EU AI Code of Practice Safety and Security	Anthropic	OpenAI	Google DeepMind	Meta	xAI	DeepSeek	Z.ai	Alibaba Cloud
	Responsible Scaling Policy (2.2) May 14, 2025	Preparedness Framework (V2) April 15, 2025	Frontier Safety Framework (3.0) September 22, 2025	Frontier AI Framework (1.1) July 14, 2025	xAI Risk Management Framework August 20, 2025			
4.1 Decision Making								
Measure 4.2 Signatories will base go/no-go decisions for model development, release, and use on whether systemic risks are deemed acceptable (Measure 4.1). **Measure 8.1** Signatories will clearly define, assign and document systemic-risk responsibilities across all organizational levels, including systemic risk oversight, ownership, support and monitoring, as well as assurance. **Measure 8.2** Those who have been assigned responsibilities (Measure 8.1) should be allocated appropriate human, financial and computational resources as well as access to information.	**Go/no-go decisions made by the CEO and RSO** based on whether risks and safeguards remain acceptable under ASL thresholds. These decisions then escalate to the Board of Directors and the Long-Term Benefit Trust before moving forward.	**The Safety Advisory Group (SAG) makes expert recommendations on whether safeguards are sufficient for deployment**; however, OpenAI Leadership can approve or reject these recommendations, and the Board's Safety and Security Committee provides oversight of these decisions.	Response plan to when alert thresholds are reached will be reviewed and approved by **appropriate corporate governance bodies**, such as (1) Google DeepMind AGI Safety Council, (2) Google DeepMind Responsibility and (3) Safety Council, and/or Google Trust & Compliance Council. [Version 2.0]	After the continuous evaluation process, the team will conduct residual risk assessments, which is informed by evaluations and mitigations. The results are reviewed by research and product teams and a multidisciplinary review group (as needed). A leadership team will then decide whether to approve, require further testing, or halt release, guided by the risk thresholds.	Deployment is gated by benchmark-linked thresholds and a tiered-access strategy; functionality can be restricted to only trusted parties. Where warranted, xAI may revoke accounts, temporarily shut down systems, or notify authorities to prevent materially unjustified risk increases. The RMF does not explicitly define how deployment decisions are reached, arguing that "the expected benefits of model deployment may outweigh the risks identified by a particular benchmark," suggesting that risk assessment and capability evaluation results may not automatically trigger decision to pause development and stop deployment.	No safety framework publicly found.	No safety framework publicly found.	No safety framework publicly found.
4.2 Advisory and Challenge								
Measure 8.1 Signatories will designate at least one member from the management body to support and monitor systemic-risk management, including conducting risk assessments and mitigations	RSO is designated to be responsible for reducing catastrophic risk, primarily by ensuring that the policy is designed and implemented effectively. Its specific duties are also clearly defined, covering the full life stages of policy development to policy enforcement.	The SAG is the internal cross-functional advisory body that reviews threat models, Capability Reports, Safeguards Reports and makes recommendations to OpenAI Leadership regarding the level and type of safeguards required for deploying frontier capabilities safely and securely.	The DeepMind AGI Safety Council will periodically review the implementation of the Framework. [Version 2.0]	It is unclear which leadership team will be responsible for supporting and monitoring the systemic risk management.	No internal body has been appointed or identified to support and monitor the systemic risk management. But the RMF integrates the approach of designating risk owners, who are responsible also for proactively mitigating identified risks.	No safety framework publicly found.	No safety framework publicly found.	No safety framework publicly found.
