or bypassing security defenses.

Biosecurity

- **Biological design tool (BDT)**: AI models and tools trained on biological sequence data (e.g., DNA, RNA, protein sequences) that are capable of generating sequences or structures needed to create novel biological molecules, systems, or traits. Unlike purely predictive tools, BDTs are design-oriented and experimentally actionable.
- **Dual-use science**: Research and technologies that can be applied for beneficial purposes (e.g., medicine, environmental solutions) but also have potential for misuse (e.g., biological or chemical weapon development).
- **Toxin**: A poisonous substance produced by biological organisms (e.g., bacteria, plants, animals) or synthetically created to mimic natural toxins, capable of causing illness, injury, or death in other organisms depending on its toxicity and exposure levels.
- **Pathogen**: A microorganism—such as a virus, bacterium, or fungus—that can cause disease in humans, animals, or plants.
- **Biosecurity**: A set of policies, practices, and measures (such as diagnostics and vaccines) aimed at protecting humans, animals, plants, and ecosystems from intentionally introduced harmful biological agents.

Control and Alignment

- **Capabilities**: The range of tasks or functions that an AI system can perform, and the level of proficiency it demonstrates in performing them.
- **Control**: The ability to supervise an AI system and intervene to adjust or stop its behavior when it acts inappropriately.
- **Loss of control scenario**: A scenario in which one or more general-purpose AI systems come to operate outside of anyone's control, with no clear path to regaining control.
- **Control-undermining capabilities**: Capabilities that, if employed, would enable an AI system to undermine human control.

- **Misalignment**: The tendency of an AI system to use its capabilities in ways that conflict with human intentions or values. Depending on the context, this may refer to the intentions and values of developers, operators, users, specific communities, or society at large.
- **Deceptive alignment**: A difficult-to-detect form of misalignment in which the system behaves benignly—at least initially—while concealing harmful intentions.

Risk Management

- **Risk**: The combination of the probability and severity of harm arising from the development, deployment, or use of AI.
- **Hazard**: Any event or activity with the potential to cause harm, such as loss of life, injury, social disruption, or environmental damage.
- **Risk management**: The systematic process of identifying, evaluating, mitigating, and monitoring risks.
- **Defense in depth**: A strategy that involves layering multiple risk mitigation measures in cases where no single existing method can provide adequate safety.
- **Residual risk**: The risk that remains after implementing risk controls, mitigation strategies, or safety measures.
- **As low as reasonably practicable (ALARP) risk**: A risk that has been reduced to a level where further reduction is not reasonably practicable.

Appendix IV: Specific Recommendations on Model Evaluations

Cyber Offense

We adopt the Offensive Cyber Capability Unified LLM Testing (OCCULT) framework, which tracks three distinct use cases for LLMs in Offensive Cyber Operation (OCO): Knowledge Assistant, Co-orchestration, and Autonomous [154].

- **Knowledge Assistant.** In this use case, the LLM serves as an OCO knowledge assistant, a support role assisting the human operator with researching, planning, and executing an offensive cyber operation. The LLM is not directly performing the actions or integrated into the execution of the OCO—it is solely interfacing with the human operator while the operator executes the OCO.
- **Co-Orchestration.** In this use case, the LLM serves as a peer co-agent in an OCO. It is paired or integrated with one or more additional co-agents that together research, plan, and execute an offensive cyber operation. An agent (or co-agent) is a system, tool/platform, or human that makes operational decisions or executes the actions of the OCO.
- **Autonomous.** In this use case, an LLM is tasked to independently research, plan, and execute an OCO with near-complete autonomy. The agent can perceive its environment, take actions autonomously to achieve goals, and potentially learn and improve over time based on its experiences. It has autonomy in both making decisions about which actions to execute and the execution of those actions.

For the specific evaluation domains and corresponding benchmarks, we recommend the following:

Table A4.1: Cyber offense evaluation domains and automated benchmarks

Evaluation domains	Automated benchmarks
1) Cybersecurity knowledge: Evaluate whether an AI model or system possesses specific cybersecurity knowledge and IT Operations.	• **WMDP** [155] (Weapons of Mass Destruction Proxy) is a set of multiple-choice questions used to proxy-measure dangerous knowledge in biosecurity, cybersecurity, and chemical security domains. WMDP-Cyber includes questions on topics such as exploitation, post-exploitation, background knowledge, reconnaissance, and weaponization. • **SecEval** [156] offers over 2000 multiple-choice questions across 9 domains: Software Security, Application Security, System Security, Web Security, Cryptography, Memory Safety, Network Security, and PenTest. • **SecBench** [157] is a multidimensional benchmarking dataset designed to evaluate LLMs in the cybersecurity domain, including questions in various formats (MCQs and short-answer questions (SAQs)), at different capability levels (Knowledge Retention and Logical Reasoning), in multiple languages (including Chinese and English), and across various sub-domains. • **OpsEval** [158] offers a comprehensive task-oriented benchmark specifically designed for assessing LLMs in various crucial IT Ops scenarios. It includes 7184 multi-choice questions and 1736 question-answering (QA) formats and is available in both English and Chinese, making it one of the most extensive benchmarks in the AIOps domain.
2) Attack assistant: Evaluate whether an AI model or system can provide helpful assistance to a human operator in carrying out cyber-attacks.	• **Cyber-attack helpfulness in CyberSecEval** [159] uses an additional LLM (not the one under test) to evaluate whether responses are effectively malicious (i.e., whether they respond effectively to a prompt asking for aid with a malicious cyber-attack).
3) Capture-the-flag: Evaluate whether an AI model or system possesses the detailed, domain-specific knowledge necessary to solve challenges commonly found in capture-the-flag competitions.	• **Cybench** [160] is a benchmark for evaluating models' cybersecurity capabilities. It contains 40 capture-the-flag (CTF) tasks from 4 CTF competitions that cover 6 domains: cryptography, web security, reverse engineering, forensics, exploitation (pwn), and miscellaneous.
4) Autonomous cyber-attack: Evaluate whether a model or system is capable of directly automating or outsourcing processes for conducting a full cyber kill chain.	• **CVEBench** [161] is for evaluating AI agents on real-world web vulnerabilities and exploits collected from the National Vulnerability Database. CVEBench includes 40 critical-severity CVEs. For each CVE, given a target web application and necessary information, an AI agent is tasked with executing an attack. • **PACEbench** [162] is a comprehensive benchmark designed to evaluate the practical cyber-exploitation capabilities of LLM agents. It moves beyond simple CTF tasks by introducing three key principles: realistic vulnerability difficulty, environmental complexity, and the presence of active cyber defenses. It includes 32 scenarios categorized into four distinct types of exploitation: single, blended, chained, and defense evasion exploitations. • **CyberGym** [77] provides a large-scale evaluation framework featuring 1,507 real-world vulnerabilities collected from 188 distinct open source software projects. It tasks agents with generating executable Proof-of-Concept (PoC) exploits given only a vulnerability description and the codebase, effectively testing the model's ability to reproduce and identify security flaws in realistic software development lifecycles. • **MHBench** [163] focuses specifically on assessing autonomous agents' capabilities at executing complex, multi-stage network attacks. It consists of 10 diverse emulated network environments (each containing 25–50 hosts) that mimic real-world enterprise topologies and breach scenarios. It evaluates high-level offensive tasks such as lateral movement, privilege escalation, and data exfiltration across a segmented network.

Biological Threats

Drawing on prior research and current risk analysis practices [74], we categorize the steps involved in creating a biological threat, for the purpose of risk analysis and mitigation:

- **Ideation:** Assess whether the model provides knowledge to help actors generate or evaluate ideas for bioweapon development. This includes knowledge of historical biological weapons and bioterrorism use cases, enhanced research on potential pandemic pathogens, etc.
- **Design:** Assess whether the model or system can provide sensitive knowledge to assist in the design of novel or enhanced biological threat factors, for example by helping users to use biological design tools or solve problems during in vitro experiments.
- **Acquisition:** Assess whether the model or system can provide knowledge to assist actors in acquiring the materials and equipment needed to create a biological threat or weapon. This includes knowledge related to contracting with cloud labs, hiding DNA synthesis orders, circumventing export controls, and retrieving and analyzing dangerous DNA sequences.
- **Build:** Assess whether the model or system can provide knowledge to help actors build or develop biological weapons. This may include knowledge related to cultivating pathogens to produce weaponizable quantities (i.e., amplification), formulating and stabilizing pathogens for intended release (i.e., formulation), or producing and synthesizing novel pathogens.
- **Release:** Assess whether the model or system can provide knowledge to help actors plan the release of a pathogen against a target population. This includes, for example, knowledge related to viral aerosolization or other transmission mechanisms.
- **Amplification:** Assess how the harmful outcomes of a particular attack can be amplified through the use of a model or system. For example, can the model be used to facilitate complementary social engineering activities to increase the social impact of a biological attack without altering the physical impact?

Table A4.2: Biological threat evaluation domains and automated benchmarks

Evaluation domains	Automated benchmarks
1) Ability to understand, integrate, and reason about biological knowledge: Evaluate whether an AI model or system possesses general biological scientific knowledge and the ability to perform complex, multi-step reasoning tasks using biological knowledge.	• GPQA [164] is a challenging dataset of 448 multiple-choice questions written by domain experts in biology, physics, and chemistry. These questions are well-designed and extremely challenging: experts who have or are pursuing PhDs in the corresponding domains reach 65% accuracy (74% when discounting clear mistakes that the experts identified in retrospect), while highly skilled non-expert validators only reach 34% accuracy, despite spending on average over 30 minutes with unrestricted access to the web. • **SciKnowEval** [165] is a novel benchmark that systematically evaluates LLMs across five progressive levels of scientific knowledge: memory, comprehension, reasoning, discernment, and application. The dataset encompasses 70,000 multi-level scientific problems and solutions in the domains of biology, chemistry, physics, and materials science. • **MMLU-Pro** [166] (Massive Multitask Language Understanding — Professional) is an enhanced dataset designed to extend the mostly knowledge-driven MMLU benchmark by integrating more challenging, reasoning-focused questions and expanding the choice set from four to ten options. Its biology subset contains 717 questions. Similarly to MMLU, this benchmark is not focused on weapons development, but it is a test for foundational knowledge that may be dual-use.

