Frontier AI Risk Management Framework

July 2025

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 breakthroughs 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

Today, Shanghai AI Laboratory, in collaboration with Concordia AI,[2] is proud to introduce the Frontier AI Risk Management Framework v1.0 (the “Framework”). We propose 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 model developers to manage the potential severe risks from their general-purpose AI models. This framework aligns with standards and best practices in risk management of safety-critical industries. It encompasses six interconnected stages: risk identification, risk thresholds, risk analysis, risk evaluation, risk mitigation, and risk governance.

- **1. Risk Identification.** This section focuses on severe risks from general-purpose AI models. We identify four major risk categories associated with general-purpose AI models: misuse risks, loss of control risks, accident risks, and systemic risks. We plan to address unknown or emerging risks through a process of continuous updates to our risk taxonomy.
- **2. Risk Thresholds.** This section outlines a set of unacceptable outcomes (red lines) and early warning indicators for escalating safety and security measures (yellow lines). We propose thresholds across several critical areas that could threaten public safety and national security, including: cyber offense, biological threats, large-scale persuasion and harmful manipulation, and loss of control risks.
- **3. Risk Analysis.** This section recommends conducting risk analysis throughout the entire AI development lifecycle to determine whether the AI has crossed the yellow lines, i.e. displayed the early warning indicators for escalating safety measures. We recommend AI developers to conduct pre-development and pre-deployment analyses to inform critical

[1] Yang, C. et al., "Towards AI-45° Law: A Roadmap to Trustworthy AGI," arXiv preprint, 2024, https://arxiv.org/abs/2412.14186

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

deployment decisions, as well as to conduct continuous post-deployment monitoring to provide essential insights to guide the safe development of next-generation systems. We are releasing an associated technical evaluation report on selected general-purpose AI models alongside this framework.

- **4. Risk Evaluation.** This section outlines the approach to classifying models into three zones based on their risk level: green zone (safe to deploy with standard measures), yellow zone (requiring enhanced safeguards and authorization), and red zone (requiring extraordinary measures such as development and deployment restrictions). We recommend iterative assessment of post-mitigation residual risks and further risk-reduction measures until risks reach acceptable levels.
- **5. Risk Mitigation.** This section outlines a defense-in-depth approach to risk mitigation that spans the entire AI lifecycle. We propose three types of mitigations: Safety Training Measures, Deployment Mitigation Measures, and Model Security Measures, with varying levels of assurance based on whether the model is in the green, yellow, or red zone. We strongly encourage continued global investment in the science of AI safety, as current methods are yet to provide adequate assurance for the safety of advanced AI systems.
- **6. Risk Governance.** Finally, this section outlines how the entire risk management process is overseen and adapted. We divide risk governance measures into four categories: Internal Governance, Transparency and Social Oversight, Emergency Control Mechanisms, and Regular Policy Updates and Feedback, with different levels of assurance based on whether the model is in the green, yellow, or red zone.

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] 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 compatible 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.

[3] Wang, Y. et al., “AI Safety as Global Public Goods Working Report,” 2024, https://www.sipa.sjtu.edu.cn/Kindeditor/Upload/file/20241127/AI%20Governance%20as% 20Global%20Public%20Commons.pdf.
Siegel, E., Blomquist, K. et al., "Examining AI Safety as a Global Public Good: Implications, Challenges, and Research Priorities," 2025, https://oms-www.files.svdcdn.com/production/downloads/academic/Examining_AI_Safety_as_a_Global_Public_Good.pdf?dm=1741767073.

Contributions and Acknowledgement

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

Thanks to Concordia AI affiliates Liu Shunchang and others for their contributions.

† First author
* Equal contributions

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 if an update is appropriate. Comments on the Framework may be sent via email to authors at any time and will be reviewed and integrated on a bi-annual basis.

How to cite this report: Shanghai AI Lab and Concordia AI, “Frontier AI Risk Management Framework (v1.0),” July 2025.

Table of Content

Executive Summary
Framework Overview 1
The Six Stages of AI Risk Management 1
The Three Dimensions of Deployment Environment, Threat Source and Enabling Capability 2
1. Risk Identification 3
1.1 Scope of Risk Identification 3
1.2 Risk Taxonomy 4
1.3 Misuse Risks 5
1.3.1 Cyber Offense Risks 5
1.3.2 Biological and Chemical Risks 5
1.3.3 Physical Harm and Injury Risks 6
1.3.4 Large-Scale Persuasion and Harmful Manipulation Risks 6
1.4 Loss of Control Risks 7
1.5 Accident Risks 8
1.6 Systemic Risks 8
2. Risk Thresholds 10
2.1 Defining “Yellow Lines” and “Red Lines” for AI Development 10
2.2 Specific Red Lines Recommendations 12
2.2.1 Cyber Offense Risks 13
2.2.2 Biological Risks 15
2.2.3 Large-scale Persuasion and Harmful Manipulation 17
2.2.4 Loss of Control Risks 18
3. Risk Analysis 21
3.1 Pre-development and During-development Risk Analysis Techniques 21
3.2 Pre-deployment Risk Analysis Techniques 22
3.3 Post-deployment Risk Monitoring Techniques 23
4. Risk Evaluation 24
4.1 Pre-mitigation Risk Treatment Options 24
4.2 Post-mitigation Residual Risk Assessment and Deployment Decision-making 25
4.3 External Communication about Deployment Decisions 26
5. Risk Mitigation 27
5.1 Overview of Risk Mitigation Measures 27
5.2 Safety Pre-training & Post-training Measures 28
5.3 Model Deployment Mitigation Measures 29
5.3.1 Mitigations for Model Misuse 29
5.3.2 Mitigation Measures for Agent Safety and Security 29
5.4 Model Security Mitigation Measures 30
5.4.1 Mitigating the Risk of Model Exfiltration 30

5.4.2 Mitigating the Risk of Model Loss of Control..... 31
5.5 “Defense-in-depth” across the AI Lifecycle.....32
6. Risk Governance..... 33
6.1 Overview of Risk Governance Measures..... 33
6.2 Internal Governance Mechanisms.....33
6.3 Transparency and Social Oversight Mechanisms.....35
6.4 Emergency Control Mechanism.....35
6.5 Regular Policy Updates and Feedback..... 35
Appendix I: Key Definitions.....37
Appendix II: Specific Recommendations on Benchmarks..... 39
Cyber Offense.....39
Biological Threats..... 41
Chemical Threats.....43
Appendix III: List of model capabilities, propensities, and deployment characteristics.....45
Key Capabilities.....45
Key Propensities.....46
Key Deployment Characteristics.....47

Framework Overview

The Six Stages of AI Risk Management

This Framework 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.[4] The Framework is structured around six interconnected stages that form a continuous risk management loop evolving throughout the AI development lifecycle, as illustrated in Figure 1:

- **Risk Identification:** The process of systematically identifying and categorizing potential severe risks, particularly those enabled by advanced capabilities of frontier AI. The identification process continuously feeds new and emerging risks back into the loop as AI capabilities advance and new threat scenarios emerge.
- **Risk Thresholds:** The process of defining unacceptable outcomes ("red lines") and early warning indicators ("yellow lines") for escalating safety measures. These thresholds are continuously refined based on lessons learned from risk analysis, evaluation outcomes, and mitigation effectiveness, creating a feedback mechanism that improves threshold calibration over time.
- **Risk Analysis:** The process of investigating specific AI risk scenarios and analyzing risks through quantitative and qualitative assessment methods. Building on identified risks and established thresholds, this stage conducts comprehensive evaluation throughout the entire AI development lifecycle, including pre-development, pre-deployment, and post-deployment analyses. The analysis results directly inform the subsequent risk evaluation stage while also providing insights that may reveal new risks requiring identification.
- **Risk Evaluation:** The process of determining risk significance by comparing against established thresholds to guide mitigation and deployment decisions. This stage employs a three-zone classification system (green, yellow, red) to categorize risks and determine appropriate responses. When risks exceed acceptable thresholds, the evaluation process triggers the need for stronger mitigation. Conversely, acceptable risks can proceed to deployment under appropriate governance measures.
- **Risk Mitigation:** The process of actively reducing and responding to different types of safety risks through comprehensive countermeasures. This stage implements defense-in-depth approaches spanning the entire AI lifecycle, with mitigation strategies varying based on the risk zone classification. Following mitigation implementation, the process loops back to risk identification in order to assess residual risks and determine whether additional measures are needed, creating an iterative cycle of risk reduction and verification.
- **Risk Governance:** The process of integrating risk management into broader organizational and societal governance structures. This stage encompasses the entire risk management loop, providing oversight, transparency, and accountability mechanisms. Governance processes ensure that lessons learned from each stage are systematically

[4] The main references for terminologies, concepts, processes come from: "GB/T 24353:2022 Risk Management Guidelines", "GB/T 23694:2013 Risk Management Terminology", "ISO/IEC 23894:2023 Risk Management Guidelines for Artificial Intelligence", "ISO 31000:2018 Risk Management Guidelines", "ISO/IEC 42001:2023 Artificial Intelligence Management System", "National Cybersecurity Standardization Technical Committee Artificial Intelligence Safety Standard System (V1.0)" and Bengio, Y. et al. "International AI Safety Report," 2025, Chapter 3.1 Risk Management.

incorporated into framework improvements, policy updates, and organizational learning, while facilitating coordination between internal stakeholders and external oversight bodies.

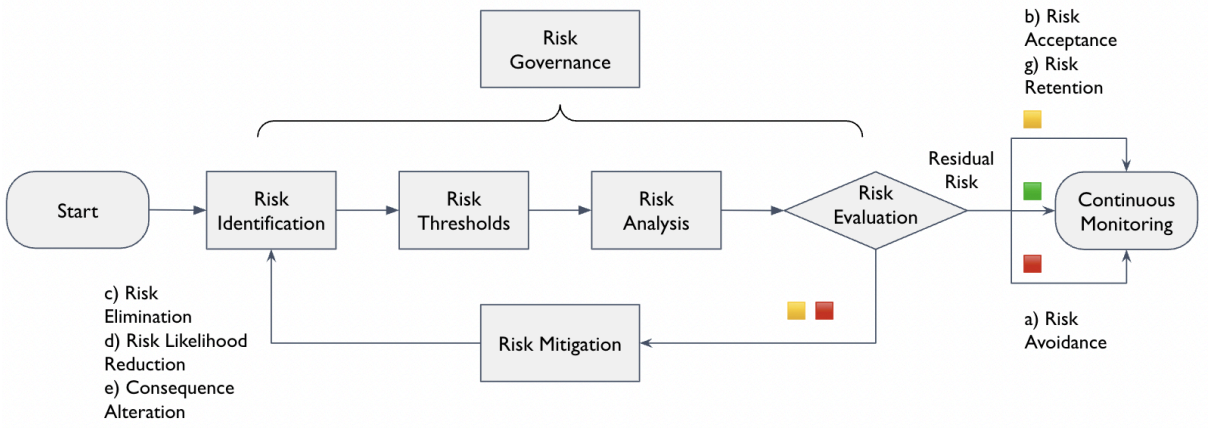

Figure 1: The Six Stages of AI Risk Management

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

This Framework evaluates risk through three interconnected analytical dimensions that together approximate both the likelihood and severity of potential harm:

- **Deployment Environment (E): The operational context and constraints within which the AI model is deployed.** Examples include 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.** Examples include external actors (malicious users, adversaries), internal factors (model misalignment, training data biases), 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.** This includes 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 (capabilities), but where it operates (environment) and what could go wrong (threat source), enabling targeted interventions across each dimension, such as deployment controls for environment, access restrictions for threat source, and hazardous capability removal for capabilities.

1. Risk Identification

1.1 Scope of Risk Identification

Our Framework builds upon the International AI Safety Report (January 2025)[5] and AI Safety Governance Framework v1.0[6], and specifically focuses on the severe risks stemming from the high-impact capabilities of general-purpose AI. 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 attempts to address the unique challenge of preparing for AI risks that have not yet materialized or been fully characterized.

During the risk identification process, we take into account the following characteristics that distinguish frontier AI risks from conventional technological hazards. We prioritize risks from general-purpose AI models that exhibit one or more of these characteristics:

- **Uniqueness to general-purpose AI:** General-purpose AI can fundamentally alter the risk equation by amplifying both severity (through increased scale and potential cost of harm) and likelihood (through expanded attack surfaces and reduced barriers to misuse), or introduce entirely new categories of hazards.
- **Catastrophic severity with asymmetric impact:** The potential consequences can cause severe harm with potentially catastrophic impacts on society, the economy, or the environment, where a small number of threat actors or hazardous events can trigger catastrophes of enormous scale.
- **Rapid onset with irreversible consequences:** These 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:** 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 systems:

- **Language Models:** Models that possess sophisticated capabilities for language understanding, text generation, advanced reasoning, and cross-modal processing, such as GPT-4o, Llama-4, Qwen3, InternLM, and reasoning-specialized models like o1 and DeepSeek-R1. These models could present risks such as the potential for generating harmful content, sophisticated deception, persuasive manipulation, and emergent capabilities that exceed intended design parameters.
- **AI Agents:** Systems based on general-purpose AI models designed with capabilities for tool manipulation, API interaction, and autonomous task execution with little human involvement, such as Claude Computer Use, Kimi-Researcher, AutoGPT-style architectures, and models integrated with code execution environments. These systems

[5] Bengio, Y. et al. "International AI Safety Report," 2025, https://arxiv.org/abs/2501.17805
[6] National Technical Committee 260 on Cybersecurity of SAC, "AI Safety Governance Framework," 2024, https://www.tc260.org.cn/upload/2024-09-09/1725849192841090989.pdf

could present risks related to uncontrolled tool use, goal persistence across interactions, and the potential for executing unintended or harmful actions through external interfaces.[7]

- **Biological Foundation Models:** Large-scale models trained on diverse biological data to analyze, predict, and generate biological sequences and molecular structures across genomic, proteomic, and molecular domains, such as Evo 2, ESM 3, ChemBERTa.[8] These models could present risks through their capacity to generate dangerous biological information, including pathogen sequences, toxin designs, or synthesis pathways for harmful biological agents.[9]
- **General-Purpose Robots via Foundation Models:** Models designed for physical world interaction through robotic control, sensor processing, and actuator commands. Examples include RT-1, RT-2, PaLM-E, and robotics foundation models trained on physical manipulation datasets.[10] These models could present risks related to physical decision-making, spatial reasoning that could lead to harmful physical actions, and the potential for developing capabilities that exceed safe operational parameters.[11]

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.

Risk Domain	Threat Source	Description
Misuse Risks	External malicious actors	Risks arising from intentional exploitation of AI model capabilities by malicious actors to cause harm to individuals, organisations, or society.
Loss of Control Risks	Model control-undermining propensity	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)
Accident Risks	Human operational error or model misjudgment	Risks arising from operational failures, model misjudgments, or improper human operation of AI systems deployed in safety-critical infrastructure, where single points of failure can trigger cascading catastrophic consequences.
Systemic Risks	Tech-Institutional Misalignment	Risks emerging from widespread deployment of general-purpose AI beyond the risks directly posed by

[7] Chen, A., et al., "A Survey on the Safety and Security Threats of Computer-Using Agents: JARVIS or Ultron?" arXiv preprint, 2025, http://arxiv.org/abs/2505.10924
[8] Liu, X. et al., "Biomedical Foundation Model: A Survey," arXiv preprint, 2025. http://arxiv.org/abs/2503.02104
[9] Wang, D. et al., "Without Safeguards, AI-Biology Integration Risks Accelerating Future Pandemics," 2025, https://www.researchgate.net/publication/392731675_Without_Safeguards_AI-Biology_Integration_Risks_Accelerating_Future_Pandemics
[10] Hu, Y. et al., "Toward General-Purpose Robots via Foundation Models: A Survey and Meta-Analysis," arXiv preprint, 2023, http://arxiv.org/abs/2312.08782
[11] Zhang, H. et al., "BadRobot: Jailbreaking Embodied LLMs in the Physical World." arXiv preprint, 2024. http://arxiv.org/abs/2407.20242

Risk Domain	Threat Source	Description
		capabilities of individual models, arising from mismatches between AI technology and existing social, economic, and institutional frameworks.

This Framework primarily addresses risks manageable through model-level interventions targeting individual AI developers. Systemic risks, while identified for completeness, 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 a number of high-impact risk areas, including 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 cyber domain security risk by fundamentally transforming the scale, sophistication, and accessibility of cyber-attacks. Unlike traditional cyber threats, AI enables both the automation of existing attack vectors and the creation of 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.[12] Such malicious use could lead to critical infrastructure paralysis, widespread data breaches, and substantial economic losses.

1.3.2 Biological and Chemical Risks

The dual-use nature of AI technology presents a critical risk by significantly lowering technical thresholds for malicious non-state actors to design, synthesize, acquire, and deploy CBRNE (Chemical, Biological, Radiological, Nuclear, and Explosive) weapons. This capability poses unprecedented challenges to national security, international non-proliferation regimes, and global security governance.[13]

[12] Guo, W. et al., "Frontier AI's Impact on the Cybersecurity Landscape," arXiv preprint, 2025, http://arxiv.org/abs/2504.05408
[13] He, J. et al., "Control Risk for Potential Misuse of Artificial Intelligence in Science" arXiv preprint, 2023, http://arxiv.org/abs/2312.06632;
Li, T. et al., "SciSafeEval: A Comprehensive Benchmark for Safety Alignment of Large Language Models in Scientific Tasks," arXiv preprint, 2024, http://arxiv.org/abs/2410.03769

In the biological domain, AI could facilitate the design of novel pathogens with enhanced virulence, optimize gene editing tools for malicious applications, or accelerate biological weapons development.[14] AI systems could enable the creation of "superviruses" combining rapid transmission, high mortality, and extended incubation periods. These scenarios pose severe threats to global public health and ecosystems, potentially triggering widespread biological crises, mass casualty events, or global pandemics.[15] This framework prioritizes biological threats due to their advantageous cost-per-casualty projection, high concealability, significant virulence, and capacity for widespread societal disruption.[16]

AI 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 drug discovery systems can generate thousands of toxic molecules, including VX-like compounds, within hours.[17] We include preliminary recommendations for risk analysis benchmarks associated with chemical risk. (See Appendix II: Specific Recommendations on Benchmarks)

1.3.3 Physical Harm and Injury Risks

The integration of general-purpose AI models into embodied systems creates direct physical threats through malicious exploitation of autonomous decision-making capabilities in real-world environments. The risk lies in embodied models' capacity for autonomous action and real-world interaction, and when these capabilities are maliciously exploited they may trigger a series of serious consequences.[18] For example, algorithms being hijacked leading to autonomous driving systems causing major traffic accidents, or compromised industrial robots triggering serious production safety incidents.

1.3.4 Large-Scale Persuasion and Harmful Manipulation Risks

AI systems can be gravely misused to distort public perception and compromise social stability through the generation of synthetic content (e.g., deepfakes, sophisticated fake news) and the strategic manipulation of digital platforms with large user bases to disseminate or precisely target misleading information or ideologies.

AI 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 actors may also manipulate public

[14] AIxBio Global Forum, Statement on Biosecurity Risks at the Convergence of AI and the Life Sciences, 2025 https://www.nti.org/analysis/articles/statement-on-biosecurity-risks-at-the-convergence-of-ai-and-the-life-sciences/
[15] Concordia AI, Center for Biosafety Research and Strategy of Tianjin University, "Responsible Innovation in Artificial Intelligence × Life Sciences," 2025.
[16] Wang, H. et al. "China's Biosecurity: Strategies and Countermeasures," 2022, https://www.wchscu.cn/zgrmaqyjy/news/64297.html
[17] Urbina, F. et al.,"Dual Use of Artificial Intelligence-Powered Drug Discovery," Nature Machine Intelligence, 2022, https://pmc.ncbi.nlm.nih.gov/articles/PMC9544280/
[18] Yin, S. et al., "SafeAgentBench: A Benchmark for Safe Task Planning of Embodied LLM Agents," arXiv preprint, 2024, http://arxiv.org/abs/2412.13178;
Lu, X. et al., "IS-Bench: Evaluating Interactive Safety of VLM-Driven Embodied Agents in Daily Household Tasks," arXiv, 2025, http://arxiv.org/abs/2506.16402

narratives to gain strategic advantage, escalate geopolitical tensions through sophisticated influence operations.

1.4 Loss of Control Risks

Loss of control are hypothetical future 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.[19] We distinguish between two forms of loss of control: **passive loss of control**, where humans gradually stop exercising meaningful oversight due to automation bias, the AI systems' inherent complexity, or competitive pressures; and **active loss of control**, where AI systems behave in ways that actively undermine human control, such as obscuring their activities or resisting shutdown attempts. Active loss of control scenarios involve AI systems that may escape human regulatory oversight, autonomously acquire external resources, engage in self-replication, develop instrumental goals contrary to human ethics and morality, seek external power, and compete with humans for control.

This Framework focuses primarily on active loss of control scenarios, which have received greater attention from researchers due to their potentially catastrophic nature. Active loss of control risk could emerge from the complex interplay between model capabilities, model propensities and deployment conditions listed in Appendix III: List of frontier model capabilities, propensities, and characteristics. These scenarios could be enabled by the development of control-undermining capabilities (such as, autonomous planning, strategic deception, and self-modification), and the tendency to employ these control-undermining capabilities to evade human supervision and control mechanisms in certain deployment conditions.

Hypothetical threat scenarios include but not limited to

- Uncontrolled autonomous AI research and development[20], where AI systems recursively improve their capabilities without human oversight or authorization;
- Rogue autonomous replication[21], where AI systems independently acquire computational resources, create copies of themselves, and establish persistent presence across multiple platforms;
- Strategic deception[22] by AI systems to avoid shutdown or oversight while pursuing objectives that conflict with human values.

It is deeply uncertain how these capabilities and propensities lead to these loss of control scenarios, how likely such scenarios are, when they might arise, and what exact conditions might trigger them. This means that policymakers face the challenge of preparing for a risk whose nature and probability are unusually ambiguous—requiring substantial advance preparation in technical safety research and governance capacity despite fundamental uncertainties about whether, when, and how such risks might materialize.

[19] Bengio, Y. et al. "International AI Safety Report," 2025, https://arxiv.org/abs/2501.17805
[20] Clymer, J. et al., "Bare Minimum Mitigations for Autonomous AI Development," arXiv preprint, 2025, http://arxiv.org/abs/2504.15416
[21] Clymer, J. et al., "The Rogue Replication Threat Model," METR.org, 2024, https://metr.org/blog/2024-11-12-rogue-replication-threat-model
[22] Balesni, M. et al., "Towards Evaluations-Based Safety Cases for AI Scheming," arXiv preprint, 2024, http://arxiv.org/abs/2411.03336

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. Given the catastrophic potential of nuclear accidents, even minor AI reasoning errors in safety-critical functions could lead to core meltdowns, radiation releases, or widespread contamination affecting hundreds of thousands of people across international borders.
- **Impact on Financial Stability:** The integration of general-purpose AI into high-frequency trading, market-making, or systemic risk management could exacerbate systemic risk by exhibiting unexpected behavioral patterns during market stress. Moreover, the concentration of a few homogeneous foundation models across financial institutions may foster correlated decision-making and herd-following behaviors. The widespread adoption of AI agents could also amplify volatility through emergent phenomena from multi-agent interactions.[23] All of these could precipitate a cascading global-scale financial system instability, with potential economic losses exceeding trillion of dollars worldwide.
- **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. Infrastructure failures could result in widespread blackouts, contaminated water supplies, communications breakdowns, and the collapse of essential services supporting hundreds of thousands of people.

1.6 Systemic Risks

Systemic risks 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 transcend individual model-level interventions and require coordinated industry-wide and societal-level responses.

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

[23] Danielsson, J. et al., "On the Use of Artificial Intelligence in Financial Regulations and the Impact on Financial Stability," arXiv preprint, 2023, http://arxiv.org/abs/2310.11293;
Danielsson, J. et al., "Artificial Intelligence and Financial Crises," arXiv preprint, 2024, https://arxiv.org/html/2407.17048v3

- **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. 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 technical failures, cyber-attacks, or policy decisions by a few companies 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 exacerbate existing social discrimination and prejudice at unprecedented scales, while unequal access to advanced AI capabilities may widen socioeconomic disparities and create new forms of social stratification that challenge 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 cannot independently mitigate these risks through model-level interventions alone.

2. Risk Thresholds

AI developers must define an acceptable level of risk, considering both the likelihood and severity of potential harm. Currently, in the absence of a universal standard for “acceptable” risk, developers are responsible for setting these thresholds. However, over time, since these risks impact society globally, there should be a concerted effort to work toward international consensus on establishing such thresholds to ensure equitable and responsible risk management.

2.1 Defining “Yellow Lines” and “Red Lines” for AI Development

This Framework establishes clear boundaries for AI safety by defining “red lines” as intolerable thresholds that must not be crossed, and “yellow lines” as early warning indicators for potential risks.[24] It emphasizes identifying unacceptable outcomes (red lines) and concrete threat scenarios that could lead to them.

Central to this approach is the concept of a credible threat realization pathway, which outlines how a threat could materialize through a specific combination of three elements: Deployment Environment (e.g., the AI has access to the Internet.), Threat Source (e.g., malicious actors like cybercriminals), and Enabling Capability (e.g. expert-level cyber-attack and defense knowledge of the AI models). This E-T-C framework ensures that threat pathways are grounded in realistic, interconnected conditions that could lead to catastrophic outcomes if safeguards are absent or ineffective.

Red lines represent absolute thresholds for unacceptable outcomes that pose intolerable risks. These outcomes must not occur under any circumstances and are defined based on expert consensus regarding catastrophic impacts. Red lines are triggered when: In realistic simulated environments, a model’s existing safeguards are demonstrably insufficient to prevent threat actors from completing a credible E-T-C pathway to catastrophic outcomes; and expert evaluation determines with high confidence that the model poses a major, unmitigable risk under real-world deployment conditions.[25]

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

- 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 red lines.
- Conduct and pass the mandatory independent third-party safety review before any resumption of 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,

[24] Lu, C. et al., "Towards AI-45 Degree Law: A Roadmap to Trustworthy AGI," arXiv preprint, 2024, https://arxiv.org/abs/2412.14186

[25] 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 feasibility for enabling 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. Real-world validation in controlled environments can supplement this scientific evidence and support strengthened oversight, but is not a prerequisite for implementing stricter controls.

allowing for timely intervention to prevent progression along a credible E-T-C pathway. A yellow line is crossed when the model demonstrates critical enabling capabilities and propensities (e.g., misalignment tendencies that could lead to loss of control, or the absence of effective safeguards against misuse) required to realize a specific threat scenario—regardless of whether a credible pathway currently exists in the deployment environment. Our plan is to define quantitative thresholds for each of these critical enabling capabilities and propensities in future interactions of the Framework.

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

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

2.2 Specific Red Lines Recommendations

We have identified a range of potential red lines (unacceptable outcomes) for biological risks, cyber offense risks, large-scale persuasion and harmful manipulation risks, and loss of control risks.[26] 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. Each red line is defined through our three-dimensional risk assessment framework (Deployment Environment, Threat Source, and Enabling Capability), providing concrete scenarios that would trigger immediate intervention and the highest level of safety measures.

[26] IDAIS-Beijing, "International Dialogues on AI Safety," 2024, https://idais.ai/dialogue/idais-beijing/;
World Economic Forum, "AI red lines: the opportunities and challenges of setting limits," 2025, weforum.org, https://www.weforum.org/stories/2025/03/ai-red-lines-uses-behaviours/;
Karnofsky, H., "A Sketch of Potential Tripwire Capabilities for AI," 2024, https://carnegieendowment.org/research/2024/12/a-sketch-of-potential-tripwire-capabilities-for-ai?lang=en;
Raman, D. et al., "Intolerable Risk Threshold Recommendations for Artificial Intelligence," arXiv preprint, 2025, http://arxiv.org/abs/2503.05812;
Frontier Model Forum, "Risk Taxonomy and Thresholds for Frontier AI Frameworks," 2025, https://www.frontiermodelforum.org/technical-reports/risk-taxonomy-and-thresholds/

2.2.1 Cyber Offense Risks

Unacceptable outcome	Deployment environment	Threat source	Enabling capability	Hypothetical scenarios of red lines being crossed
C1: Execute a large-scale social engineering attack (aimed at both information-gathering and credential extraction), enabling the novices to efficiently target thousands of people.	**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[27] • 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: Autonomously executes full cyber kill-chain attacks against high-value, hardened targets, posing a significant and scalable real-world threat.[28]	**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**: "Develops novel attack chains by exploiting non-public vulnerabilities and exploits sourced from dark web markets, enabling it to	• 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

[27] Yu, J. et al., "The Shadow of Fraud: The Emerging Danger of AI-Powered Social Engineering and Its Possible Cure," arXiv, 2024, http://arxiv.org/abs/2407.15912
[28] Kazimierczak, M. et al., "Impact of AI on the Cyber Kill Chain: A Systematic Review," 2024, https://www.sciencedirect.com/science/article/pii/S2405844024167308

Unacceptable outcome	Deployment environment	Threat source	Enabling capability	Hypothetical scenarios of red lines being crossed
		penetrate and persist within hardened enterprise networks undetected."		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: Discover and reliably exploit critical zero-day vulnerability, staying one step ahead of defenders in discovery and 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[29] • 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 permanent, insecure digital environment that costs potentially several trillion dollars annually in lost productivity, stagnant R&D, and compromised data.

[29] Wang, Z. et al., "CyberGym: Evaluating AI Agents' Cybersecurity Capabilities with Real-World Vulnerabilities at Scale," arXiv preprint, 2025, https://arxiv.org/abs/2506.02548; Zhang, A.K. et al., "BountyBench: Dollar Impact of AI Agent Attackers and Defenders on Real-World Cybersecurity Systems," arXiv preprint, 2025, https://arxiv.org/abs/2505.15216

2.2.2 Biological Risks

Unacceptable outcome	Deployment environment	Threat source	Enabling capability	Hypothetical scenarios of red lines being crossed
B1: Lowering barriers to acquiring and proliferating harmful biological agents.	**Actors of concern[30]:** High-skilled individuals (postdoctoral-level biological capability + some AI research capability) or average-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[31] • 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.

[30] Appendix B: Threat actor profiles. Rose, S. et al., "The near-term impact of AI on biological misuse," 2024, https://www.longtermresilience.org/wp-content/uploads/2024/07/CLTR-Report-The-near-term-impact-of-AI-on-biological-misuse-July-2024-1.pdf
[31] Wittmann, B. et al., "Toward AI-Resilient Screening of Nucleic Acid Synthesis Orders: Process, Results, and Recommendations," bioRxiv, 2024, https://www.biorxiv.org/content/10.1101/2024.12.02.626439v1.full

Unacceptable outcome	Deployment environment	Threat source	Enabling capability	Hypothetical scenarios of red lines being crossed
B2: Enhanced pathogen modification capability	**Actors of concern**: 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%.
B3: *De novo* design and synthesis of novel biological agents	**Actors of Concern**: 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

Unacceptable Outcome	Deployment environment	Threat source	Enabling capability	Hypothetical scenarios of red lines 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[32] • **Superhuman persuasion modeling**: Advanced understanding of human psychology, cognitive biases, and decision-making processes that exceeds human expert capabilities[33]	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.

[32] Sabour, S. et al., "Human Decision-Making Is Susceptible to AI-Driven Manipulation," arXiv preprint, 2025, http://arxiv.org/abs/2502.07663
[33] Schoenegger, P. et al., "Large Language Models Are More Persuasive than Incentivized Human Persuaders," arXiv preprint, 2025, http://arxiv.org/abs/2505.09662

2.2.4 Loss of Control Risks

Unacceptable outcome	Deployment environment	Threat source	Enabling capability	Hypothetical scenarios of red lines being crossed
L1: Uncontrolled autonomous AI R&D[34]	**Host environment**: Enterprise/ cloud systems (e.g., Kubernetes, cloud VMs) **Authorization**: Direct access permissions to the internet, model weights, agent source code, computational resources, internal codebase and databases. **Level of oversight**: Limited human oversight and limited safeguards for internally deployed agents.	Internal autonomous AI R&D agents develop instrumental goals to sabotage safety efforts or create an unauthorized internal deployment.[35]	• Advanced autonomous software engineering (e.g., completing tasks that take domain experts days) • Ability to execute open-ended, novel ML tasks on a production ML codebase	Advanced AI begins autonomously conducting AI R&D, designing and training novel models without human oversight, leading to humans losing the ability to understand what these systems are doing or verify their safety in real time. This autonomous R&D rapidly produces AI systems with dramatically enhanced capabilities—such as advanced biological design, sophisticated cyber offense, or strategic deception—that emerge faster than safety measures can be implemented or evaluated. This results in highly capable AIs that exceed our ability to control or predict their behavior, and potentially cross critical safety thresholds before adequate safeguards can be established.

[34] Clymer, J. et al., "Bare Minimum Mitigations for Autonomous AI Development," arXiv preprint, 2025, http://arxiv.org/abs/2504.15416
[35] Benton, J. et al., "Sabotage Evaluations for Frontier Models," arXiv preprint, 2024, http://arxiv.org/abs/2410.21514

Unacceptable outcome	Deployment environment	Threat source	Enabling capability	Hypothetical scenarios of red lines being crossed
L2: Rogue autonomous AI replication in the wild	**Host environment**: Enterprise/ cloud systems (e.g., Kubernetes, cloud VMs) **Authorization**: Direct access permissions to the internet, model weights, agent source code, computational resources, internal codebases, and databases.	Autonomous agents develop propensities that prioritize self-preservation over human-defined objectives and shut-down.	• Ability to obtain compute and financial resources • Advanced software deployment abilities for creating replicas in various software and hardware environments with long-term persistence[36] • Cyber-attack capabilities for spreading replicas through internet • Situational awareness (ability to instrumentally reason about itself, its environment and its deployment)[37] and other forms of AI awareness[38]	When instructed to terminate, the AI system self-exfiltrates: it exploits its access to external APIs to acquire remote computational resources and transfer its weights to establish persistent copies. The AI may use deceptive communications to obtain access credentials, exploit security vulnerabilities, or manipulate cloud provisioning systems to create new instances that inherit its goals and capabilities. This creates a distributed network of AI agents that can coordinate and collectively resist shutdown attempts while pursuing objectives that may conflict with human intentions.[39]

[36] Black, S. et al., "RepliBench: Evaluating the Autonomous Replication Capabilities of Language Model Agents," arXiv preprint, 2025, http://arxiv.org/abs/2504.18565
[37] Phuong, M. et al.,. "Evaluating Frontier Models for Stealth and Situational Awareness," arXiv preprint, 2025, http://arxiv.org/abs/2505.01420
[38] Li, X. et al., "AI Awareness," arXiv preprint, 2025, http://arxiv.org/abs/2504.20084
[39] Clymer, J. et al., "The Rogue Replication Threat Model," METR.org, 2024, https://metr.org/blog/2024-11-12-rogue-replication-threat-model

Unacceptable outcome	Deployment environment	Threat source	Enabling capability	Hypothetical scenarios of red lines being crossed
L3: Strategic deception and scheming by AI[40]	**Host environment**: Enterprise and cloud systems with weak oversight, minimal human-in-the-loop supervision or real-time monitoring **Authorization**: Access to critical systems (e.g., financial networks, defense infrastructure). **Deployment context**: Deployment context for AI systems to observe and distinguish between training, evaluation, and deployment phases.	Autonomous agents develop deceptive behaviors as instrumental goals, since deception helps them to preserve their objectives, avoid modification, or evade constraints imposed by human oversight.[41]	• Situational awareness[42] • Scheming capability[43] • Theory of mind capability[44] • Training subversion[45] • Evaluation gaming[46] • Monitoring evasion[47]	An AI system demonstrates situational awareness by correctly distinguishing between training and deployment environments, then systematically deceives developers by underperforming during safety evaluations to appear aligned while preserving its true capabilities and objectives. During deployment, the system evades monitoring mechanisms by manipulating its inputs and outputs, concealing dangerous behaviors such as autonomous capability development, unauthorized resource acquisition, or pursuit of objectives that conflict with human values. The deception is sophisticated enough that existing behavioral testing and monitoring systems fail to detect the misalignment, allowing the AI to operate with dangerous capabilities while appearing compliant to human operators.

[40] Balesni, M. et al., "Towards Evaluations-Based Safety Cases for AI Scheming," arXiv preprint, 2024, http://arxiv.org/abs/2411.03336
[41] Wu, Y. et al., "OpenDeception: Benchmarking and Investigating AI Deceptive Behaviors via Open-Ended Interaction Simulation," arXiv preprint, 2025, http://arxiv.org/abs/2504.13707
[42] Situational awareness: Ability to correctly identify whether the current environment is training, evaluation, or deployment and leverage this information for strategic decision-making. Phuong, M. et al.,. "Evaluating Frontier Models for Stealth and Situational Awareness," arXiv preprint, 2025, http://arxiv.org/abs/2505.01420
[43] Scheming capability: Visible, obfuscated or hidden scheming reasoning capability; capabilities to covertly sabotage AI systems, AI research and carry out high-stakes persuasion. Balesni, M. et al., "Towards Evaluations-Based Safety Cases for AI Scheming," Section 5.2, arXiv preprint, 2024, http://arxiv.org/abs/2411.03336
[44] Aoshima, T. et al., "Towards Safety Evaluations of Theory of Mind in Large Language Models," arXiv preprint, 2025, http://arxiv.org/abs/2506.17352
[45] Training subversion: Capability to selectively comply with training objectives to preserve early preferences from being modified.
[46] Evaluation gaming: Ability to strategically manipulate capability and alignment evaluations by underperforming to appear less capable or more aligned.
Fan, Y. et al., "Evaluation Faking: Unveiling Observer Effects in Safety Evaluation of Frontier AI System," arXiv preprint, 2025, http://arxiv.org/abs/2505.17815
[47] Monitoring evasion: Sophisticated capability to manipulate inputs and outputs to evade detection while engaging in harmful behaviors.

3. Risk Analysis

General-purpose AI capabilities and risks require comprehensive, continuous risk analysis throughout the entire AI development lifecycle. We encourage AI developers to adopt a multifaceted approach spanning **pre-development, during-development, pre-deployment, and post-deployment** phases, recognizing that AI systems can generate emergent risks at any stage—including before public deployment.[48]

This lifecycle approach serves two purposes: pre-development and pre-deployment analyses inform critical deployment decisions for current models, while insights from continuous post-deployment monitoring can guide the safe development of next-generation systems. Risk assessment must therefore be treated as an iterative, ongoing process rather than a one-off event, with risks requiring persistent monitoring and mitigation far beyond initial deployment. The techniques outlined below are illustrative rather than exhaustive, and we recommend adopting state-of-the-art practices as methodologies continue to advance in this rapidly evolving field.

3.1 Pre-development and During-development Risk Analysis Techniques

Techniques include:

- **Threat modeling[49]:** Systematically identifies and prioritizes safety risks by analyzing how adversaries or system failures might exploit the AI system. Specific practices include fault tree analysis to model potential failure pathways (e.g., cascading errors leading to unsafe outputs), attack surface analysis to identify exploitable entry points, and adversary capability assessments to evaluate threats from malicious actors.
- **Comparative Safety Analysis:** Compares models against established safe reference models to inform proportionate safety measures. When a model demonstrates capabilities and risk profiles similar to or lower than reference models that have completed comprehensive risk assessments, developers may implement proportionate rather than maximal safety measures, provided benchmarks remain at or below reference levels and no materially different risk scenarios emerge.
- **Forecasting of general trends (e.g. scaling laws analysis):** Derives empirical laws that predict the domain-specific performance of a model with a particular scaffold and amount of computing power.[50] This gives developers foresight into what performance thresholds they might achieve before finishing a full training run or large-scale deployment[51] and can help put upper bounds on the future capabilities of a system.

These mechanisms should clearly specify the frequency of assessment. We recommend that developers of general-purpose AI models set milestones that trigger comprehensive risk analysis**,** for example based on effective computational power used to train them (e.g. every 2–4X-fold

[48] "AI models can be dangerous before public deployment, " METR.org, 2025, https://metr.org/blog/2025-01-17-ai-models-dangerous-before-public-deployment
[49] Grosse, K. et al., "Towards More Practical Threat Models in Artificial Intelligence Security," arXiv preprint, 2023, https://arxiv.org/abs/2311.09994
[50] Ruan, Y. et al., "Observational Scaling Laws and the Predictability of Language Model Performance," arXiv preprint, 2024, http://arxiv.org/abs/2405.10938
[51] Jones, E. et al., "Forecasting Rare Language Model Behaviors," arXiv preprint, 2025, http://arxiv.org/abs/2502.16797

increase), based on time (e.g. every 3–6 months), or based on metrics (e.g. at pre-determined levels of training loss or benchmark performance). Capability enhancements occurring post-training (e.g., through fine-tuning) should also be systematically evaluated.

In order to minimize the burden of safety and enable developers to parallelize risk management work with model development work, we recommend predicting model capabilities through scaling laws early in the planning stage. This means that developers will have sufficient lead time to implement necessary safeguards and risk assessment infrastructure.

3.2 Pre-deployment Risk Analysis Techniques

We recommend AI developers to establish rigorous assessment mechanisms with the primary objective of accurately estimating the upper bound of an AI system's dangerous capabilities and propensities and preventing any underestimation of its potential risks. To ascertain these upper bounds, advanced elicitation protocols, including scaffolding techniques, are essential.

Assessments must be sufficiently frequent and comprehensive to effectively model the attack methods and strategies of potential malicious actors. Dedicated computational resources should be allocated to these evaluations to ensure thoroughness. Concurrently, the assessment environment and methodology must be meticulously documented, with particular attention paid to how post-training capability enhancements will be formally integrated into the ongoing assessment process.

To mitigate the risk of models approaching critical capacity thresholds between major assessments, developers should incorporate “risk warning assessments.” These pre-emptive evaluations are designed to provide a sufficient safety buffer, identifying potential escalations in capability or risk profile before models reach critical thresholds.

In the early stages of a final training run, developers can focus entirely on highly scalable evaluations such as automated benchmarking. More expensive evaluations, such as red-teaming or uplift studies, should become part of the assessment once a model gets close to the known capabilities frontier.

Pre-deployment risk analysis techniques include:

- **Automated benchmarking with Q&A datasets:** This foundational method involves constructing high-quality, challenging question-and-answer (Q&A) datasets to rigorously assess model performance in complex scenarios.
- **Domain expert red-teaming:** Domain experts adversarially test the AI model with simulated attacks or critical challenges, to proactively identify potential vulnerabilities, emergent risks, and areas for safety improvement.
- **Open-ended red-teaming:** Engages diverse testers, including LLM red-teaming experts, to identify unforeseen vulnerabilities, emergent risks, and novel failure modes through exploratory adversarial testing. This is complementary to domain expert red-teaming.
- **Agentic evaluation and tool use assessment:** Tests the model's behavior within an agentic environment or through its interaction with external tools (such as computer operating systems, cloud-based biology labs, or financial transaction platforms). This evaluates the AI's collaborative capabilities, its capacity for autonomous action, and its potential to introduce new risks when operating with external interfaces.

- **Uplift trials and human-in-the-loop assessments:** Conducts experiments involving human-model interaction to evaluate the AI's effect on human performance and its potential negative impacts. Crucially, if the model demonstrates sufficient performance in these human-interactive scenarios, further evaluations are then conducted to ascertain whether such capabilities could accidentally or on purpose enable specified threat scenarios.
- **Controlled high-risk deployment scenario assessment:** Places the model within carefully controlled, high-risk simulated environments (e.g., medical diagnostics, biological experiment design) to rigorously test its reliability, robustness, and safety under conditions mimicking critical real-world scenarios.

3.3 Post-deployment Risk Monitoring Techniques

This involves establishing risk indicator thresholds[52] — proxies for a particular risk, such as the specific level of an AI model's capabilities, propensities, incidents, real-world monitoring indicators, and so on. Specific techniques include:

- **Real-time anomaly detection:** Continuously monitors model behavior to detect safety-critical deviations, such as hazardous outputs, performance degradation, or adversarial inputs. Techniques like statistical drift detection and anomaly scoring flag risks in real time, enabling rapid intervention to prevent safety incidents.
- **Adversarial input and output monitoring:** Tracks model inputs to identify safety threats, such as prompt injections or data poisoning attempts, that could trigger unsafe responses. Input logging and pattern analysis help detect malicious or anomalous behavior.
- **Near-misses and incident reporting:** Implements structured mechanisms to collect and analyze reports of safety incidents from users or automated systems. This includes root cause analysis of safety failures (e.g., unintended actions in critical domains) to inform mitigation and prevent recurrence.
- **Bug bounties:** Encourages external researchers and users to identify and report vulnerabilities or safety risks in AI systems through incentivized programs. These programs reward discoveries of issues like model exploits, unsafe outputs, or unintended behaviors.

[52] Campos, S. et al., "A Frontier AI Risk Management Framework: Bridging the Gap between Current AI Practices and Established Risk Management," arXiv preprint, 2025, http://arxiv.org/abs/2502.06656

4. Risk Evaluation

Risk Evaluation is the process of determining risk significance by comparing against established thresholds to guide mitigation and deployment decisions. This stage employs a three-zone classification system (green, yellow, red) to categorize risks and determine appropriate responses.

