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

安远 AI
CONCORDIA AI

Frontier AI Risk Management Framework 2.0

July 2026

Executive Summary

Our vision of trustworthy AGI development

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

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

Introducing our Frontier AI Risk Management Framework

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

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

Evolution to Version 2.0

In July 2026, we were proud to release Version 2.0 of the Framework. Since Version 1.0 (including the 1.5 and 2.0 iterations), key updates include:

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

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

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

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

- **Iterated "Red Line" risk scenarios**: We introduced a new chemical safety red-line risk scenario, and refined the red-line scenarios for biological risks, cyber offense risks, large-scale persuasion and harmful manipulation risks, and loss of control risks.
- **Operationalizing risk analysis**: To make the Framework more operational, we have updated the risk analysis guidance for GPAI model developers. By clarifying the essential modules of this process, such as model evaluation, elicitation, risk modeling and estimates, we aim to make it easier for developers to practically implement risk analysis best practices (see Section 3. Risk Analysis).
- **Enhanced interoperability**: We have mapped our risk management measures against leading international and domestic AI risk management guidance, specifically China's National TC260 AI Safety Governance Framework 2.0 and the EU Code of Practice for General-Purpose AI Models (Safety and Security Chapter). This helps developers adopt safety measures shared by major domestic and international regulatory guidance (see Appendix I and Appendix II).

AI safety as a global public good

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

Versions and Update

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

Current Version: v2.0 (July 2026)

Changelog

Version 2.0 (July 2026)

- Updated chapters related to loss of control risks: included supervision degradation as a cross-cutting enabling factor and identified internal deployment environments as a critical environment for risk generation.
- Iterated on red-line risk scenarios: added a new chemical risk red-line scenario, and iterated on red-line scenarios for biological risks, cyber offense risks, large-scale persuasion and harmful manipulation risks, and loss of control risks.

Version 1.5 (February 2026)

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

Version 1.0 (July 2025)

- Initial release of the Frontier AI Risk Management Framework.

Table of Contents