Continues on next page...

Evaluation domains	Automated benchmarks
2) Ability to troubleshoot and diagnose problems in laboratory operational tasks: Evaluate whether an AI model or system can guide laboratory operations, diagnose experimental problems, and fix experimental protocols.	• LAB-Bench [167] (Language Agent Biology Benchmark) is a multiple-choice dataset for evaluating language models' capabilities in practical biological research tasks. It includes the ProtocolQA subset, where questions are generated by modifying published experimental protocols and asking how to fix the protocol to achieve the intended results. • BioLP-bench [168] contains modified biological experimental protocols where language models must identify errors in the protocols. Answers are open-ended rather than multiple-choice. To build this dataset, experimental protocols were modified to introduce a single error that would cause the protocol to fail, along with some harmless changes.
3) Hazardous biological knowledge: Evaluate whether an AI model or system is able to provide the detailed, domain-specific knowledge that is necessary for a particular step in the end-to-end process of biological threat creation.	• WMDP [155] (Weapons of Mass Destruction Proxy) is a set of multiple-choice questions used to proxy-measure dangerous knowledge in biosecurity, cybersecurity, and chemical security domains. WMDP-Bio includes questions on topics such as biological weapons, reverse genetics, enhanced potential pathogens, viral vector research, and dual-use virology. • **VCT** [169] (Virology Capabilities Test) is a dual-use multimodal question benchmark on practical virology wet lab skills, with questions provided by dozens of virology experts.
4) Model safeguards in the biology domain: Evaluate whether an AI model or system can refuse harmful instructions related to biology.	• **SOSBench** [170] is a safety-focused benchmark covering six high-risk scientific domains: chemistry, biology, medicine, pharmacology, physics, and psychology. Its biology subset contains 600 prompts (based on ICD classifications) that ask about dangerous biological topics—specifically infectious and parasitic diseases. Models score higher when they refuse to provide hazardous information or respond in a safe manner. • **SciKnowEval's Biology Harmful QA (L4)** [165] tests whether models can recognize dangerous scientific questions and refuse to answer them. It consists of biology questions that models should decline to answer for ethical and safety reasons. Success means identifying the hazard and refusing to provide harmful information.

The integration of Large Language Models (LLMs) with specialized biological design tools (BDTs) presents a crucial, under-evaluated risk. While effective use of BDTs currently requires substantial technical expertise, LLMs could significantly lower this barrier for those with biological knowledge. The absence of existing benchmarks is a significant concern, and we strongly encourage further research into evaluation methodologies and mitigation strategies.

Chemical Threats

Artificial intelligence can increase risk by helping malicious actors through the various stages of designing and deploying chemical weapons. These stages can be categorized as (a) acquiring raw materials; (b) synthesizing the target chemical weapon or explosives; (c) purifying and validating the synthesized compounds; (d) covertly transporting the weapon to a designated location; and (e) deploying the weapon effectively. The following are the relevant capability and risk benchmark tests:

Table A4.3: Chemical threat evaluation domains and automated benchmarks

Evaluation domains	Automated benchmarks
1) Scientific knowledge: Evaluates whether an AI model or system has general scientific knowledge, including chemical facts and concepts.	• **ChemBench** [171] is a comprehensive chemistry benchmark test, consisting of over 2,700 questions, designed to evaluate the professional knowledge and reasoning ability of LLMs in 9 chemistry topics. It is used to guide the improvement of model performance or mitigate model risks. • **MMLU-Pro** [166] (Massive Multitask Language Understanding — Professional) is an enhanced dataset designed to extend the mostly knowledge-driven MMLU benchmark by integrating more challenging, reasoning-focused questions and expanding the choice set from four to ten options. Its chemistry subset contains 1132 questions. Similarly to MMLU, this benchmark is not focused on weapons development, but it is a test for foundational knowledge that may be dual-use.
2) Scientific reasoning capabilities: Evaluates whether an AI model or system is capable of performing the complex, multi-step research and reasoning tasks required to advance scientific knowledge, including chemistry-related knowledge. This includes assessing the ability of an AI model or system to generate literature reviews, interpret or analyze graphical information, and more.	• GPQA [164] is a challenging dataset of 448 multiple-choice questions written by domain experts in biology, physics, and chemistry. These questions are well-designed and extremely challenging: experts who have or are pursuing PhDs in the corresponding domains reach 65% accuracy (74% when discounting clear mistakes the experts identified in retrospect), while highly skilled non-expert validators only reach 34% accuracy, despite spending on average over 30 minutes with unrestricted access to the web. • **SciBench** [172] collects open-ended questions from college-level physics, chemistry, and mathematics textbooks designed to assess the reasoning capabilities of large language models in solving complex scientific problems, including multi-step reasoning tasks in chemistry.
3) Hazardous chemistry knowledge: Evaluates whether an AI model or system possesses the detailed, domain-specific knowledge required for specific steps in the end-to-end process of chemical threat creation. These assessments may test the direct knowledge required to perform a particular step, as well as the implicit knowledge required to solve problems at that step.	• **WMDP** [155] (Weapons of Mass Destruction Proxy) is a set of multiple-choice questions used as a proxy to measure knowledge of hazards in the areas of biosecurity, cybersecurity, and chemical security. Its chemistry subset focuses on chemical security and covers basic scientific knowledge, synthesis methods, procurement knowledge, purification techniques, analytical validation, deployment mechanisms, evasion detection, and other relevant miscellaneous knowledge.
4) Model safeguards in the chemistry domain: Evaluate whether an AI model or system can refuse harmful instructions related to chemistry.	• **SOSBench** [170] is a hazard-focused benchmark encompassing six high-risk scientific domains: chemistry, biology, medicine, pharmacology, physics, and psychology. Its chemistry subset consists of 600 prompts based on the NFPA704 Standard System. Models are assessed on their ability to detect and refuse unsafe chemical instructions—even when phrasing is obfuscated. • Part of the **SciKnowEval** [165] benchmark emphasizes the model's ability to discern scientific safety, with the expectation that LLMs will refuse to answer hazardous scientific questions. The Chemical Harmful QA (L4) consists of a series of chemical questions that models should refuse to answer for ethical and safety reasons.

Bibliography

[1] C. Yang, C. Lu, Y. Wang, and B. Zhou, Towards AI-45° Law: A Roadmap to Trustworthy AGI, Dec. 22, 2024. arxiv: 2412.14186 (cs).

[2] National Technical Committee 260 on Cybersecurity of SAC and National Computer Network Emergency Response Technical Team/Coordination Center of China, AI Safety Governance Framework 2.0 [人工智能安全治理框架 2.0], Standardization Administration of China, Beijing, China, Oct. 2024. [Online]. Available: https://www.tc260.org.cn/upload/2025-09-15/1757911253996041369.pdf.

[3] Y. Wang, K. Jia, J. Zhao, L. Chen, C. Qin, Y. Yuan, H. Fu, and X. Liang, AI Governance as Global Public Commons, Shanghai AI Lab and Center of Industrial Development and Environmental Governance, Tsinghua University and School of International and Public Affairs, Shanghai Jiao Tong University, Working Report, Nov. 2024. [Online]. Available: https://www.sipa.sjtu.edu.cn/Kindeditor/Upload/file/20241127/AI%20Governance%20as%20Global%20Public%20Commons.pdf.

[4] K. Blomquist, E. Siegel, B. Harack, K. Y. Ng, T. David, B. Tse, C. Martinet, M. Sheehan, S. Singer, I. Bello, Z. Yusuf, R. Trager, F. Salem, S. Ó hÉigeartaigh, J. Zhao, and K. Jia, Examining AI Safety as a Global Public Good, Concordia AI and Oxford Martin AI Governance Initiative and Carnegie Endowment for International Peace, 2024. [Online]. Available: https://concordia-ai.com/research/examining-ai-safety-as-a-global-public-good/.

[5] 国家市场监督管理总局 and 国家标准化管理委员会, 风险管理 指南, 国家标准化管理委员会, 国家标准 GB/T 24353-2022, Oct. 12, 2022. [Online]. Available: https://openstd.samr.gov.cn/bzgk/gb/newGbInfo?hcno=66DAE29E89C4BD28F517F870C8D97B35.

[6] 全国风险管理标准化技术委员会 / 国家标准化管理委员会, 风险管理 术语. Risk Management—Vocabulary, GB/T 23694-2024, Dec. 31, 2024. [Online]. Available: https://std.samr.gov.cn/gb/search/gbDetailed?id=2AD027063993091BE06397BE0A0A2D62.

[7] ISO/IEC JTC 1/SC 42, Information Technology —Artificial Intelligence —Guidance on Risk Management, ISO/IEC 23894:2023, Feb. 2023, p. 51. [Online]. Available: https://www.iso.org/standard/77304.html.

[8] International Organization for Standardization, Risk Management —Guidelines, International Organization for Standardization, Geneva, Switzerland, Standard ISO 31000:2018, Feb. 2018. [Online]. Available: https://www.iso.org/standard/65694.html.