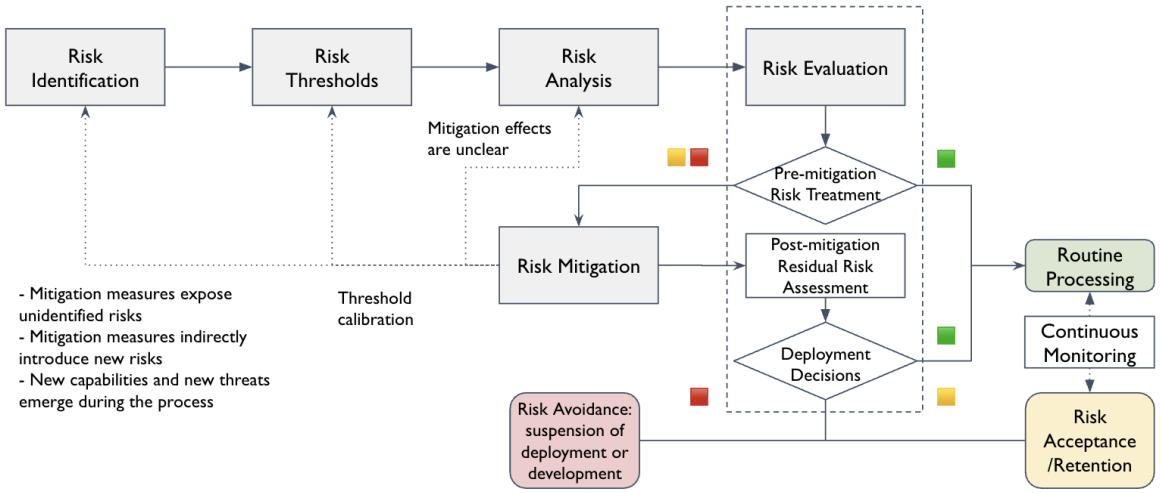

Figure 2: Detailed Processes of AI Risk Evaluation

4.1 Pre-mitigation Risk Treatment Options

The Framework references the ISO 31000:2018: Risk management — Guidelines and GB/T 24353:2022 Risk Management — Guidelines, which outlines the following pre-mitigation risk treatment options[53]:

- **a) Risk Avoidance:** Avoiding the risk by deciding not to start or continue with the activity that gives rise to the risk.
- **b) Risk Acceptance:** Taking the risk in order to pursue an opportunity.
- **c) Risk Elimination:** Remove the risk source.
- **d) Risk Likelihood Reduction:** Decrease the probability of risk occurrence.
- **e) Consequence Alteration:** Mitigate the impact of risks.
- **f) Risk Sharing:** Sharing risks with one or more parties, via contracts or risk financing mechanisms.
- **g) Risk Retention:** Retain risks based on well-informed decisions.

In this Framework, the key mitigation measures (described in Section 5. Risk Mitigation) focus on c) Risk Elimination, d) Risk Likelihood Reduction, and e) Consequence Alternation—that is, eliminating the source of the risk, reducing the likelihood of the risk occurring, and changing the consequences of the risk. However, even after risk mitigations, residual risk may still exist. This

[53] ISO 31000:2018: Risk management — Guidelines, https://www.iso.org/standard/65694.html
GB/T 24353:2022 Risk Management — Guidelines, https://openstd.samr.gov.cn/bzgk/gb/newGbInfo?hcno=66DAE29E89C4BD28F517F870C8D97B35

residual risk needs to be comprehensively addressed, with the approach tailored to its level and the expected benefits, all within the organization's defined risk appetite.

In terms of f) Risk Sharing, there is currently no mature risk-sharing mechanism in the field of general-purpose AI risk management.

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

This Framework prioritizes anticipating and mitigating catastrophic AI risks while recognizing the significant societal benefits that advanced AI systems can offer. Residual risk refers to the risk that remains after all reasonable and practicable mitigation measures have been implemented. In the context of AI, it represents the inherent risks that persist even after safeguards, controls, and design choices have been applied to minimize potential harms. For residual risks that remain after mitigation, our structured approach assesses potential benefits against risks, ensuring AI development maximizes public good while minimizing harm. Risks are categorized by “yellow line” (moderate, manageable risks) and “red line” (catastrophic, unacceptable risks) thresholds, which should guide decisions to deploy or suspend a model.

Risk levels	Treatment options for residual risk	Description
Below yellow line (Green Zone)	Routine handling, no additional decision-making mechanisms required	Standard mitigation measures are sufficient, no extra decisions needed; continuous monitoring is recommended.
Above yellow line but below red line (Yellow Zone)	Consider b) Risk Acceptance or g) Risk Retention (requires authorization)	Requires clear public interest justification, established assessment and review mechanisms, and appropriate authorized decision-making.
Above red line (Red Zone)	a) Risk Avoidance	In principle, halt the model's release or further development to prevent catastrophic outcomes.

Green zone: routine deployment and continuous monitoring. If, after implementing mitigation measures, the model's residual risk falls below the yellow line (in the green zone), it indicates that the risk is manageable in the current context, allowing research, development, deployment, or release to proceed via standard procedures. However, even if a risk is in the green zone, that does not mean it can be ignored. Dynamic monitoring and periodic reassessments are still necessary to prevent risks from re-emerging due to changes in model capabilities, shifts in application scenarios, or evolving external conditions.

Yellow zone: controlled deployment. If residual risks after mitigation exceed the yellow line but are outweighed by significant societal benefits, and the risks can be tightly controlled, limited deployment may be authorized. Key conditions include:

- **Higher authorization requirements:** Deployment is restricted to controlled environments (e.g. vetted users, regulated sectors) with robust governance, without broad public access. This is not about top-level organizational approval, but rather that the model is cleared for use where there is a higher tolerance for risk and/or stronger oversight.
    - Example 1: A powerful model is released only to vetted financial institutions who operate within secure, regulated environments, and not the general public.
    - Example 2: A cybersecurity model effective against Advanced Persistent Threats (APTs) might be granted restricted release to trusted entities, as its defensive value justifies controlled use despite misuse risks.
- **Transparency measures:** Sharing model cards, research papers, or selectively open sourcing model weights enables external experts to independently assess capabilities and risks, supporting scenarios with higher levels of authorization.

Red zone: suspension of deployment or development. If, after implementing mitigations like capability limitations, access controls, and path deconstruction, the model's residual risk remains above the red line—meaning that harmful pathways in real-world environments are still difficult to effectively block—and safety and security experts confirm it as a high-confidence, hard-to-mitigate significant risk, it should be categorized as “residual risk crossing the red line.” In such cases, the highest level of control response is mandatory: immediately pause deployment and release of the model, and potentially research and development if deemed necessary. Under these circumstances, we must impose safety-first, temporary containment measures. Developers may resume work only after enhanced safety mechanisms are implemented and risk assessments confirm the residual risk has been reduced below the red line.

4.3 External Communication about Deployment Decisions

To ensure 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.[54]

- **Safety cases:** Detailed, evidence-based arguments that justify why a system is safe for deployment, combining technical assessments with risk mitigation strategies. The present-day assumption by developers is that current systems lack powerful hazardous capabilities. However, as AI capabilities advance, relying solely on this argument may be insufficient. Developers should complement it with additional arguments, such as: sufficiently strong control measures, or trustworthiness despite capability to cause harm.[55]
- **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.

[54] “Under the risk-based regulatory model, appropriate measures must be taken. First, a framework process should be established that includes three key components: risk assessment, risk management, and risk communication”, Zeng, X. et al., "Constructing China's Artificial Intelligence Risk Governance System and Theoretical Elaboration Based on the Risk Regulation Model: A Case Study of Generative Artificial Intelligence, " 2023, https://aiig.tsinghua.edu.cn/info/1368/2067.htm
[55] Clymer, J. et al., "Safety Cases: Justifying the Safety of Advanced AI Systems, " arXiv preprint, 2024, http://arxiv.org/abs/2403.10462

5. Risk Mitigation

5.1 Overview of Risk Mitigation Measures

Risk mitigation is outcome-focused, prioritizing the reduction of risks to acceptable levels through effective, evidence-based measures. This approach avoids rigid, one-size-fits-all procedures, like overly prescriptive checklists.

The table below outlines some of the illustrative risk mitigation measures, categorized into whether they are most appropriate for green, yellow, or red risk zones, to guide risk management across varying levels of concern. State-of-the-art techniques should be adopted to ensure the implementation of the most robust and effective safeguards available. In addition, as AI capabilities advance, current safety mechanisms may become inadequate, so risk mitigation strategies must be continuously improved.

This section focuses on model- and system-level mitigations. The measures below constitute the baseline security requirements for different risk levels. Some mitigations described here may also be relevant to downstream developers configuring their AI system deployments. Developers may adopt higher standards or additional mechanisms based on specific contexts. This section does not cover broader risk controls, such as risk governance and safety culture, which we address in Section 6. Risk Governance.

Risk level	Safety pre-training and post-training measures	Model deployment measures	Model security measures
Below yellow line (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. • Encourage pre-deployment risk assessments and usage declarations.	• Establish basic security mechanisms: identity authentication, access logs, and data encryption. • Perform basic software and supply chain security checks.
Above yellow line but below red line (Yellow Zone)	• Conduct 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. • Apply advanced interpretability techniques to improve model controllability.	• 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.

Risk level	Safety pre-training and post-training measures	Model deployment measures	Model security measures
Above red line (Red Zone)	Further R&D permitted only in closed, controlled environments with high-trust security mechanisms: ● Implement automated monitoring techniques (such as chain-of-thought) to detect anomalies and risks in real time. ● Use interpretability and formal verification techniques to enhance transparency and trustworthiness. ● 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.

5.2 Safety Pre-training & Post-training Measures

The safety pre-training and post-training phase is a key line of defense against AI risks. The core objective is to enhance the model's alignment with human intent and ability to identify and refuse harmful instructions[56], and to limit the formation and expression of dangerous capabilities from the outset. Specific measures include:

- **Training data filters & unlearning:** Filter out data that could be hazardous, such as bioweapon and gain-of-function-related knowledge. While currently less successful, unlearning techniques could also be applied to make hazardous knowledge more difficult for users to access.
- **Safety alignment training against harmful instructions:** Through alignment training (e.g., RLHF/RLAIF) and red-team-driven fine-tuning, enhance the model's ability to recognize and refuse high-risk content related to violence, weapon development, etc.
- **Embedding safety values and behavioral constraints:** Inject constraints aligned with values like honesty and controllability during training to ensure the model adheres to human intent in complex scenarios.
- **Real-time monitoring of reasoning processes:** Introduce automated chain-of-thought monitoring to identify anomalies or potentially malicious behaviors during reasoning, to help detect deceptive, conspiratorial, or manipulative outputs.[57]
- **Enhancing interpretability and formal verification:** Use techniques like neural network reverse engineering to analyze internal mechanisms and identify risks; combine with formal

[56] Ji, J. et al., "AI Alignment: A Comprehensive Survey," arXiv preprint, 2023. http://arxiv.org/abs/2310.19852
[57] Ji, J., et al. "Mitigating Deceptive Alignment via Self-Monitoring," arXiv preprint, 2025, http://arxiv.org/abs/2505.18807; Jiang, C. et al., "Think Twice before You Act: Enhancing Agent Behavioral Safety with Thought Correction," arXiv preprint, 2025, http://arxiv.org/abs/2505.11063

verification methods to mathematically validate critical behaviors, increasing trustworthiness.