4.3 Audit								
Measure 8.1 Signatories will designate an assurance role (e.g., Chief Audit Executive or Head of Internal Audit) that is tasked with providing assurance on the adequacy of systemic-risk processes to the board or its supervisory function. This individual is supported by internal audit and, where appropriate, external auditors.	ASL-3 Security requires the mechanism to (1) audit and assess the design and implementation of the security program and (2) share these findings with management on an appropriate cadence. The following methods have been recommended: independent validation of threat modeling and risk assessment results; a sampling-based audit of the operating effectiveness of the defined controls; periodic, broadly scoped, and independent testing with expert red-teamers who are industry-renowned and have been recognized in competitive challenges.	The framework requires auditing and transparency mechanisms as part of the security controls for High capability models. These measures include independent security audits to security controls and practices are validated regularly by third-party auditors to ensure compliance with relevant standards and robustness against identified threats.	Auditing was mentioned as an example of the suite of safeguards targeting the capability, although it is not a formal part of the deployment mitigations.	There is no mention of internal or external audit functions in the Framework.	There is no mention of internal or external audit functions in the Framework.	No safety framework publicly found.	No safety framework publicly found.	No safety framework publicly found.
4.4 Oversight								
Measure 8.1 Signatories will assign a specific committee of the management body in its supervisory function or one or more multiple suitable independent bodies to oversee its systemic risk management processes and measures.	**Oversight is provided by the Board of Directors**, including the Long-Term Benefit Trust, which review risk determinations, safeguard implementation, and deployment decisions under the RSP.	**Oversight is provided by the Board's Safety & Security** Committee, which receives information on process and decisions and "may reverse a decision or mandate a revised course of action" if necessary.	Appropriate corporate governance bodies such as the Google DeepMind AGI Safety Council, Google DeepMind Responsibility and Safety Council, and/or Google Trust & Compliance Council will review and approve response plans, while Google DeepMind AGI Safety Council will periodically review the implementation. [Version 2.0]	A leadership team will then decide whether to approve, require further testing, or halt release, guided by the risk thresholds, although it is unclear who will make up the leadership team.	No oversight body has been identified in the RMF.	No safety framework publicly found.	No safety framework publicly found.	No safety framework publicly found.
4.5 Culture								
Measure 8.3 Signatories will promote a healthy risk culture and take appropriate measures to ensure that actors who have been assigned responsibilities for managing the systemic risks stemming from their models (**Measure 8.1**) take a reasoned and balanced approach to systemic risk. Examples include leadership priority, clear communication and challenge of decisions concerning systemic risks, active internal reporting channels, no retaliation, incentives and structural independence for objective risk assessment and less excessive risk-taking, and easy public access and regular reminder of whistleblower policy.	Anthropic protects employees' ability to raise safety and compliance concerns without retaliation by maintaining anonymous reporting channels for noncompliance to the RSO and the Board of Directors and prohibiting non-disparagement clauses that could discourage speaking up about safety issues. Anthropic has multiple teams working on AI safety research including alignment science, interpretability, frontier red team, safeguards team and more.	OpenAI's employees can access summaries of Safety Advisory Group (SAG) testing results and recommendations, within confidentiality limits. All potential policy violations or implementation issues can be reported under the Raising Concerns Policy, and each report is tracked, investigated, and addressed with proportional corrective actions. (The whistleblower policy will be discussed more in detail in "Governance and Accountability" Section).	No internal reporting or anti-retaliation mechanisms are referenced in the Framework.	No internal reporting or anti-retaliation mechanisms are referenced in the Framework.	Employees can raise concerns to relevant government agencies regarding imminent threats to public safety based on whistleblower policy.		Z.ai's safety team is made up of Zhipu Evaluation Team, Zhipu Safety Team, Zhipu Posttraining Team. The teams do not have team websites and prefer not to disclose mission and scope. There are 20-30 technical FTEs for safety teams.	