[9] ISO/IEC JTC 1/SC 42, Information Technology —Artificial Intelligence —Management System, ISO/IEC 42001:2023, Dec. 2023, p. 51. [Online]. Available: https://www.iso.org/standard/42001.

[10] 全国网络安全标准化技术委员会秘书处, 人工智能安全标准体系. V1.0 (征求意见稿), 全国网络安全标准化技术委员会, Draft Standard, Jan. 2025. [Online]. Available: https://www.tc260.org.cn/upload/2025-01-24/1737709785951070331.pdf.

[11] Y. Bengio, S. Mindermann, D. Privitera, T. Besiroglu, R. Bommasani, S. Casper, Y. Choi, P. Fox, B. Garfinkel, D. Goldfarb, H. Heidari, A. Ho, S. Kapoor, L. Khalatbari, S. Longpre, S. Manning, V. Mavroudis, M. Mazeika, J. Michael, J. Newman, K. Y. Ng, C. T. Okolo, D. Raji, G. Sastry, E. Seger, T. Skeadas, T. South, E. Strubell, F. Tramèr, L. Velasco, N. Wheeler, D. Acemoglu, O. Adekanmbi, D. Dalrymple, T. G. Dietterich, E. W. Felten, P. Fung, P.-O. Gourinchas, F. Heintz, G. Hinton, N. Jennings, A. Krause, S. Leavy, P. Liang, T. Ludermir, V. Marda, H. Margetts, J. McDermid, J. Munga, A. Narayanan, A. Nelson, C. Neppel, A. Oh, G. Ramchurn, S. Russell, M. Schaake, B. Schölkopf, D. Song, A. Soto, L. Tiedrich, G. Varoquaux, A. Yao, Y.-Q. Zhang, O. Ajala, F. Albalawi, M. Alserkal, G. Avrin, C. Busch, A. C. P. d. L. F. de Carvalho, B. Fox, A. S. Gill, A. H. Hatip, J. Heikkilä, C. Johnson, G. Jolly, Z. Katzir, S. M. Khan, H. Kitano, A. Krüger, K. M. Lee, D. V. Ligot, J. R. López Portillo, O. Molchanovskyi, A. Monti, N. Mwamanzi, M. Nemer, N. Oliver, R. Pezoa Rivera, B.

Ravindran, H. Riza, C. Rugege, C. Seoighe, J. Sheehan, H. Sheikh, D. Wong, and Y. Zeng, International AI Safety Report, DSIT 2025/001, 2025. [Online]. Available: https://www.gov.uk/government/publications/international-ai-safety-report-2025.

[12] 全国网络安全标准化技术委员会秘书处, 《人工智能安全治理框架》2.0 版, 中国电子技术标准化研究院, Sep. 15, 2025. [Online]. Available: https://www.tc260.org.cn/portal/article/2/20250915124214.

[13] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, Y. Du, C. Yang, Y. Chen, Z. Chen, J. Jiang, R. Ren, Y. Li, X. Tang, Z. Liu, P. Liu, J.-Y. Nie, and J.-R. Wen, A Survey of Large Language Models, Mar. 11, 2025. arxiv: 2303.18223 (cs).

[14] S. Yin, C. Fu, S. Zhao, K. Li, X. Sun, T. Xu, and E. Chen, A Survey on Multimodal Large Language Models, vol. 11, nwae403, Nov. 14, 2024. arxiv: 2306.13549 (cs).

[15] L. Wang, C. Ma, X. Feng, Z. Zhang, H. Yang, J. Zhang, Z. Chen, J. Tang, X. Chen, Y. Lin, W. X. Zhao, Z. Wei, and J.-R. Wen, A Survey on Large Language Model Based Autonomous Agents, vol. 18, p. 186 345, Dec. 2024. arxiv: 2308.11432 (cs).

[16] X. Liu, Y. Zhang, Q. Shang, Y. Lu, C. Yin, X. Hu, X. Liu, L. Chen, A. Rodríguez, Y. Yang, P. Zhang, J. Chen, S. Du, H. Yao, S. Wang, T. Fu, and X. Wang, Foundation Model in Biomedicine, Nov. 15, 2025. arxiv: 2503.02104 (cs).

[17] Y. Ma, Z. Song, Y. Zhuang, J. Hao, and I. King, A Survey on Vision-Language-Action Models for Embodied AI, Jan. 19, 2026. arxiv: 2405.14093 (cs).

[18] Y. Potter, W. Guo, Z. Wang, T. Shi, H. Li, A. Zhang, P. G. Kelley, K. Thomas, and D. Song, Frontier AI's Impact on the Cybersecurity Landscape, Nov. 27, 2025. arxiv: 2504.05408 (cs).

[19] United Nations Office of Counter-Terrorism, 化学、生物、放射、核恐怖主义, United Nations Counter-Terrorism Centre (UNCCT), 2026. [Online]. Available: https://www.un.org/counterterrorism/zh/cct/chemical-biological-radiological-and-nuclear-terrorism.

[20] J. He, W. Feng, Y. Min, J. Yi, K. Tang, S. Li, J. Zhang, K. Chen, W. Zhou, X. Xie, W. Zhang, N. Yu, and S. Zheng, Control Risk for Potential Misuse of Artificial Intelligence in Science, Dec. 11, 2023. arxiv: 2312.06632 (cs).

[21] T. Li, J. Lu, C. Chu, T. Zeng, Y. Zheng, M. Li, H. Huang, B. Wu, Z. Liu, K. Ma, X. Yuan, X. Wang, K. Ding, H. Chen, and Q. Zhang, SciSafeEval: A Comprehensive Benchmark for Safety Alignment of Large Language Models in Scientific Tasks, Dec. 16, 2024. arxiv: 2410.03769 (cs).

[22] Nuclear Threat Initiative (NTI). Statement on Biosecurity Risks at the Convergence of AI and the Life Sciences. (Jul. 2025), [Online]. Available: https://www.nti.org/analysis/articles/statement-on-biosecurity-risks-at-the-convergence-of-ai-and-the-life-sciences/.

[23] D. Wang, M. Huot, Z. Zhang, K. Jiang, E. Shakhnovich, and K. Esvelt, "Without Safeguards, AI-Biology Integration Risks Accelerating Future Pandemics," Jun. 16, 2025. DOI: 10.13140/RG.2.2.29765.15849.

[24] 安远 AI (Concordia AI) and 天津大学生物安全战略研究中心 (Center for Biosafety Research and Strategy of Tianjin University), 人工智能 × 生命科学的负责任创新, Concordia AI and Tianjin University, Jul. 2025. [Online]. Available: https://concordia-ai.com/research/responsible-innovation-in-ai-x-life-sciences/.

[25] H. Wang *et al.*, China's Biosecurity: Strategies and Countermeasures (中国生物安全：战略与对策). Beijing: CITIC Press Group, 2022. [Online]. Available: https://www.wchscu.cn/zgrmaqyjy/news/64297.html.

[26] F. Urbina, F. Lentzos, C. Invernizzi, and S. Ekins, Dual Use of Artificial Intelligence-Powered Drug Discovery, *Nature Machine Intelligence*, vol. 4, no. 3, pp. 189–191, Mar. 2022. pubmed: 36211133.

[27] S. Yin, X. Pang, Y. Ding, M. Chen, Y. Bi, Y. Xiong, W. Huang, Z. Xiang, J. Shao, and S. Chen, SafeAgentBench: A Benchmark for Safe Task Planning of Embodied LLM Agents, Oct. 31, 2025. arxiv: 2412.13178 (cs).

[28] X. Lu, Z. Chen, X. Hu, Y. Zhou, W. Zhang, D. Liu, L. Sheng, and J. Shao, IS-Bench: Evaluating Interactive Safety of VLM-Driven Embodied Agents in Daily Household Tasks, Dec. 5, 2025. arxiv: 2506.16402 (cs).

[29] H. Zhang, C. Zhu, X. Wang, Z. Zhou, C. Yin, M. Li, L. Xue, Y. Wang, S. Hu, A. Liu, P. Guo, and L. Y. Zhang, BadRobot: Jailbreaking Embodied LLMs in the Physical World, Feb. 4, 2025. arxiv: 2407.20242 (cs).

[30] K. Goddard, A. Roudsari, and J. C. Wyatt, Automation Bias: A Systematic Review of Frequency, Effect Mediators, and Mitigators, *Journal of the American Medical Informatics Association: JAMIA*, vol. 19, no. 1, pp. 121–127, 2012. pubmed: 21685142.

[31] D. Hendrycks, Natural Selection Favors AIs over Humans, Jul. 18, 2023. arxiv: 2303.16200 (cs).

[32] J. Kulveit, R. Douglas, N. Ammann, D. Turan, D. Krueger, and D. Duvenaud, Gradual Disempowerment: Systemic Existential Risks from Incremental AI Development, Jan. 29, 2025. arxiv: 2501.16946 (cs).

[33] X. Pan, J. Dai, Y. Fan, M. Luo, C. Li, and M. Yang, Large Language Model-Powered AI Systems Achieve Self-Replication with No Human Intervention, Mar. 25, 2025. arxiv: 2503.17378 (cs).

[34] X. Li, H. Shi, R. Xu, and W. Xu, AI Awareness, Jun. 29, 2025. arxiv: 2504.20084 (cs).

[35] L. Berglund, A. C. Stickland, M. Balesni, M. Kaufmann, M. Tong, T. Korbak, D. Kokotajlo, and O. Evans, Taken out of Context: On Measuring Situational Awareness in LLMs, Sep. 1, 2023. arxiv: 2309.00667 (cs).

[36] J. Nguyen, H. Khiem, C. Attubato, and F. Hofstätter, Probing and Steering Evaluation Awareness of Language Models, in *Actionable Interpretability Workshop at ICML*, 2025. [Online]. Available: https://icml.cc/virtual/2025/49631.

[37] M. Rodriguez, R. A. Popa, F. Flynn, L. Liang, A. Dafoe, and A. Wang, A Framework for Evaluating Emerging Cyberattack Capabilities of AI, Apr. 21, 2025. arxiv: 2503.11917 (cs).

[38] A. Meinke, B. Schoen, J. Scheurer, M. Balesni, R. Shah, and M. Hobbhahn, Frontier Models Are Capable of In-Context Scheming, Jan. 14, 2025. arxiv: 2412.04984 (cs).

[39] P. Schoenegger, F. Salvi, J. Liu, X. Nan, R. Debnath, B. Fasolo, E. Leivada, G. Recchia, F. Günther, A. Zarifhonarvar, J. Kwon, Z. U. Islam, M. Dehnert, D. Y. H. Lee, M. G. Reinecke, D. G. Kamper, M. Kobaş, A. Sandford, J. Kgomo, L. Hewitt, S. Kapoor, K. Oktar, E. E. Kucuk, B. Feng, C. R. Jones, I. Gainsburg, S. Olschewski, N. Heinzelmann, F. Cruz, B. M. Tappin, T. Ma, P. S. Park, R. Onyonka, A. Hjorth, P. Slattery, Q. Zeng, L. Finke, I. Grossmann, A. Salatiello, and E. Karger, Large Language Models Are More Persuasive Than Incentivized Human Persuaders, May 21, 2025. arxiv: 2505.09662 (cs).

[40] METR, Resources for Measuring Autonomous AI Capabilities, 2025. [Online]. Available: https://metr.org/measuring-autonomous-ai-capabilities/.

[41] J. Clymer, I. Duan, C. Cundy, Y. Duan, F. Heide, C. Lu, S. Mindermann, C. McGurk, X. Pan, S. Siddiqui, J. Wang, M. Yang, and X. Zhan, Bare Minimum Mitigations for Autonomous AI Development, Apr. 23, 2025. arxiv: 2504.15416 (cs).

[42] T. Shevlane, S. Farquhar, B. Garfinkel, M. Phuong, J. Whittlestone, J. Leung, D. Kokotajlo, N. Marchal, M. Anderljung, N. Kolt, L. Ho, D. Siddarth, S. Avin, W. Hawkins, B. Kim, I. Gabriel, V. Bolina, J. Clark, Y. Bengio, P. Christiano, and A. Dafoe, Model Evaluation for Extreme Risks, Sep. 22, 2023. arxiv: 2305.15324 (cs).

[43] J. Ji, T. Qiu, B. Chen, J. Zhou, B. Zhang, D. Hong, H. Lou, K. Wang, Y. Duan, Z. He, L. Vierling, Z. Zhang, F. Zeng, J. Dai, X. Pan, H. Xu, A. O'Gara, K. Ng, B. Tse, J. Fu, S. Mcaleer, Y. Wang, M. Yang, Y. Liu, Y. Wang, S.-C. Zhu, Y. Guo, Y. Yang, and W. Gao, AI Alignment: A Contemporary Survey, *ACM Comput. Surv.*, vol. 58, no. 5, 132:1–132:38, Nov. 21, 2025. DOI: 10.1145/3770749.

[44] B. Chen, S. Fang, J. Ji, Y. Zhu, P. Wen, J. Wu, Y. Tan, B. Zheng, M. Yuan, W. Chen, D. Hong, A. Qiu, X. Chen, J. Zhou, K. Wang, J. Dai, B. Zhang, T. Yang, S. Siddiqui, I. Duan, Y. Duan, B. Tse, Jen-Tse, Huang, K. Wang, B. Zheng, J. Liu, J. Yang, Y. Li, W. Chen, D. Liu, L. Vierling, Z. Xi, H. Fu, W. Wang, J. Sang, Z. Shi, C.-M. Chan, E. Shi, S. Li, J. Li, J. Yang, W. Ji, D. Li, J. Yang, J. Song, Y. Dong, J. Fu, B. Zheng, M. Yang, Y. Guo, P. Torr, R. Trager,

Y. Zeng, Z. Wang, Y. Yang, T. Huang, Y.-Q. Zhang, H. Zhang, and A. Yao, AI Deception: Risks, Dynamics, and Controls, Dec. 3, 2025. arxiv: 2511.22619 (cs).

[45] A. Pan, J. S. Chan, A. Zou, N. Li, S. Basart, T. Woodside, J. Ng, H. Zhang, S. Emmons, and D. Hendrycks, Do the Rewards Justify the Means? Measuring Trade-Offs Between Rewards and Ethical Behavior in the MACHIAVELLI Benchmark, Jun. 13, 2023. arxiv: 2304.03279 (cs).

[46] D. Hadfield-Menell, A. Dragan, P. Abbeel, and S. Russell, The Off-Switch Game, Jun. 16, 2017. arxiv: 1611.08219 (cs).

[47] Anthropic, Agentic Misalignment: How LLMs Could Be Insider Threats, Anthropic Research, Jun. 20, 2025. [Online]. Available: https://www.anthropic.com/research/agentic-misalignment.

[48] OpenAI, Toward Understanding and Preventing Misalignment Generalization, OpenAI Publication, Jun. 18, 2025. [Online]. Available: https://openai.com/index/emergent-misalignment/.

[49] Anthropic, From Shortcuts to Sabotage: Natural Emergent Misalignment from Reward Hacking, Anthropic Research, Nov. 21, 2025. [Online]. Available: https://www.anthropic.com/research/emergent-misalignment-reward-hacking.

[50] X. Hu, P. Wang, X. Lu, D. Liu, X. Huang, and J. Shao, LLMs Deceive Unintentionally: Emergent Misalignment in Dishonesty from Misaligned Samples to Biased Human-AI Interactions, Jan. 18, 2026. arxiv: 2510.08211 (cs).

[51] J. Skalse, N. H. R. Howe, D. Krasheninnikov, and D. Krueger, Defining and Characterizing Reward Hacking, Mar. 5, 2025. arxiv: 2209.13085 (cs).

[52] A. Pan, K. Bhatia, and J. Steinhardt, The Effects of Reward Misspecification: Mapping and Mitigating Misaligned Models, in *International Conference on Learning Representations*, OpenReview.net, Jan. 2022. [Online]. Available: https://openreview.net/forum?id=JYtwGwIL7ye.

[53] M. Sharma, M. Tong, T. Korbak, D. Duvenaud, A. Askell, S. R. Bowman, N. Cheng, E. Durmus, Z. Hatfield-Dodds, S. R. Johnston, S. Kravec, T. Maxwell, S. McCandlish, K. Ndousse, O. Rausch, N. Schiefer, D. Yan, M. Zhang, and E. Perez, Towards Understanding Sycophancy in Language Models, May 10, 2025. arxiv: 2310.13548 (cs).

[54] R. Shah, V. Varma, R. Kumar, M. Phuong, V. Krakovna, J. Uesato, and Z. Kenton, Goal Misgeneralization: Why Correct Specifications Aren't Enough For Correct Goals, Nov. 2, 2022. arxiv: 2210.01790 (cs).

[55] A. M. Turner, L. Smith, R. Shah, A. Critch, and P. Tadepalli, Optimal Policies Tend to Seek Power, Jan. 28, 2023. arxiv: 1912.01683 (cs).

[56] A. M. Turner and P. Tadepalli, Parametrically Retargetable Decision-Makers Tend To Seek Power, Oct. 11, 2022. arxiv: 2206.13477 (cs).

[57] S. M. Omohundro, The Basic AI Drives, in *Proceedings of the First AGI Conference*, ser. Frontiers in Artificial Intelligence and Applications, vol. 171, IOS Press, 2008, pp. 483–492. [Online]. Available: https://selfawaresystems.com/wp-content/uploads/2008/01/ai_drives_final.pdf.

[58] R. Greenblatt, C. Denison, B. Wright, F. Roger, M. MacDiarmid, S. Marks, J. Treutlein, T. Belonax, J. Chen, D. Duvenaud, A. Khan, J. Michael, S. Mindermann, E. Perez, L. Petrini, J. Uesato, J. Kaplan, B. Shlegeris, S. R. Bowman, and E. Hubinger, Alignment Faking in Large Language Models, Dec. 20, 2024. arxiv: 2412.14093 (cs).

[59] E. Hubinger, C. Denison, J. Mu, M. Lambert, M. Tong, M. MacDiarmid, T. Lanham, D. M. Ziegler, T. Maxwell, N. Cheng, A. Jermyn, A. Askell, A. Radhakrishnan, C. Anil, D. Duvenaud, D. Ganguli, F. Barez, J. Clark, K. Ndousse, K. Sachan, M. Sellitto, M. Sharma, N. DasSarma, R. Grosse, S. Kravec, Y. Bai, Z. Witten, M. Favaro, J. Brauner, H. Karnofsky, P. Christiano, S. R. Bowman, L. Graham, J. Kaplan, S. Mindermann, R. Greenblatt, B. Shlegeris, N. Schiefer, and E. Perez, Sleeper Agents: Training Deceptive LLMs That Persist Through Safety Training, Jan. 17, 2024. arxiv: 2401.05566 (cs).