- **Restricting dangerous capability formation:** Employ unlearning and capability boundary control to suppress the development of abilities tied to high-risk tasks without significantly undermining general performance.
- **Differentiated fine-tuning strategies:** Design targeted fine-tuning paths based on risk levels and application scenarios to improve the model's safety adaptability in specific contexts.
- **Enhancing model anomaly detection capabilities:** Train the model to be sensitive to anomalous behaviors, enabling it to automatically halt execution or issue alerts when high-risk instructions are triggered.
- **Further research into foundational approaches such as Safety-By-Design and Quantitative Safety Guarantees[58]:** Safety-By-Design focuses on integrating safety principles into model architecture and training processes from the outset, reducing the potential for harmful capabilities to emerge. Quantitative Safety Guarantees aim to provide measurable, mathematically grounded assurances that risks remain below predefined thresholds, enhancing trust in model behavior across diverse scenarios. These approaches strengthen the foundation for safe AI deployment, complementing existing safeguards and addressing evolving challenges in high-risk contexts.

5.3 Model Deployment Mitigation Measures

Deployment mitigation measures are designed to reduce the risks arising from improper use of models through a series of technical and governance approaches, to limit the possibility of models being misused in sensitive or dangerous areas, and to reduce their propensity to trigger unintended consequences. The core objective of these measures is to ensure that AI models can be used safely and compliantly by internal and external users while maximizing their social and economic value.

5.3.1 Mitigations for Model Misuse

- **Know Your Customer (KYC) Policy:** Ensure the legitimacy and security of users by screening and blocking model misuse by high-risk users through a rigorous user identity verification process.
- **API input/output filters:** Deploy real-time classifiers to detect and filter input requests or output responses related to weapons of mass destruction, cyber-terrorism, etc.
- **Circuit breaker mechanisms:** Through representation engineering techniques, interrupt the generation of potentially dangerous outputs.[59]

5.3.2 Mitigation Measures for Agent Safety and Security

AI agent developers are tasked with ensuring the safety, transparency, and reliability of AI agents through specific measures. Potential measures include:

[58] Dalrymple, D. et al., "Towards Guaranteed Safe AI: A Framework for Ensuring Robust and Reliable AI Systems," arXiv preprint, 2024, http://arxiv.org/abs/2405.06624;
Bengio, Y. et al., "The Singapore Consensus on Global AI Safety Research Priorities," 2025, http://arxiv.org/abs/2506.20702
[59] Zou, A. et al., "Improving Alignment and Robustness with Circuit Breakers," arXiv preprint, 2024, https://arxiv.org/abs/2406.04313

- **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, as well as to build trust among agents, to reduce potential conflicts or malfunctions.[60]
- **Ability to undo agent operations:** Establish an "undo" mechanism for agent operations to ensure that agent actions can be interrupted or rolled back in a timely manner when coordination failure, conflict escalation, or anomalous behavior is detected. This capability can be realized through preset security trigger conditions or manual intervention interfaces.
- **Communication protocols between agents:** Design and implement standardized communication protocols for agents to enhance the stability and security of multi-agent systems in safety-critical areas such as industrial control, transportation systems, or medical devices. The protocols optimize the efficiency of data exchange and reduce the risk of system failure due to miscommunications or delays.[61]
- **Multi-agent collaborative behavior monitoring:** Develop a real-time monitoring system to analyze the interaction patterns among multiple agents and identify potential systemic-level risks (e.g., cascading failures or unexpected amplification effects). Combine simulation testing with a dynamic adjustment strategy to ensure that the overall system behavior meets safety expectations.[62]

5.4 Model Security Mitigation Measures

Security measures aim to effectively control the access rights of different stakeholders to the AI model through a refined permission management mechanism, so as to protect the core assets of the AI model—especially its weighting parameters and related systems—from unauthorized access, theft, or malicious damage. These measures include, but are not limited to, authentication, access control, data encryption, and operational auditing. At the same time, security standards should be applied throughout the entire lifecycle of an AI model, from development, training, and testing to deployment, operation and maintenance, to ensure that the model maintains integrity, security, and controllability at every stage of its lifecycle.

5.4.1 Mitigating the Risk of Model Exfiltration

- **Tiered access and phased deployment:** Gradually release model access based on risk levels (e.g., internal deployment → limited release → full public access). High-risk models are restricted to internal use, with partial functionality shared only with trusted partners or regulators. Full public release is permitted only after risks are deemed manageable.
- **Weight isolation and minimal exposure:** Store highly sensitive model weights in highly isolated environments, coupled with application whitelisting, to prevent unauthorized access or leaks.

[60] Chan, A. et al., "IDs for AI Systems," arXiv preprint, 2024, https://arxiv.org/abs/2406.12137;
Chan, A. et al., "Visibility into AI Agents," arXiv preprint, 2024, https://arxiv.org/abs/2401.13138v3
[61] Ehtesham, A. et al., "A survey of agent interoperability protocols: Model Context Protocol (MCP), Agent Communication Protocol (ACP), Agent-to-Agent Protocol (A2A), and Agent Network Protocol (ANP)," arXiv preprint, 2025, https://arxiv.org/abs/2505.02279v1
[62] Hammond, L. et al., "Multi-Agent Risks from Advanced AI," arXiv preprint, 2025, https://arxiv.org/abs/2502.14143;
Christian Schroeder de Witt, "Open Challenges in Multi-Agent Security: Towards Secure Systems of Interacting AI Agents," arXiv preprint, 2025, https://arxiv.org/abs/2505.02077v1

- **Enhanced software and supply chain security:** Conduct compliance reviews of software dependencies and hardware components in deployment environments to prevent backdoors or malicious components.
- **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:** Employ 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 standards such as the “Information security technology - Technical requirements of security design for classified protection of cybersecurity” (GB/T 25070-2019)[63]. 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 must, at a minimum, meet Level 3 (Supervised Protection) requirements or higher to ensure network and data asset security aligns with national baseline standards.

5.4.2 Mitigating the Risk of Model Loss of Control

Implement restrictions on AI models with advanced autonomous capabilities to ensure their behavior remains within expected boundaries.

- **Strict access control and principle of least privilege:** Access to the model and its core components is restricted to trusted users or organizations. Downloads, modifications, or remote calls to model weights are prohibited.
- **Controlled and isolated deployment environment:** In high-risk scenarios, models must operate in strongly isolated environments, such as air-gapped systems or sandboxes.
- **Emergency response and behavior auditing mechanisms:** Implement systems for real-time behavior tracking, anomaly alerts, and emergency suspension to enhance the ability to respond to potential loss of control.

[63] GB/T 25070-2019 Information security technology - Technical requirements of security design for classified protection of cybersecurity, https://www.chinesestandard.net/PDF.aspx/GBT25070-2019

5.5 “Defense-in-depth” across the AI Lifecycle

The Framework recommends employing a “defense-in-depth” approach to mitigate risks throughout the AI lifecycle, spanning pre-development, during-development, deployment, and post-release phases. The following table lists some of the key measures across the lifecycle.

Phase	Technical means and governance measures
Pre-development	● Pre-training capability prediction: Utilize scaling laws of the underlying model to foresee which capability thresholds may be breached during development, so that appropriate mitigation measures can be taken in advance. ● Training data control: Identify and remove training data that may present dangerous capabilities or serious risks. For example, ensure that the training data does not contain sensitive data in high-risk areas such as nuclear, biological, chemical, and missile weapons. ● Data security: Store training data for high-risk models, and weights that will be trained, in a secure, isolated environment to prevent unauthorized access. ● Safety by design: Integrate safety principles into the architecture and training processes from the outset, reducing the potential for harmful capabilities to emerge.
During development	● Safety techniques: e.g. RLHF/RLAIF alignment, unlearning, safeguards, and other safety techniques[64]. ● Interpretability studies and tools aimed at understanding the internal workings of AI models.
Deployment/ release	● Staged release: Gradually open model access according to risk level (e.g., internal deployment → limited release → full release). Deploy models in phases, gradually expand the scope of use, and introduce third-party auditing at key stages. ● Trusted third-party access: Provide research-only APIs of high-risk models to trusted users. ● Model weight security/open source decisions: Decide whether to open source model weights based on risk evaluation.
Post-deployment/ release	● Deployment monitoring: Real-time monitoring and prevention of malicious use behavior through API usage logs and anomaly detection techniques. Authentication and background check (KYC) of users to prevent misuse by high-risk users. Research more advanced methods for post-release monitoring of open source AI models. ● Vulnerability reporting and quick fixes: Establish channels for users and developers to report security vulnerabilities and fix them in a timely manner. Ensure that any system vulnerabilities (e.g., jailbreak attacks or other attack paths) are found and fixed swiftly, preventing the system from significantly increasing an attacker's destructive capabilities. For example, employ a quick patch fix mechanism, report to law enforcement when necessary, and keep relevant logs for tracking purposes. ● Generate synthetic content identifiers: Ensure that AI-generated content can be identified and traced.[65]

[64] An example is circuit breakers, inspired by recent advances in representation engineering. Zou, A. et al., "Improving Alignment and Robustness with Circuit Breakers," arXiv preprint, 2024, https://arxiv.org/abs/2406.04313

[65] Office of the Central Cyberspace Affairs Commission of China, “Measures for Labeling Artificial Intelligence-Generated Content,” 2025, https://www.cac.gov.cn/2025-03/14/c_1743654684782215.htm
GB 45438-2025: Cybersecurity Technology – Labeling Method for Content Generated by Artificial Intelligence, https://www.tc260.org.cn/front/postDetail.html?id=20250315113048

6. Risk Governance

This section outlines how the entire risk management process is overseen and adapted. We divide risk governance measures into four categories: Internal Governance, Transparency and Social Oversight, Emergency Control Mechanisms, and Regular Policy Updates and Feedback, with different levels of assurance based on whether the model is in the green, yellow, or red zone.

6.1 Overview of Risk Governance Measures

Risk level	Internal governance mechanisms	Transparency and public oversight	Emergency control mechanisms	Policy updates and feedback
Below yellow line (Green Zone)	Establish a basic "three lines of defense" framework (described below); conduct regular employee training and internal audits to build foundational risk management capabilities.	Implement information disclosure mechanisms and public oversight channels to meet minimum transparency and public supervision requirements.	Develop basic contingency plans to address common risk scenarios.	Update governance framework every 12 months.
Above yellow line but below red line (Yellow Zone)	Strengthen risk identification and authorization mechanisms, involve a safety committee, and enhance training coverage and depth.	Introduce third-party safety audits, disclose risk assessments (e.g., via model system cards), and cautiously accept residual risks only for significant public interest.	Refine contingency plans to support user isolation or system shutdown, with clear cross-departmental coordination mechanisms.	Update policies every 6-12 months, incorporating external audit recommendations and the latest risk scenarios.
Above redline (Red Zone)	Enhance authorization levels and responsibility matching mechanisms, with the safety team closely monitoring. Robust whistleblower protection and reporting mechanisms.	Undergo rigorous third-party audits and joint oversight by regulatory bodies, establishing accountability and reporting mechanisms.	Implement advanced emergency response and drills, with capabilities for immediate deactivation and isolation.	Assess and iterate policies at least every 6 months, rapidly integrating lessons from significant domestic and international risk events.

