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			
1.1 Classification of Applicable Known Risks								
Measure 2.1 (Appendix 1.1 to 1.4) Signatories will identify systemic risk through two approaches. (1) Following the specified structured process to compile a list of identified systemic risks, taking into consideration model-independent data and analysing relevant characteristics such as nature of the systemic risk and sources of the systemic risk (including model capabilities, model propensities, and model affordances) (Appendix 1.1-1.3). (2) Four risks are treated as specified systemic risks that are always identified: CBRN risks, loss of control, cyber offense, and harmful manipulation (Appendix 1.4) **Measure 2.2** Signatories will develop appropriate systemic risk scenarios for each identified systemic risk. **Measure 3.2** Model evaluations should [...] should include open-ended testing of the model, to improve the understanding of the systemic risk, with a view to identifying unexpected behaviours, capability boundaries, or emergent properties.	Anthropic identifies CBRN weapons and Autonomous AI R&D as its two most pressing catastrophic risks. In addition, it also designates cyber operations as an emerging risk category under ongoing evaluation. Although it recognizes potential risks of highly persuasive AI models, active consultation with experts lead to the conclusion that this capability is "not yet sufficiently understood to include in the current commitments." Anthropic prioritizes these risks through the process of external engagements such as commissioned research reports, discussions with domain experts, input from expert forecasters, public research, conversations with other industry actors through the Frontier Model Forum, and internal discussions.	OpenAI uses a structured risk-assessment process to evaluate whether frontier AI capabilities could lead to severe harm, which is defined as death of thousands or hundreds of billions of dollars in economic damage. The process relies on its own internal research and signals, and where appropriate incorporates feedback from academic researchers, independent domain experts, industry bodies such as the Frontier Model Forum, and the U.S. government and its partners, as well as relevant legal and policy mandates. It assigns identified risks to categories: (1) Tracked Categories: currently including Biological & Chemical, Cybersecurity, AI Self-improvement and; (2) Research Categories, including Long-range Autonomy, Nuclear & Radiological for further work.	DeepMind's Framework identifies misuse risks in three domains: Misuse (CBRN, Cyber, and Harmful Manipulation), ML R&D, as well as Misalignment (exploratory) risk. These risks are organized by the framework around capability thresholds called "Critical Capability Levels" (CCLs). The selection is attributed to "early research" that judged these areas most likely to lead to severe harm from future models if unmitigated, but the framework does not describe a formal methodology or process for how these risk domains were identified.	Meta adopts an outcome-based approach described in high levels where it proceeds by (1) defining catastrophic outcomes; (2) maps the causal pathways that could produce them; (3) locate threat scenarios that are potentially sufficient to realize the outcome. The most urgent catastrophic outcomes identified are in the domains of cybersecurity and chemical and biological weapons.	xAI focuses on two overarching systemic risks—malicious use and loss of control—and organizes concrete risk scenarios across abuse potential (e.g., vulnerability to jailbreaks), concerning propensities (e.g., a propensity for deceiving the user), and dual-use capabilities (e.g., offensive cyber capabilities). It does not spell out a formal risk-identification process, but it does quantify "catastrophic malicious-use events" using thresholds for expected fatalities and economic damage.	No safety framework publicly found.	No safety framework publicly found.	No safety framework publicly found.
1.2 Identification of Unknown Risks								
	The Responsible Scaling Policy does not specify pre-deployment measures to identify novel risk domains for the frontier model, although Anthropic has implemented adversarial testing, red-teaming, and bug bounty programs that can help the company identify unknown threats.	**The Preparedness Framework** mentions that OpenAI conducts adversarial testing, red-teaming, and bug bounty programs to proactively identify and mitigate unknown vulnerabilities and emerging threats across its corporate, research, and product systems.	The Frontier Safety Framework explicitly states that it will "continue to assess whether there are other risk domains where severe risks may arise and will update our approach as appropriate," