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			
3.1 Mitigation Measures								
Measure 5.1 Signatories will implement safety mitigations that are appropriate along the entire model lifecycle, to ensure systemic risks stemming from the model are acceptable. (Commitment 4) **Commitment 6** Signatories will implement adequate cybersecurity protection for models and physical infrastructure along the entire lifecycle, to ensure systemic risks stemming from their models from unauthorised releases or access, and/or model theft are acceptable. **Measure 6.1** Signatories will define a goal that specifies the threat actors that their security mitigations are intended to protect against. **Measure 6.2** Signatories will implement appropriate security mitigations to meet the security goal, including the security mitigations pursuant to Appendix 4, such as general security mitigations, protection of unreleased model weights, hardening interface-access to unreleased model parameters, insider threats, and security assurance.	The latest model Sonnet 4.5 is deployed under ASL-3, according to the system card. In the Framework, ASL-3 Security Standards have clearly defined threat actors within scope and out of scope. It requires mitigation measures such as threat modeling, security framework, including parameters and access controls around sensitive assets, life cycle security, ongoing and effective monitoring, sufficient resourcing, and existing guidance, audits and documenting compliance when models are deployed in third-party environments. In addition, ASL-3 Deployment Standard requires threat modeling, defense in depth, red-teaming, rapid remediation, monitoring, trusted users, and documenting compliance when models are deployed in third-party environments. These measures are described as high-level outcomes and do not include actionable and measurable protocols.	The Preparedness Framework includes only illustrative examples of safeguards against malicious users, against a misaligned model, and security controls It also includes corresponding efficacy assessments for these safeguards. The latest model ChatGPT-5 is deployed under High Capability threshold for the Biological and Chemical Risks. The deployment includes multilayered mitigations—such as refusal and safe-completion training, real-time monitoring classifiers, account-level enforcement, and API safety identifiers.	The Framework does not specify the mitigation measures for the security controls at a level generally aligned with "security standards such as RAND SL 2, RAND SL 3, and RAND SL 4." It explained such decisions are due to the fact that they "expect the concrete [security] measures implemented to reach each level of security to evolve substantially." Deployment mitigations involve processes that are "designed to ensure that residual risk remains at acceptable levels," which involves (1) the development and assessment of mitigations; (2) pre-deployment review of safety case; (3) post-deployment where safety cases and mitigations may be updated if deemed necessary by post-market monitoring. The latest model Gemini 2.5 Pro did not reach the CBRN Uplift Level 1 CCL, Cyber Autonomy Level 1 and Uplift Level 1, Machine Learning R&D Uplift Level 1 and Autonomy Level 1. However, alert thresholds for the model's alert threshold for Cyber Uplift Level 1 prompted proactive measures—specifically, increased evaluation cadence and accelerated mitigation deployment.	The Framework states that the full mitigation strategy will be informed by the risk assessment, the frontier AI's particular capabilities, and the release plans. It does not prescribe a fixed set of mitigations, but list a few examples include certain examples including fine-tuning, misuse filtering, response protocols, sanctions screening and geo-gating, staged release to prepare the external ecosystem. Meta has not updated Llama 4 Maverick's system cards to reflect these changes.	The RMF references mitigations measures on a high level, including: (1) safety training, system prompts, and input & output filters for malicious use risks (2) safety training for controllability, and system prompt for loss of control risks. These mitigations do not correlate with the aforementioned threshold. The latest model Grok-4 have implemented safeguards in particular for the Bio & Chem risk, including (1) narrow, topically-focused filters for bioweapon abuses and chemical weapons-related abuses; (2) existing system prompts against radiological and nuclear weapons development.	No safety framework publicly found.	No safety framework publicly found.	No safety framework publicly found.
3.2 Continuous Monitoring and Comparing Results with Predetermined Thresholds								
	Anthropic's capability assessment for the most pressing risks has three stages: (1) preliminary testing, (2) comprehensive evaluation, and (3) a capability decision. - Models showing (1) a 4× increase in Effective Compute or (2) six months of fine-tuning trigger full testing. - Comprehensive evaluation covers threat modeling, empirical testing, elicitation under attacker scenarios, and forecasting. - Results are reviewed by the Responsible Scaling Officer (RSO) and CEO to decide if escalation is needed. Accordingly, it also assesses the safeguards of the Deployment and Security Standard. After evaluating their implementations, the CEO and the RSO (1) make the ultimate determination as to whether we have satisfied the Required Safeguards and (2) decide any deployment-related issues after soliciting internal and external expert feedback on the evaluation. Safeguards will be revisited and re-approved at least annually. There also exists follow-up capability assessment that rechecks to ensure model capabilities remain below higher thresholds and updates the policy if new risks emerge.	Before deployment, every model covered by the Framework undergoes a structured suite of Scalable Evaluations: automated tests that measure capability proxies tied to risk thresholds. The results of which will be compiled into a Capabilities Report that is submitted to the SAG. The report will be reviewed by the SAG to decide on the next steps, which can include (1) Capability threshold is crossed, recommending to implement and assess corresponding safeguards; (2) Capability threshold has not been met, (3) Recommend deep Dive evaluations, such as expert red-teaming or third-party assessments, to validate those results. Accordingly, it also assesses the safeguards through a Safeguards Report, which compiles all identified pathways by which severe harm could occur, the corresponding mitigations, their measured efficacy, the residual risk after controls are applied, and notable limitations. The SAG reviews this report to determine whether the safeguards in place sufficiently minimize the risks associated with the model's capability level and deployment context, drawing on internal and external expert input as needed.	Google conducts regular early-warning evaluations to monitor whether models are nearing critical capability levels. These evaluations use predefined alert thresholds and are adjusted in frequency or sensitivity as model progress accelerates. When needed, they are supplemented by additional assessments to ensure an accurate understanding of capability proximity and to maintain a sufficient safety buffer before deployment.	Meta lays out the monitoring and continuous evaluation process in the following procedure: (1) Ensuring robust evaluation environment (2) Conducting evaluations for performance and safety, against our expectations for the reference class as well as the enabling capabilities we have identified in our threat scenarios. - The indicators trigger for further evaluations as capabilities develop. (3) Evaluations are repeated as a frontier model is close to or completes training.	xAI continuously measures model's safety properties through public benchmarks and monitors live use through public deployment (e.g. Grok on X) It also regularly evaluates the adequacy and reliability of such benchmarks, including by comparing them against other benchmarks that we could potentially utilize, to determine and apply effective benchmarks available at the time of evaluation.	No safety framework publicly found.	No safety framework publicly found.	No safety framework publicly found.