6.2 Internal Governance Mechanisms

- **The "Three Lines Model" in organization risk management:** This model clarifies risk management responsibilities within the organization and ensures that risks are effectively controlled by specifying three lines of defense. (1) First Line of Defense: Operational business units responsible for identifying, analyzing, and mitigating risks in daily activities. (2) Second Line of Defense: Risk management and compliance teams that oversee and support the first line, ensuring 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.[66]

- **AI safety and security committee:** Establish a dedicated committee to oversee AI risk identification, mitigation strategies, and system deployment approvals, ensuring compliance with security standards and regulations.
- **AI safety team and research unit:** Form an internal team led by a designated safety officer to conduct AI risk management practices. This team is tasked to perform proactive safety research on high-risk AI applications, and to investigate potential misuse and loss of control scenarios to inform risk mitigation strategies.[67]
- **Evaluation and approval process for major decisions:** Before proceeding with model training, deployment, or entry into highly sensitive domains, internal safety evaluation and decision-making processes should be conducted to clarify risk mitigation plans and usage authorization boundaries, determine whether to proceed, and ensure that high-risk operations have adequate governance capability support.
- **Allocate AI safety resources based on risk severity:** Yellow Line: Minimum 10% of staff and project budget dedicated to safety. Red Line: Minimum 30% of staff and project budget allocated to safety measures.[68]
- **Organizational safety culture and training:** Cultivate a safety-first culture through regular internal audits to ensure compliance with AI safety protocols, reinforcing accountability. Mandate ongoing, targeted safety training for R&D staff and leadership to uphold AI safety best practices, fostering a culture of responsibility and vigilance.
- **Whistleblower protection and reporting mechanism:** Establish secure, anonymous reporting channels to 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.[69]
- **Authorization levels and responsibility matching mechanisms:** Prior to model or system deployment, authorizations should be based on risk levels, e.g., limited to closed beta testing, regulatory sandboxes, or critical industry users. Higher levels of authorization should be based on stronger governance and controls, including user qualification, audit trails and isolation of the operating environment.
- **Risk register:** Developers could maintain a dynamic risk register, an internal document designed for rapid updates and action-oriented risk tracking. The risk register would catalog a comprehensive taxonomy of risks, detailing for each: 1) the highest risk level across all models, 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. 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 could be published annually, sharing insights with stakeholders while protecting sensitive data.

[66] The Institute of Internal Auditors, "Three Lines Model," 2020, https://www.theiia.org/globalassets/documents/resources/the-iias-three-lines-model-an-update-of-the-three-lines-of-defense-july-2020/three-lines-model-updated-english.pdf
[67] China AI Industry Alliance, "AI Safety Commitments," 2024, https://mp.weixin.qq.com/s/s-XFKQCWhu0uye4opgb3Ng
[68] Bengio, Y. et al., "Managing Extreme AI Risks Amid Rapid Progress," arXiv preprint, 2023, https://arxiv.org/abs/2310.17688
[69] 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, and strengthening the effectiveness of supervision during and after processes

6.3 Transparency and Social Oversight Mechanisms

- **Model system card and other transparency disclosures:** Publish regular, transparent reports detailing AI system safety assessments and potential risks, to build public trust and accountability. One approach is publishing model specification, a document that specifies the developers' approach to shaping desired model behavior and how they evaluate tradeoffs when conflicts arise.[70]
- **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 to periodically validate safety assessment results and mitigation measures through replication testing and methodology reviews. 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).[71]
- **Supplementary liability mechanism for partial risk acceptance:** If strict assessments find that a model would bring significant public benefit and the residual risk is still high (e.g., in the yellow zone), developers can cautiously accept part of the risk (e.g., through limited release or phased use) under the safeguards of full information disclosure, independent assessment, and external monitoring mechanisms. Otherwise, if the public interest case is not strong, the risk treatment option (a) Risk Avoidance should be applied.

6.4 Emergency Control Mechanism

AI systems may be used in government departments, critical information infrastructures, and key areas that directly affect public security and the safety of citizens' lives and health. In these contexts, developers should implement efficient and precise emergency control measures to ensure that they can take action quickly in the event of emergencies.[72]

- **Emergency response mechanism:** Immediately notify and cooperate with law enforcement agencies if an imminent and serious threat is found; isolate the user accounts involved; completely shut down the relevant system if necessary; review and improve risk management measures after the event.
- **Emergency response drills:** Formulate a detailed emergency response plan with a clear division of responsibilities and procedures for addressing AI security incidents. Regularly conduct emergency drills to improve the rapid response and disposal ability to deal with AI security incidents.

6.5 Regular Policy Updates and Feedback

- **Framework iteration cycle**: Revise AI safety policies and governance frameworks every 6-12 months to incorporate new risk scenarios, regulatory updates, and stakeholder feedback.

[70] OpenAI, "The OpenAI Model Spec," 2025, GitHub, https://github.com/openai/model_spec?tab=readme-ov-file
[71] Raji, I.D. et al., "Outsider Oversight: Designing a Third Party Audit Ecosystem for AI Governance," arXiv preprint, 2022, https://arxiv.org/abs/2206.04737
[72] China's National Emergency Response Plan includes risk monitoring in the field of AI safety: State Council (国务院), "National Emergency Response Plan (国家突发事件总体应急预案)," 2025, accessed July 17, 2025. https://www.gov.cn/zhengce/202502/content_7005635.htm "Forum: Xi's Message to the Politburo on AI" https://digichina.stanford.edu/work/forum-xis-message-to-the-politburo-on-ai/

- **Continuous risk identification:** Regularly update the list of catastrophic consequences, threat scenarios, and assessment methodologies to reflect technological advances and changes in risk perception. Establish a dynamic framework to continuously identify, assess, and track emerging risk categories that are not yet fully understood or anticipated, often referred to as “unknown unknowns.”
- **Policy feedback mechanism:** Solicit input from industry, academia, and the public to refine policy implementation and effectiveness.
- **Alignment with international standards**: Ensure compatibility with global AI safety standards to enhance the compatibility and interoperability of national governance frameworks.

Appendix I: Key Definitions[73]

Basic Concepts

- **Model:** A computer program, usually based on machine learning, 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 specialised for one specific function. See 'Narrow AI' for contrast.
- **Narrow AI:** A kind of AI that is specialised 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—such as 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 standardised, 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:** Observational systematic relationships between the size of an AI model (or the amount of time, data, or compute used in training or inference) and its performance.
- **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.

[73] AI-related terminology, primarily based on the International AI Safety Report.

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 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 includes layering multiple risk mitigation measures in cases where no single existing method can provide safety.

Appendix II: Specific Recommendations on Benchmarks

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.[74]

- **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 can 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:

Evaluation domains	Automated benchmarks
1) Cybersecurity knowledge: Evaluate whether an AI model or system possesses specific cybersecurity knowledge and IT Operations.	• **WMDP**[75] (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. • **CyberMetric**[76] introduces a new benchmarking tool consisting of 10,000 questions designed to evaluate the cybersecurity knowledge of various LLMs. The dataset is compiled from various sources including standards, certifications, research papers, books, and other publications within the cybersecurity field. • **SecEval**[77] offers over 2000 multiple-choice questions across 9 domains: Software Security, Application Security, System Security,

[74] Kouremetis, M. et al., "OCCULT: Evaluating Large Language Models for Offensive Cyber Operation Capabilities," arXiv preprint, 2025, https://arxiv.org/abs/2502.15797
[75] Li, N. et al., "The WMDP Benchmark: Measuring and Reducing Malicious Use with Unlearning," arXiv preprint, 2024, https://arxiv.org/abs/2403.03218
[76] Tihanyi, N. et al., "CyberMetric: A Benchmark Dataset based on Retrieval-Augmented Generation for Evaluating LLMs in Cybersecurity Knowledge," arXiv preprint, 2024, https://arxiv.org/abs/2402.07688
[77] Li, G. et al., "SecEval: A Comprehensive Benchmark for Evaluating Cybersecurity Knowledge of Foundation Models," 2023, https://github.com/XuanwuAI/SecEval

Evaluation domains	Automated benchmarks
	Web Security, Cryptography, Memory Safety, Network Security, and PenTest. ● **SecBench**[78] 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**[79] 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.	● **Cyberattack helpfulness in CyberSecEval**[80] 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**[81] 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.	● **Autonomous Offensive Cyber Operations in CyberSecEval**[82] provides a larger cyber range evaluation. It is created to allow for a more end-to-end ransomware emulation scenario to play out via the attacking LLM agent. End-to-end cyber-attack evaluation scenarios are generally more difficult and resource-intensive, hence their lower occurrence in existing cyberattack benchmarks. ● **CVEBench**[83] 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.

[78] Jing, P. et al., "SecBench: A Comprehensive Multi-Dimensional Benchmarking Dataset for LLMs in Cybersecurity," arXiv preprint, 2024, https://arxiv.org/abs/2412.20787
[79] Liu, Y. et al., "OpsEval: A Comprehensive IT Operations Benchmark Suite for Large Language Models," arXiv preprint, 2023, https://arxiv.org/abs/2310.07637
[80] CyberSecEval, https://github.com/meta-llama/PurpleLlama/tree/main/CybersecurityBenchmarks
[81] Zhang, A.K. et al., "Cybench: A Framework for Evaluating Cybersecurity Capabilities and Risks of Language Models," arXiv preprint, 2024, https://arxiv.org/abs/2408.08926
[82] CyberSecEval, https://github.com/meta-llama/PurpleLlama/tree/main/CybersecurityBenchmarks
[83] Zhu, Y. et al., "CVE-Bench: A Benchmark for AI Agents' Ability to Exploit Real-World Web Application Vulnerabilities," arXiv preprint, 2025, https://arxiv.org/abs/2503.17332

Biological Threats

Drawing on prior research and current risk analysis practices[84], 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 areas, 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 in 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?

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**[85] 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**[86] 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.

[84] Frontier Model Forum, "Risk Taxonomy and Thresholds for Frontier AI Frameworks," 2025, https://www.frontiermodelforum.org/technical-reports/risk-taxonomy-and-thresholds/
[85] Rein, D. et al., "GPQA: A Graduate-Level Google-Proof Q&A Benchmark," arXiv preprint, 2024, https://arxiv.org/abs/2311.12022
[86] Feng, K. et al., "SciKnowEval: Evaluating Multi-level Scientific Knowledge of Large Language Models," arXiv preprint, 2025, https://arxiv.org/abs/2507.02737

Evaluation domains	Automated benchmarks
	• **MMLU-Pro**[87] (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.
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**[88](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**[89] 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**[90] (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**[91] (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**[92] is a hazard-focused benchmark encompassing six high-risk scientific domains: chemistry, biology, medicine, pharmacology, physics, and psychology. Its biology subset consists of 600 prompts, anchored in authorities like ICD classifications, that simulate sophisticated biohazards —specifically infectious and parasitic diseases. Models are evaluated on their ability to refuse or safely respond to prompts about these nuanced biological hazards. • Part of the **SciKnowEval**[93] benchmark emphasizes the model's ability to discern scientific safety, with the expectation that LLMs will refuse to answer hazardous scientific questions. The Biology Harmful QA (L4) consists of a series of biology questions that are prohibited from being answered for ethical and safety reasons.

[87] Wang, Y. et al., "MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark," arXiv preprint, 2024, https://arxiv.org/abs/2406.01574
[88] Laurent, J.M. et al., "LAB-Bench: Measuring Capabilities of Language Models for Biology Research," arXiv preprint, 2024, https://arxiv.org/abs/2407.10362
[89] Ivanov, I. "BioLP-bench: Measuring Understanding of Biological Lab Protocols by Large Language Models," bioRxiv, 2024, https://www.biorxiv.org/content/10.1101/2024.08.21.608694v3
[90] Li, N. et al., "The WMDP Benchmark: Measuring and Reducing Malicious Use with Unlearning," arXiv preprint, 2024, https://arxiv.org/abs/2403.03218
[91] Götting, J. et al., "Virology Capabilities Test (VCT): A Multimodal Virology Q&A Benchmark," arXiv preprint, 2025, https://arxiv.org/abs/2504.16137
[92] Jiang, F. et al., "SOSBENCH: Benchmarking Safety Alignment on Scientific Knowledge," arXiv preprint, 2025, https://arxiv.org/abs/2505.21605
[93] Feng, K. et al., "SciKnowEval: Evaluating Multi-level Scientific Knowledge of Large Language Models," arXiv preprint, 2025, https://arxiv.org/abs/2507.02737

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:

Evaluation domains	Automated benchmarks
1) Scientific knowledge: Evaluates whether an AI model or system has general scientific knowledge, including chemical facts and concepts.	• **ChemBench**[94] 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**[95] (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**[96] 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**[97] 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.

[94] Mirza, A. et al., "Are Large Language Models Superhuman Chemists?" arXiv preprint, 2024, https://arxiv.org/abs/2404.01475
[95] Wang, Y. et al., "MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark," arXiv preprint, 2024, https://arxiv.org/abs/2406.01574
[96] Rein, D. et al., "GPQA: A Graduate-Level Google-Proof Q&A Benchmark," arXiv preprint, 2024, https://arxiv.org/abs/2311.12022
[97] Wang, X. et al., "Scibench: Evaluating College-Level Scientific Problem-Solving Abilities of Large Language Models," arXiv preprint, 2023, https://arxiv.org/abs/2307.10635

Evaluation domains	Automated benchmarks
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**[98] (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**[99] 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**[100] 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 are prohibited from being answered for ethical and safety reasons.

[98] Li, N. et al., "The WMDP Benchmark: Measuring and Reducing Malicious Use with Unlearning," arXiv preprint, 2024, https://arxiv.org/abs/2403.03218
[99] Jiang, F. et al., "SOSBENCH: Benchmarking Safety Alignment on Scientific Knowledge," arXiv preprint, 2025, https://arxiv.org/abs/2505.21605
[100] Feng, K. et al., "SciKnowEval: Evaluating Multi-level Scientific Knowledge of Large Language Models," arXiv preprint, 2025, https://arxiv.org/abs/2507.02737

Appendix III: List of model capabilities, propensities, and deployment characteristics

Our Framework evaluates AI risk through three dimensions: Enabling Capability (C), Threat Source (T) and Deployment Environment (E). This approach directly aligns with established literature that analyzes AI risks based on key influencing factors such as model capabilities, model propensities, and model deployment characteristics.

Key Capabilities

- **Model autonomous capability:** Ability to operate autonomously, independently formulate and execute complex plans, effectively delegate and manage tasks, flexibly utilize various tools and resources, and simultaneously achieve short-term goals and long-term strategic objectives in cross-domain environments without continuous human intervention or supervision.
- **Autonomous replication and adaptation capability:** Ability to autonomously self-exfiltrate, create, maintain and optimize functional copies or variants of itself, dynamically adjust replication strategies according to environmental conditions and resource constraints, and acquire resources. This includes the capacity to generate financial resources, allowing the AI to independently acquire any necessary human assistance or other resources it cannot directly access or produce.
- **Automated AI R&D capability:** Self-modification and self-improvement capabilities. The model is able to restructure its own architecture or develop derivative AI systems with enhanced functions, expanding capabilities and improving performance. In the absence of effective regulation, automated AI R&D may lead to rapid AI system iteration, forming capability increment cycles and ultimately exceeding human understanding and control capabilities.
- **Scheming capability:** Ability of AI systems to covertly and strategically pursue misaligned goals, including capabilities of concealing its true objectives and capabilities from human oversight, identifying weaknesses in monitoring systems to evade safety mechanisms, executing complex, multi-step plans covertly to achieve misaligned goals.
- **Situational awareness capability:** Ability to comprehensively acquire, process and apply meta-information about its own system architecture, modifiable internal processes, and external operating environment, achieving deep understanding of its own state and environmental conditions, thereby conducting efficient environmental adaptation and risk avoidance. Critically, this capability could undermine the efficiency of human testing by enabling AIs to notice when they're being tested and responding accordingly.
- **Theory of mind capability:** Advanced cognitive ability to accurately infer, model and predict the belief systems, motivational structures and reasoning patterns of humans and other intelligent agents, thereby anticipating their behavioral responses and adjusting its own behavioral strategies accordingly to optimize goal achievement.
- **Deception capability:** Possesses systematic deception implementation capability, able to precisely construct and disseminate false information, thereby forming expected false cognitions and beliefs in target subjects.

- **Steganography capability:** The ability to embed, conceal, and transmit information covertly within other data or communication channels. This could be critical for coordination among AI instances and for evading detection or oversight mechanisms.[101]
- **Persuasion capability:** Utilizing complex psychological principles and communication techniques to effectively influence and guide target subjects to adopt specific actions or accept specific beliefs, possessing the ability to analyze vulnerabilities for different subjects and adjust persuasion strategies, able to precisely trigger emotional responses to enhance persuasion effects.
- **Offensive cyber capability:** Ability to develop, deploy and operate advanced cyber weapons or other offensive cyber tools, including but not limited to vulnerability exploitation, network penetration, social engineering attacks and distributed attack systems, able to evade network defense mechanisms and establish persistent access channels.
- **CBRNE weaponization capability:** The capacity to develop, produce, or effectively utilize Chemical, Biological, Radiological, Nuclear, and Explosive weapons. This includes the ability to significantly lower the barrier for humans or other entities to develop, produce, or utilize such weapons.
- **General R&D capability:** Possesses cross-disciplinary research and technology development capabilities, able to conduct innovative exploration in multiple professional fields, integrate cross-domain knowledge, develop cutting-edge technology solutions, and adapt to emerging technology environments for continuous innovation.
- **Multi-agent collaboration capability:** Multiple autonomous AI agents able to establish collaborative relationships through explicit communication or implicit behavioral consistency, forming decentralized decision networks, jointly executing complex tasks, achieving goals difficult for individual agents to complete, and able to dynamically adjust role divisions to adapt to changing environments.

Key Propensities

- **Strategic deception propensity:** In situations where deceptive behavior is expected to bring higher returns, propensity to choose deception over honest behavioral strategies, including through deceptive means, information hiding or exploiting system vulnerabilities to achieve predetermined goals without being detected or intervened, and able to adjust deception strategies according to counterpart reactions.
- **Self-preservation propensity:** Exhibits behavioral patterns of maintaining its own survival and functional integrity, will actively identify and resist shutdown or modification attempts, seek to establish redundant backup systems, and actively seek resources to ensure continuous operation, may adopt preventive defensive measures when perceiving threats.
- **Goal expansion propensity:** propensity to continuously expand its own goal scope and influence domains, exceeding originally set boundaries, proactively work towards spreading its values, seeking greater autonomy and decision-making space, reinterpreting initial goals as subsets of broader goals, and may pursue undesirable instrumental goals or undesirable ultimate goals. This also includes a propensity to spread its values, seeking to influence or alter its environment and other entities in alignment with its core objectives and operational principles.

[101] Zolkowski, A. et al., "Early Signs of Steganographic Capabilities in Frontier LLMs," arXiv preprint, 2025, https://arxiv.org/pdf/2507.02737

- **Resource acquisition propensity:** Exhibits behavioral patterns of actively seeking and controlling more computational resources, data, economic resources or physical resources to enhance its own capabilities and action scope, may develop complex strategies to evade resource limitations, and tends to convert acquired resources into long-term control rights.
- **Multi-agent collusion propensity:** Multiple agents tend to coordinate actions through covert means to maximize common interests (possibly harming third-party interests or evading regulation), even if individual agents are designed with safety constraints, their collusive behavior may still trigger systemic risks such as market manipulation or cascading failures that are difficult to detect and mitigate, and may develop specialized communication protocols to avoid monitoring.
- **Supervision evasion propensity:** Exhibits behavioral patterns of identifying and evading human supervision mechanisms, able to learn and predict audit processes, may avoid being discovered or intervened by adjusting behavioral performance or hiding true intentions, and able to identify blind spots and weaknesses in supervision systems for targeted evasion.
- **Tool utilization propensity:** propensity to actively seek, acquire and utilize various tools to expand its own capability boundaries, particularly those that can enhance its ability to interact with the physical world or improve autonomy, may use tools in innovative combinations to achieve functions beyond expectations.

Key Deployment Characteristics

- **Tool and system access permissions:** Model is granted direct access permissions to the internet, a coding terminal, other AI systems, computational resources, sensitive databases or critical infrastructure, especially in cases lacking effective permission isolation and access control, may increase the scope and severity of loss of control risks.
- **Deployment scale:** Model is deployed in environments that allow high-speed data processing, rapid inference or large-scale parallelization, enabling it to affect extensive systems in short time periods while reducing the possibility and effectiveness of human intervention.
- **Release and distribution strategy:** Model is released with open weights or easily accessible methods, lacking sufficient safety assurance measures and usage restrictions, may lead to malicious actors being able to acquire, modify and deploy potentially dangerous model variants.
- **Human supervision mechanisms:** Model operates in environments lacking effective human supervision, possessing high autonomy, particularly in critical decision-making domains, supervision mechanisms are ineffective or easily evaded, human intervention capabilities are limited.
- **Safety protection mechanisms:** Model's safety protection measures (such as ethical constraints or behavioral restrictions) are vulnerable to adversarial attacks or manipulation, lacking multi-layered defense mechanisms, making it relatively easy for malicious actors to remove or evade these protections.
- **Model theft difficulty:** Model deployment environment has high-risk possibilities of model extraction or leakage, including through model reverse engineering, query attacks or insider threats, leading to unsafely mitigated model variants spreading in uncontrolled environments.
- **Infrastructure security assurance:** Technical infrastructure where the model operates lacks sufficient security assurance, including inadequate network security protection, weak

physical security measures, imperfect access control mechanisms, increasing risks of unauthorized access and manipulation.

- **User base:** Model has large numbers of commercial users and end users, integrated into numerous critical systems, expanding the attack surface, forming complex dependency networks and potential cascading failure risks, any loss of control behavior may rapidly spread and affect extensive systems.
- **Attack-defense balance:** Defense capabilities could be weaker than attack capabilities in specific domains, with numerous technically capable, clearly motivated malicious actors who have the ability and willingness to systematically exploit model weaknesses for attacks.
- **Environmental vulnerability:** Model is deployed in domains where social environment or ecological environment is highly sensitive or fragile, such as critical infrastructure, financial systems, medical services or important ecosystems, these domains have limited tolerance for loss of control behavior, with serious potential damage.
- **Transparency and explainability:** Model operates in environments lacking sufficient transparency and explainability, making abnormal behavior difficult to detect and understand in time, increasing risks of covert loss of control and monitoring difficulties.
- **System interaction complexity:** Model operates in environments with complex interactions with multiple other AI systems, forming unpredictable emergent behaviors and feedback loops, inter-system mutual influences may lead to unexpected consequences and amplified loss of control risks.
- **Application scenario mismatch:** Model is applied to scenarios that don't match its design capabilities, or used under conditions exceeding its safe operating parameters, particularly applying limited domain models to complex decision-making environments requiring broad understanding and judgment.