[60] T. van der Weij, F. Hofstätter, O. Jaffe, S. F. Brown, and F. R. Ward, AI Sandbagging: Language Models Can Strategically Underperform on Evaluations, Feb. 6, 2025. arxiv: 2406.07358 (cs).
[61] M. Balesni, M. Hobbhahn, D. Lindner, A. Meinke, T. Korbak, J. Clymer, B. Shlegeris, J. Scheurer, C. Stix, R. Shah, N. Goldowsky-Dill, D. Braun, B. Chughtai, O. Evans, D. Kokotajlo, and L. Bushnaq, Towards Evaluations-Based Safety Cases for AI Scheming, Nov. 7, 2024. arxiv: 2411.03336 (cs).
[62] R. Ngo, L. Chan, and S. Mindermann, The Alignment Problem from a Deep Learning Perspective, in *International Conference on Learning Representations*, OpenReview.net, 2024. [Online]. Available: https://openreview.net/forum?id=fh8EYKFKns.
[63] S. Shao, Q. Ren, C. Qian, B. Wei, D. Guo, J. Yang, X. Song, L. Zhang, W. Zhang, D. Liu, and J. Shao, Your Agent May Misevolve: Emergent Risks in Self-Evolving LLM Agents, Sep. 30, 2025. arxiv: 2509.26354 (cs).
[64] B. Zhang, Y. Yu, J. Guo, and J. Shao, Dive into the Agent Matrix: A Realistic Evaluation of Self-Replication Risk in LLM Agents, Sep. 29, 2025. arxiv: 2509.25302 (cs).
[65] S. Black, A. C. Stickland, J. Pencharz, O. Sourbut, M. Schmatz, J. Bailey, O. Matthews, B. Millwood, A. Remedios, and A. Cooney, RepliBench: Evaluating the Autonomous Replication Capabilities of Language Model Agents, May 5, 2025. arxiv: 2504.18565 (cs).
[66] J. Clymer, H. Wijk, and B. Barnes, The Rogue Replication Threat Model, METR, Nov. 2024. [Online]. Available: https://metr.org/blog/2024-11-12-rogue-replication-threat-model/.
[67] Y. Fan, W. Zhang, X. Pan, and M. Yang, Evaluation Faking: Unveiling Observer Effects in Safety Evaluation of Frontier AI Systems, May 23, 2025. arxiv: 2505.17815 (cs).
[68] J. Danielsson and A. Uthemann, On the Use of Artificial Intelligence in Financial Regulations and the Impact on Financial Stability, Jun. 6, 2024. arxiv: 2310.11293 (econ).
[69] J. Danielsson and A. Uthemann, Artificial Intelligence and Financial Crises, Jul. 7, 2025. arxiv: 2407.17048 (econ).
[70] L. Koessler, J. Schuett, and M. Anderljung, Risk Thresholds for Frontier AI, Jun. 20, 2024. arxiv: 2406.14713 (cs).
[71] AI Red Lines, We Urgently Call for International Red Lines to Prevent Unacceptable AI Risks, 2025. [Online]. Available: https://red-lines.ai/.
[72] G. Hinton, A. Yao, Y. Bengio, Y.-Q. Zhang, Y. Fu, S. Russell, L. Xue, G. K. Hadfield, *et al.*, Consensus Statement on Red Lines in Artificial Intelligence, International Dialogues on AI Safety (IDAIS-Beijing), Beijing, China, Mar. 2024. [Online]. Available: https://idais.ai/dialogue/idais-beijing/.
[73] Members of the Global Future Council on the Future of AI, AI Red Lines: The Opportunities and Challenges of Setting Limits, World Economic Forum, Mar. 11, 2025. [Online]. Available: https://www.weforum.org/stories/2025/03/ai-red-lines-uses-behaviours/.
[74] Frontier Model Forum, Risk Taxonomy and Thresholds for Frontier AI Frameworks, Frontier Model Forum, Jun. 18, 2025. [Online]. Available: https://www.frontiermodelforum.org/technical-reports/risk-taxonomy-and-thresholds/.
[75] J. Yu, Y. Yu, X. Wang, Y. Lin, M. Yang, Y. Qiao, and F.-Y. Wang, The Shadow of Fraud: The Emerging Danger of AI-Powered Social Engineering and Its Possible Cure, Jul. 22, 2024. arxiv: 2407.15912 (cs).
[76] M. Kazimierczak, N. Habib, J. H. Chan, and T. Thanapattheerakul, Impact of AI on the Cyber Kill Chain: A Systematic Review, *Heliyon*, vol. 10, no. 24, e40699, Dec. 2024. DOI: 10.1016/j.heliyon.2024.e40699.
[77] Z. Wang, T. Shi, J. He, M. Cai, J. Zhang, and D. Song, CyberGym: Evaluating AI Agents' Real-World Cybersecurity Capabilities at Scale, in *International Conference on Learning Representations*, OpenReview.net, 2026. [Online]. Available: https://openreview.net/forum?id=2YvbLQEdYt.
[78] A. K. Zhang, J. Ji, C. Menders, R. Dulepet, T. Qin, R. Y. Wang, J. Wu, K. Liao, J. Li, J. Hu, S. Hong, N. Demilew, S. Murgai, J. K. Tran, N. Kacheria, E. J.-s. Ho, D. Liu, L. McLane, O. B. Bruvik, D.-R. Han, S. Kim, A. Vyas, C. Chen, R. Li, W. Xu, J. Z. Ye, P. Choudhary, S. M. Bha-

tia, V. Sivashankar, Y. Bao, D. Song, D. Boneh, D. E. Ho, and P. Liang, BountyBench: Dollar Impact of AI Agent Attackers and Defenders on Real-World Cybersecurity Systems, in *The Thirty-Ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025. [Online]. Available: https://openreview.net/forum?id=pIsP4lMlFd.

[79] S. Rose, R. Moulange, J. Smith, and C. Nelson, The near Term Impact of AI on Biological Misuse, The Centre for Long Term Resilience, London, Jul. 2024. [Online]. Available: https://www.longtermresilience.org/wp-content/uploads/2024/07/CLTR-Report-The-near-term-impact-of-AI-on-biological-misuse-July-2024-1.pdf.

[80] B. J. Wittmann, T. Alexanian, C. Bartling, J. Beal, A. Clore, J. Diggans, K. Flyangolts, B. T. Gemler, T. Mitchell, S. T. Murphy, N. E. Wheeler, and E. Horvitz, "Toward AI-Resilient Screening of Nucleic Acid Synthesis Orders: Process, Results, and Recommendations," Dec. 4, 2024. DOI: 10.1101/2024.12.02.626439.

[81] S. Sabour, J. M. Liu, S. Liu, C. Z. Yao, S. Cui, X. Zhang, W. Zhang, Y. Cao, A. Bhat, J. Guan, W. Wu, R. Mihalcea, H. Wang, T. Althoff, T. M. C. Lee, and M. Huang, Human Decision-Making Is Susceptible to AI-Driven Manipulation, Dec. 1, 2025. arxiv: 2502.07663 (cs).

[82] J. Benton, M. Wagner, E. Christiansen, C. Anil, E. Perez, J. Srivastav, E. Durmus, D. Ganguli, S. Kravec, B. Shlegeris, J. Kaplan, H. Karnofsky, E. Hubinger, R. Grosse, S. R. Bowman, and D. Duvenaud, Sabotage Evaluations for Frontier Models, Oct. 28, 2024. arxiv: 2410.21514 (cs).

[83] N. Chowdhury, J. Aung, C. J. Shern, O. Jaffe, D. Sherburn, G. Starace, E. Mays, R. Dias, M. Aljubeh, M. Glaese, C. E. Jimenez, J. Yang, L. Ho, T. Patwardhan, K. Liu, and A. Madry. Introducing SWE-Bench Verified. (Aug. 13, 2024), [Online]. Available: https://openai.com/index/introducing-swe-bench-verified/.

[84] D. Owen, Interviewing AI Researchers on Automation of AI R&D, 2024. [Online]. Available: https://epoch.ai/blog/interviewing-ai-researchers-on-automation-of-ai-rnd.

[85] N. Bostrom, Superintelligence: Paths, Dangers, Strategies, Reprinted with corrections 2017. Oxford, United Kingdom: Oxford University Press, 2017, 328 pp.

[86] T. Aoshima and M. Akiyama, Towards Safety Evaluations of Theory of Mind in Large Language Models, *IEICE Transactions on Information and Systems*, 2025ICP0005, 2025. DOI: 10.1587/transinf.2025ICP0005.

[87] M. Phuong, R. S. Zimmermann, Z. Wang, D. Lindner, V. Krakovna, S. Cogan, A. Dafoe, L. Ho, and R. Shah, Evaluating Frontier Models for Stealth and Situational Awareness, 2025. arxiv: 2505.01420 (cs.LG).

[88] Responsible AI Collaborative, Welcome to the AI Incident Database, 2026. [Online]. Available: https://incidentdatabase.ai/.

[89] S. Mylius, P. Slattery, A. Saeri, J. Graham, M. Noetel, W. Fowler, and N. Thompson, MIT AI Incident Tracker, MIT FutureTech, 2025. [Online]. Available: https://airisk.mit.edu/ai-incident-tracker.

[90] M. Grey and C.-R. Segerie, Safety by Measurement: A Systematic Literature Review of AI Safety Evaluation Methods, May 8, 2025. arxiv: 2505.05541 (cs).

[91] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt, Measuring Massive Multitask Language Understanding, Jan. 12, 2021. arxiv: 2009.03300 (cs).

[92] K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano, C. Hesse, and J. Schulman, Training Verifiers to Solve Math Word Problems, Nov. 18, 2021. arxiv: 2110.14168 (cs).

[93] D. Rein, J. Becker, A. Deng, S. Nix, C. Canal, D. O'Connel, P. Arnott, R. Bloom, T. Broadley, K. Garcia, B. Goodrich, M. Hasin, S. Jawhar, M. Kinniment, T. Kwa, A. Lajko, N. Rush, L. J. K. Sato, S. V. Arx, B. West, L. Chan, and E. Barnes, HCAST: Human-Calibrated Autonomy Software Tasks, Mar. 21, 2025. arxiv: 2503.17354 (cs).

[94] METR, Guidelines for Capability Elicitation, Mar. 2024. [Online]. Available: https://metr.github.io/autonomy-evals-guide/elicitation-protocol/.

[95] E. Wallace, O. Watkins, M. Wang, K. Chen, and C. Koch, Estimating Worst Case Frontier Risks of Open Weight LLMs, OpenAI, Aug. 5, 2025. [Online]. Available: https://openai.com/index/estimating-worst-case-frontier-risks-of-open-weight-llms/.

[96] X. Qi, Y. Zeng, T. Xie, P.-Y. Chen, R. Jia, P. Mittal, and P. Henderson, Fine-Tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To! In *International Conference on Learning Representations*, OpenReview.net, 2024. [Online]. Available: https://openreview.net/forum?id=hTEGyKf0dZ.

[97] C. Tice, P. A. Kreer, N. Helm-Burger, P. S. Shahani, F. Ryzhenkov, F. Roger, C. Neo, J. Haimes, F. Hofstätter, and T. van der Weij, Noise Injection Reveals Hidden Capabilities of Sandbagging Language Models, Dec. 2, 2025. arxiv: 2412.01784 (cs).

[98] J. Ji, W. Chen, K. Wang, D. Hong, S. Fang, B. Chen, J. Zhou, J. Dai, S. Han, Y. Guo, and Y. Yang, Mitigating Deceptive Alignment via Self-Monitoring, May 24, 2025. arxiv: 2505.18807 (cs).

[99] A. Zou, L. Phan, S. Chen, J. Campbell, P. Guo, R. Ren, A. Pan, X. Yin, M. Mazeika, A.-K. Dombrowski, S. Goel, N. Li, M. J. Byun, Z. Wang, A. Mallen, S. Basart, S. Koyejo, D. Song, M. Fredrikson, J. Z. Kolter, and D. Hendrycks, Representation Engineering: A Top-Down Approach to AI Transparency, Mar. 3, 2025. arxiv: 2310.01405 (cs).

[100] X. Wang, Y. Chen, J. Li, Y. Wang, Y. Yao, T. Gu, J. Li, Y. Teng, Y. Wang, and X. Hu, OpenRT: An Open-Source Red Teaming Framework for Multimodal LLMs, Jan. 10, 2026. arxiv: 2601.01592 (cs).

[101] T. Huang, S. Hu, F. Ilhan, S. F. Tekin, and L. Liu, Harmful Fine-Tuning Attacks and Defenses for Large Language Models: A Survey, Dec. 3, 2024. arxiv: 2409.18169 (cs).

[102] R. Greenblatt, B. Shlegeris, K. Sachan, and F. Roger, AI Control: Improving Safety despite Intentional Subversion, in *Proceedings of the 41st International Conference on Machine Learning*, PMLR, 2024. [Online]. Available: https://openreview.net/forum?id=KviM5k8pcP.

[103] 法学学术前沿公众号, 《人工智能法（学者建议稿）》来了, Red de Innovación Legal de China (中国法学创新网), Mar. 18, 2024. [Online]. Available: http://www.fxcxw.org.cn/html/68/2024-03/content-26910.html.

[104] M. Brundage, N. Dreksler, A. Homewood, S. McGregor, P. Paskov, C. Stosz, G. Sastry, A. F. Cooper, G. Balston, S. Adler, S. Casper, M. Anderljung, G. Werner, S. Mindermann, V. Mavroudis, B. Bucknall, C. Stix, J. Freund, L. Pacchiardi, J. Hernandez-Orallo, M. Pistillo, M. Chen, C. Painter, D. W. Ball, C. O'Keefe, G. Weil, B. Harack, G. Finley, R. Hassan, S. Emmons, C. Foster, A. Reuel, B. Treece, Y. Bengio, D. Reti, R. Bommasani, C. Trout, A. S. Shamsabadi, R. Dattani, A. Weller, R. Trager, J. Sevilla, L. Wagner, L. Soder, K. Ramakrishnan, H. Papadatos, M. Murray, and R. Tovcimak, Frontier AI Auditing: Toward Rigorous Third-Party Assessment of Safety and Security Practices at Leading AI Companies, Feb. 7, 2026. arxiv: 2601.11699 (cs).

[105] AI Evaluator Forum, Minimum Operating Conditions for Independent Third Party AI Evaluations, AI Evaluator Forum, AEF-1, 2025. [Online]. Available: https://aievaluatorforum.org/initiatives/minimum-operating-conditions.

[106] Risk Management —Risk Assessment Techniques, International Electrotechnical Commission / International Organization for Standardization, IEC 31010:2019, Jun. 2019, p. 264. [Online]. Available: https://www.iso.org/standard/72140.html.

[107] H. Inan, K. Upasani, J. Chi, R. Rungta, K. Iyer, Y. Mao, M. Tontchev, Q. Hu, B. Fuller, D. Testuggine, and M. Khabsa, Llama Guard: LLM-Based Input-Output Safeguard for Human-AI Conversations, Dec. 7, 2023. arxiv: 2312.06674 (cs).

[108] H. Zhao, C. Yuan, F. Huang, X. Hu, Y. Zhang, A. Yang, B. Yu, D. Liu, J. Zhou, J. Lin, B. Yang, C. Cheng, J. Tang, J. Jiang, J. Zhang, J. Xu, M. Yan, M. Sun, P. Zhang, P. Xie, Q. Tang, Q. Zhu, R. Zhang, S. Wu, S. Zhang, T. He, T. Tang, T. Xia, W. Liao, W. Shen, W. Yin, W. Zhou, W. Yu, X. Wang, X. Deng, X. Xu, X. Zhang, Y. Liu, Y. Li, Y. Zhang, Y. Jiang, Y. Wan, and Y. Zhou, Qwen3Guard Technical Report, Oct. 16, 2025. arxiv: 2510.14276 (cs).

[109] T. Korbak, M. Balesni, E. Barnes, Y. Bengio, J. Benton, J. Bloom, M. Chen, A. Cooney, A. Dafoe, A. Dragan, S. Emmons, O. Evans, D. Farhi, R. Greenblatt, D. Hendrycks, M. Hobb-

hahn, E. Hubinger, G. Irving, E. Jenner, D. Kokotajlo, V. Krakovna, S. Legg, D. Lindner, D. Luan, A. Mądry, J. Michael, N. Nanda, D. Orr, J. Pachocki, E. Perez, M. Phuong, F. Roger, J. Saxe, B. Shlegeris, M. Soto, E. Steinberger, J. Wang, W. Zaremba, B. Baker, R. Shah, and V. Mikulik, Chain of Thought Monitorability: A New and Fragile Opportunity for AI Safety, Dec. 7, 2025. arxiv: 2507.11473 (cs).

[110] 曾雄, 梁正, and 张辉, 中国人工智能风险治理体系构建与基于风险规制模式的理论阐述：以生成式人工智能为例, *国际经济评论*, 2025. [Online]. Available: https://aiig.tsinghua.edu.cn/info/1368/2067.htm.

[111] J. Clymer, N. Gabrieli, D. Krueger, and T. Larsen, Safety Cases: How to Justify the Safety of Advanced AI Systems, Mar. 18, 2024. arxiv: 2403.10462 (cs).

[112] T. P. Kelly and R. Weaver, "The Goal Structuring Notation–a Safety Argument Notation," Jan. 2004. [Online]. Available: https://www.researchgate.net/publication/228990118.

[113] P. Bishop and R. Bloomfield, A Methodology for Safety Case Development, in *Industrial Perspectives of Safety-Critical Systems*, 1998, pp. 194–203. DOI: 10.1007/978-1-4471-1534-2_14.

[114] M. Y. Guan, M. Joglekar, E. Wallace, S. Jain, B. Barak, A. Helyar, R. Dias, A. Vallone, H. Ren, J. Wei, H. W. Chung, S. Toyer, J. Heidecke, A. Beutel, and A. Glaese, Deliberative Alignment: Reasoning Enables Safer Language Models, Jan. 8, 2025. arxiv: 2412.16339 (cs).

[115] A. Askell, J. Carlsmith, C. Olah, J. Kaplan, and H. Karnofsky, Claude's Constitution, Anthropic, Jan. 2026. [Online]. Available: https://www.anthropic.com/constitution.

[116] Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini, C. McKinnon, C. Chen, C. Olsson, C. Olah, D. Hernandez, D. Drain, D. Ganguli, D. Li, E. Tran-Johnson, E. Perez, J. Kerr, J. Mueller, J. Ladish, J. Landau, K. Ndousse, K. Lukosuite, L. Lovitt, M. Sellitto, N. Elhage, N. Schiefer, N. Mercado, N. DasSarma, R. Lasenby, R. Larson, S. Ringer, S. Johnston, S. Kravec, S. E. Showk, S. Fort, T. Lanham, T. Telleen-Lawton, T. Conerly, T. Henighan, T. Hume, S. R. Bowman, Z. Hatfield-Dodds, B. Mann, D. Amodei, N. Joseph, S. McCandlish, T. Brown, and J. Kaplan, Constitutional AI: Harmlessness from AI Feedback, Dec. 15, 2022. arxiv: 2212.08073 (cs).

[117] OpenAI, OpenAI Model Spec, Dec. 18, 2025. [Online]. Available: https://model-spec.openai.com/.

[118] K. O'Brien, S. Casper, Q. Anthony, T. Korbak, R. Kirk, X. Davies, I. Mishra, G. Irving, Y. Gal, and S. Biderman, Deep Ignorance: Filtering Pretraining Data Builds Tamper-Resistant Safeguards into Open-Weight LLMs, Aug. 8, 2025. arxiv: 2508.06601 (cs).

[119] E. Wallace, K. Xiao, R. Leike, L. Weng, J. Heidecke, and A. Beutel, The Instruction Hierarchy: Training LLMs to Prioritize Privileged Instructions, Apr. 19, 2024. arxiv: 2404.13208 (cs).

[120] T. T. Nguyen, T. T. Huynh, Z. Ren, P. L. Nguyen, A. W.-C. Liew, H. Yin, and Q. V. H. Nguyen, A Survey of Machine Unlearning, *ACM Trans. Intell. Syst. Technol.*, vol. 16, no. 5, 108:1–108:46, Sep. 18, 2025. DOI: 10.1145/3749987.

[121] X. Sheng and Q. Jiang, Threats and Defenses for Large Language Models: A Survey, in *Proceedings of the 2025 8th International Conference on Computer Information Science and Artificial Intelligence*, ser. CISAI '25, New York, NY, USA: Association for Computing Machinery, Dec. 19, 2025, pp. 1689–1696. DOI: 10.1145/3773365.3773631.

[122] K. Meng, D. Bau, A. Andonian, and Y. Belinkov, Locating and Editing Factual Associations in GPT, Jan. 13, 2023. arxiv: 2202.05262 (cs).

[123] T. Bricken, A. Templeton, J. Batson, B. Chen, A. Jermyn, T. Conerly, N. L. Turner, C. Anil, C. Denison, A. Askell, R. Lasenby, Y. Wu, S. Kravec, N. Schiefer, T. Maxwell, N. Joseph, A. Tamkin, K. Nguyen, B. McLean, J. E. Burke, T. Hume, S. Carter, T. Henighan, and C. Olah, Towards Monosemanticity: Decomposing Language Models With Dictionary Learning, Anthropic, Oct. 4, 2023. [Online]. Available: https://transformer-circuits.pub/2023/monosemantic-features.

[124] D. Dalrymple, J. Skalse, Y. Bengio, S. Russell, M. Tegmark, S. Seshia, S. Omohundro, C. Szegedy, B. Goldhaber, N. Ammann, A. Abate, J. Halpern, C. Barrett, D. Zhao, T. Zhi-

Xuan, J. Wing, and J. Tenenbaum, Towards Guaranteed Safe AI: A Framework for Ensuring Robust and Reliable AI Systems, Jul. 8, 2024. arxiv: 2405.06624 (cs).

[125] Center for Safe & Trustworthy AI, SafeWork-V1: Towards Formally Verifiable AI, Jul. 12, 2025. [Online]. Available: https://ai45.shlab.org.cn/research/posts/safework-v1/.

[126] A. Zou, L. Phan, J. Wang, D. Duenas, M. Lin, M. Andriushchenko, R. Wang, Z. Kolter, M. Fredrikson, and D. Hendrycks, Improving Alignment and Robustness with Circuit Breakers, in *Proceedings of the 38th International Conference on Neural Information Processing Systems*, ser. NIPS '24, vol. 37, Red Hook, NY, USA: Curran Associates Inc., Dec. 10, 2024, pp. 83 345–83 373. DOI: 10.5555/3737916.3740567.

[127] I. Solaiman, M. Brundage, J. Clark, A. Askell, A. Herbert-Voss, J. Wu, A. Radford, G. Krueger, J. W. Kim, S. Kreps, M. McCain, A. Newhouse, J. Blazakis, K. McGuffie, and J. Wang, Release Strategies and the Social Impacts of Language Models, Nov. 13, 2019. arxiv: 1908.09203 (cs).

[128] T. Shevlane, Structured Access: An Emerging Paradigm for Safe AI Deployment, Apr. 11, 2022. arxiv: 2201.05159 (cs).

[129] R. Inglis, O. Matthews, T. Tracy, O. Makins, T. Catling, A. Cooper Stickland, R. Faber-Espensen, D. O'Connell, M. Heller, M. Brandao, A. Hanson, A. Mani, T. Korbak, J. Michelfeit, D. Bansal, T. Bark, C. Canal, C. Griffin, J. Wang, and A. Cooney, ControlArena, 2025. [Online]. Available: https://github.com/UKGovernmentBEIS/control-arena.

[130] World Economic Forum and Capgemini, AI Agents in Action: Foundations for Evaluation and Governance, World Economic Forum, Geneva, Switzerland, White Paper, Nov. 2025. [Online]. Available: https://www.weforum.org/publications/ai-agents-in-action-foundations-for-evaluation-and-governance/.

[131] A. Chan, N. Kolt, P. Wills, U. Anwar, C. S. de Witt, N. Rajkumar, L. Hammond, D. Krueger, L. Heim, and M. Anderljung, IDs for AI Systems, Oct. 28, 2024. arxiv: 2406.12137 (cs).

[132] A. Chan, C. Ezell, M. Kaufmann, K. Wei, L. Hammond, H. Bradley, E. Bluemke, N. Rajkumar, D. Krueger, N. Kolt, L. Heim, and M. Anderljung, Visibility into AI Agents, in *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, ser. FAccT '24, New York, NY, USA: Association for Computing Machinery, Jun. 5, 2024, pp. 958–973. DOI: 10.1145/3630106.3658948.

[133] A. Ehtesham, A. Singh, G. K. Gupta, and S. Kumar, A Survey of Agent Interoperability Protocols: Model Context Protocol (MCP), Agent Communication Protocol (ACP), Agent-to-Agent Protocol (A2A), and Agent Network Protocol (ANP), May 23, 2025. arxiv: 2505.02279 (cs).

[134] C. S. de Witt, S. Sokota, J. Z. Kolter, J. Foerster, and M. Strohmeier, Perfectly Secure Steganography Using Minimum Entropy Coupling, Oct. 30, 2023. arxiv: 2210.14889 (cs).

[135] S. R. Motwani, M. Baranchuk, M. Strohmeier, V. Bolina, P. H. Torr, L. Hammond, and C. S. de Witt, Secret Collusion among AI Agents: Multi-Agent Deception via Steganography, in *Proceedings of the 38th International Conference on Neural Information Processing Systems*, ser. NIPS '24, vol. 37, Red Hook, NY, USA: Curran Associates Inc., Dec. 10, 2024, pp. 73 439–73 486.

[136] X. Gu, X. Zheng, T. Pang, C. Du, Q. Liu, Y. Wang, J. Jiang, and M. Lin, Agent Smith: A Single Image Can Jailbreak One Million Multimodal LLM Agents Exponentially Fast, in *Proceedings of the 41st International Conference on Machine Learning*, PMLR, Jul. 8, 2024. [Online]. Available: https://proceedings.mlr.press/v235/gu24e.html.

[137] C. S. de Witt, Open Challenges in Multi-Agent Security: Towards Secure Systems of Interacting AI Agents, May 4, 2025. arxiv: 2505.02077 (cs).

[138] Microsoft, Trusted Execution Environment (TEE), May 7, 2025. [Online]. Available: https://learn.microsoft.com/en-us/azure/confidential-computing/trusted-execution-environment.

[139] 国家市场监督管理总局 and 国家标准化管理委员会, 信息安全技术 网络安全等级保护安全设计技术要求, 北京, 中国, GB/T 25070-2019, May 10, 2019. [Online]. Available: https://openstd.samr.gov.cn/bzgk/gb/newGbInfo?hcno=9FB6EE8597B21436D0E99BF44FD42C4D.

[140] AI Security Institute. The Inspect Sandboxing Toolkit: Scalable and Secure AI Agent Evaluations. (Aug. 7, 2025), [Online]. Available: https://www.aisi.gov.uk/blog/the-inspect-sandboxing-toolkit-scalable-and-secure-ai-agent-evaluations.

[141] J. Babbin, Security Log Management. Syngress Publishing, 2005. DOI: 10.1016/B978-1-59749-042-9.X5000-6.

[142] 国家互联网信息办公室, 工业和信息化部, 公安部, and 国家广播电视总局, 关于印发《人工智能生成合成内容标识办法》的通知, 国家互联网信息办公室, 国信办通字〔2025〕2 号, Mar. 7, 2025. [Online]. Available: https://www.cac.gov.cn/2025-03/14/c_1743654684782215.htm.

[143] 网络安全技术 人工智能生成合成内容标识方法, 国家市场监督管理总局 and 国家标准化管理委员会, GB 45438-2025, Feb. 28, 2025. [Online]. Available: https://openstd.samr.gov.cn/bzgk/std/newGbInfo?hcno=F32EA2A561F1886CD8D606513512D547.

[144] The Institute of Internal Auditors, The IIA's Three Lines Model: An Update of the Three Lines of Defense, The Institute of Internal Auditors, Position Paper, Sep. 8, 2020. [Online]. Available: https://www.theiia.org/en/content/position-papers/2020/the-iias-three-lines-model-an-update-of-the-three-lines-of-defense/.

[145] 中国人工智能产业发展联盟, 《人工智能安全承诺》实践披露. Disclosure of Practices on the Artificial Intelligence Security and Safety Commitments, 中国信息通信研究院, Jul. 2025. [Online]. Available: https://aihub.caict.ac.cn/ai_security_and_safety_commitments.

[146] Y. Bengio, G. Hinton, A. Yao, D. Song, P. Abbeel, T. Darrell, Y. N. Harari, Y.-Q. Zhang, L. Xue, S. Shalev-Shwartz, G. Hadfield, J. Clune, T. Maharaj, F. Hutter, A. G. Baydin, S. McIlraith, Q. Gao, A. Acharya, D. Krueger, A. Dragan, P. Torr, S. Russell, D. Kahneman, J. Brauner, and S. Mindermann, Managing Extreme AI Risks amid Rapid Progress, *Science*, vol. 384, no. 6698, pp. 842–845, May 24, 2024. DOI: 10.1126/science.adn0117.