4.6 Transparency								
Commitment 7 Safety and Security Model Reports Signatories must document, justify, and continuously report the safety and security of these models to the EU AI Office. - Content Requirements (Measure 7.1-Measure 7.5), such as model description and behavior, reasons for proceeding with development, documentation of risk identification, analysis, and mitigation, external reports, and material changes to the systemic risk landscape	**Anthropic promises to share publicly key information** related to the evaluation and deployment, including (1) Capability and Safeguards Reports for deployed models, (2) plans for comprehensive capability assessments and deployment and security safeguards. **It will also ask for external input from experts** for developing and conducting the capability and safeguards assessments and third-party review of procedural commitments on an approximately annual basis.	**OpenAI promises to share with the public** summaries of capability evaluations, testing scope, reasoning behind deployment decisions, and implemented safeguards (for models at or beyond the High threshold), with redactions where needed for security or proprietary reasons.	The Frontier Safety Framework will be updated at least once a year, including the CCLs and the testing and mitigation approaches.		xAI intends to publish publicly and for third-party reviews with potentially redacted information for concerns of public safety, national security, and protection of intellectual property: (1) Updates to the RMF (2) Adherence with the RMF (3) Benchmark results (4) Internal AI Usage (5) Employee survey for important future developments of AI		Z.ai has a written formal policy to conduct regulator-only notification, where the policy mandates prompt disclosure to a competent regulatory, or supervisory authority when safety testing determines a model exceeds its "unacceptable-risk" threshold.	
- Update Duties when signatories have reasonable grounds to believe if they have reasonable grounds to believe that the justification for why the systemic risks stemming from the model are acceptable - Notifications **Measure 10.1** Signatories must maintain comprehensive internal documentation on model architecture, system integration, evaluations, and safety mitigations. They must also record processes, key risk-related decisions, and justifications for their chosen safety practices. Documentation must be kept for at least 10 years and be made available to the AI Office upon request. **Measure 1.3** Signatories will update the Framework as appropriate, including without undue delay after a Framework assessment to ensure the information for the safety framework is kept up-to-date and the Framework is at least state-of-the-art. For any update of the Framework, Signatories will include a changelog, describing how and why the Framework has been updated, along with a version number and the date of change. Signatories must document, justify, and continuously report the safety and security of these models to the EU AI Office.	**The company will also notify U.S. government authorities** if stronger protections than ASL-2 are needed. In the system card for Sonnet 4.5, Anthropic has noted that the model does not require comprehensive capability assessment since it does not meet the "notably more capable" threshold. Comprehensive automated testing, comparative capability assessment to earlier models, and conservative threshold application evaluations confidently rule out ASL-4 capabilities across all domains. The decision was overseen by the RSO and followed the company's established protocols for precautionary ASL determinations	When warranted, OpenAI will engage independent third parties to evaluate model capabilities and stress-test safeguards, particularly for high-risk deployments. The SAG may also seek independent expert opinions to inform its safety determinations before deployment. In the system card for GPT-5, OpenAI recorded both scalable and deep-dive evaluations for the model across the three Tracked Categories, including **both internal and external assessments** compiled into a Capabilities Report for the SAG. The SAG reviewed the evidence and concluded that GPT-5-Thinking reached the High threshold, requiring "safeguards sufficiently minimize associated risks" before deployment. The Preparedness Team compiled mitigations into a Safeguards Report, **validated through extensive third-party red-teaming**. The SAG, supported by OpenAI leadership and external experts, provided oversight across the evaluation and mitigation phases. **There is no written requirement to notify any external body if safety testing determines a model exceeds OpenAI's "unacceptable-risk" threshold.**	**Google DeepMind is dedicated to sharing relevant information with appropriate government authorities when a model has reached a CCL according to their assessments**. These disclosures occur under strict confidentiality and security safeguards. Such information may include model information, evaluation results, and mitigation plans. Google DeepMind also considers disclosing information to other external organizations to promote shared learning and coordinated risk mitigation, although unclear under what circumstances.	In the Framework, Meta states their continuous dedication to openly releasing models to the ecosystem, sharing relevant information about responsible development and evaluation through model cards and research papers and believes that this will allow their team to work with outside experts and allow external independent assessment of their models. However, according to a letter released by Mark Zuckerberg on July 30, 2025, the CEO of Meta noted that the company will be "careful about what we choose to open source."			