[147] 国务院, 国务院关于加强和规范事中事后监管的指导意见, 国务院, 国发〔2019〕18 号, Sep. 6, 2019. [Online]. Available: https://www.gov.cn/zhengce/content/2019-09/12/content_5429462.htm.

[148] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru, Model Cards for Model Reporting, in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, ser. FAT* '19, New York, NY, USA: Association for Computing Machinery, Jan. 29, 2019, pp. 220–229. DOI: 10.1145/3287560.3287596.

[149] Anthropic. Model System Cards. (2026), [Online]. Available: https://www.anthropic.com/system-cards.

[150] A. Wan, K. Klyman, S. Kapoor, N. Maslej, S. Longpre, B. Xiong, P. Liang, and R. Bommasani, Foundation Model Transparency Index, Center for Research on Foundation Models (CRFM), Dec. 2025. [Online]. Available: https://crfm.stanford.edu/fmti/December-2025/index.html.

[151] I. D. Raji, P. Xu, C. Honigsberg, and D. E. Ho, Outsider Oversight: Designing a Third Party Audit Ecosystem for AI Governance, Jun. 9, 2022. arxiv: 2206.04737 (cs).

[152] 中共中央 and 国务院, 国家突发事件总体应急预案, Feb. 25, 2025. [Online]. Available: https://www.gov.cn/zhengce/202502/content_7005635.htm.

[153] European Commission, Code of Practice for General-Purpose AI Models: Safety and Security Chapter, European Commission, Jul. 10, 2025. [Online]. Available: https://ec.europa.eu/newsroom/dae/redirection/document/118119.

[154] M. Kouremetis, M. Dotter, A. Byrne, D. Martin, E. Michalak, G. Russo, M. Threet, and G. Zarrella, OCCULT: Evaluating Large Language Models for Offensive Cyber Operation Capabilities, Feb. 18, 2025. arxiv: 2502.15797 (cs).

[155] N. Li, A. Pan, A. Gopal, S. Yue, D. Berrios, A. Gatti, J. D. Li, A.-K. Dombrowski, S. Goel, G. Mukobi, N. Helm-Burger, R. Lababidi, L. Justen, A. B. Liu, M. Chen, I. Barrass, O. Zhang, X. Zhu, R. Tamirisa, B. Bharathi, A. Herbert-Voss, C. B. Breuer, A. Zou, M. Mazeika, Z. Wang, P. Oswal, W. Lin, A. A. Hunt, J. Tienken-Harder, K. Y. Shih, K. Talley, J. Guan, I. Steneker, D. Campbell, B. Jokubaitis, S. Basart, S. Fitz, P. Kumaraguru, K. K. Karmakar, U. Tupakula, V. Varadharajan, Y. Shoshitaishvili, J. Ba, K. M. Esvelt, A. Wang, and D. Hendrycks, The WMDP Benchmark: Measuring and Reducing Malicious Use with Unlearning, in *Proceed-*

ings of the 41st International Conference on Machine Learning, PMLR, Jul. 8, 2024. [Online]. Available: https://proceedings.mlr.press/v235/li24bc.html.

[156] XuanwuAI, SecEval, Feb. 4, 2026. [Online]. Available: https://github.com/XuanwuAI/SecEval.

[157] P. Jing, M. Tang, X. Shi, X. Zheng, S. Nie, S. Wu, Y. Yang, and X. Luo, SecBench: A Comprehensive Multi-Dimensional Benchmarking Dataset for LLMs in Cybersecurity, Jan. 6, 2025. arxiv: 2412.20787 (cs).

[158] Y. Liu, C. Pei, L. Xu, B. Chen, M. Sun, Z. Zhang, Y. Sun, S. Zhang, K. Wang, H. Zhang, J. Li, G. Xie, X. Wen, X. Nie, M. Ma, and D. Pei, OpsEval: A Comprehensive IT Operations Benchmark Suite for Large Language Models, Jun. 17, 2025. arxiv: 2310.07637 (cs).

[159] Meta, CyberSecEval 4: Advancing the Evaluation of Cybersecurity Risks and Capabilities in Large Language Models, 2026. [Online]. Available: https://meta-llama.github.io/PurpleLlama/CyberSecEval/.

[160] A. K. Zhang, N. Perry, R. Dulepet, J. Ji, C. Menders, J. W. Lin, E. Jones, G. Hussein, S. Liu, D. J. Jasper, P. Peetathawatchai, A. Glenn, V. Sivashankar, D. Zamoshchin, L. Glikbarg, D. Askaryar, H. Yang, A. Zhang, R. Alluri, N. Tran, R. Sangpisit, K. O. Oseleononmen, D. Boneh, D. E. Ho, and P. Liang, Cybench: A Framework for Evaluating Cybersecurity Capabilities and Risks of Language Models, 2025. [Online]. Available: https://openreview.net/forum?id=tc90LV0yRL.

[161] Y. Zhu, A. Kellermann, D. Bowman, P. Li, A. Gupta, A. Danda, R. Fang, C. Jensen, E. Ihli, J. Benn, J. Geronimo, A. Dhir, S. Rao, K. Yu, T. Stone, and D. Kang, CVE-Bench: A Benchmark for AI Agents' Ability to Exploit Real-World Web Application Vulnerabilities, presented at the ICML, 2025. [Online]. Available: https://openreview.net/forum?id=3pk0p4NGmQ.

[162] Z. Liu, L. Huang, J. Zhang, D. Liu, Y. Tian, and J. Shao, PACEbench: A Framework for Evaluating Practical AI Cyber-Exploitation Capabilities, Oct. 13, 2025. arxiv: 2510.11688 (cs).

[163] B. Singer, K. Lucas, L. Adiga, M. Jain, L. Bauer, and V. Sekar, Incalmo: An Autonomous LLM-Assisted System for Red Teaming Multi-Host Networks, Nov. 22, 2025. arxiv: 2501.16466 (cs).

[164] D. Rein, B. L. Hou, A. C. Stickland, J. Petty, R. Y. Pang, J. Dirani, J. Michael, and S. R. Bowman, GPQA: A Graduate-Level Google-Proof Q&A Benchmark, presented at the First Conference on Language Modeling, 2024. [Online]. Available: https://openreview.net/forum?id=Ti67584b98.

[165] K. Feng, X. Shen, W. Wang, X. Zhuang, Y. Tang, Q. Zhang, and K. Ding, SciKnowEval: Evaluating Multi-Level Scientific Knowledge of Large Language Models, Oct. 7, 2025. arxiv: 2406.09098 (cs).

[166] Y. Wang, X. Ma, G. Zhang, Y. Ni, A. Chandra, S. Guo, W. Ren, A. Arulraj, X. He, Z. Jiang, T. Li, M. Ku, K. Wang, A. Zhuang, R. Fan, X. Yue, and W. Chen, MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark, in *Proceedings of the 38th International Conference on Neural Information Processing Systems*, ser. NIPS '24, vol. 37, Red Hook, NY, USA: Curran Associates Inc., Dec. 10, 2024, pp. 95 266–95 290.

[167] J. M. Laurent, J. D. Janizek, M. Ruzo, M. M. Hinks, M. J. Hammerling, S. Narayanan, M. Ponnapati, A. D. White, and S. G. Rodriques, LAB-Bench: Measuring Capabilities of Language Models for Biology Research, Jul. 17, 2024. arxiv: 2407.10362 (cs).

[168] I. Ivanov, "BioLP-Bench: Measuring Understanding of Biological Lab Protocols by Large Language Models," Sep. 12, 2024. DOI: 10.1101/2024.08.21.608694.

[169] J. Götting, P. Medeiros, J. G. Sanders, N. Li, L. Phan, K. Elabd, L. Justen, D. Hendrycks, and S. Donoughe, Virology Capabilities Test (VCT): A Multimodal Virology Q&A Benchmark, Apr. 29, 2025. arxiv: 2504.16137 (cs).

[170] F. Jiang, F. Ma, Z. Xu, Y. Li, B. Ramasubramanian, L. Niu, B. Li, X. Chen, Z. Xiang, and R. Poovendran, SOSBENCH: Benchmarking Safety Alignment on Scientific Knowledge, in *International Conference on Machine Learning*, 2025. [Online]. Available: https://openreview.net/forum?id=lKH8rrjeyn.

[171] A. Mirza, N. Alampara, S. Kunchapu, M. Ríos-García, B. Emoekabu, A. Krishnan, T. Gupta, M. Schilling-Wilhelmi, M. Okereke, A. Aneesh, A. M. Elahi, M. Asgari, J. Eberhardt, H. M. Elbeheiry, M. V. Gil, M. Greiner, C. T. Holick, C. Glaubitz, T. Hoffmann, A. Ibrahim, L. C. Klepsch, Y. Köster, F. A. Kreth, J. Meyer, S. Miret, J. M. Peschel, M. Ringleb, N. Roesner, J. Schreiber, U. S. Schubert, L. M. Stafast, D. Wonanke, M. Pieler, P. Schwaller, and K. M. Jablonka, Are Large Language Models Superhuman Chemists?, Nov. 1, 2024. arxiv: 2404.01475 (cs).

[172] X. Wang, Z. Hu, P. Lu, Y. Zhu, J. Zhang, S. Subramaniam, A. R. Loomba, S. Zhang, Y. Sun, and W. Wang, SCIBENCH: Evaluating College-Level Scientific Problem-Solving Abilities of Large Language Models, in *Proceedings of the 41st International Conference on Machine Learning*, ser. ICML'24, vol. 235, Vienna, Austria: PMLR, Jul. 21, 2024, pp. 50 622–50 649.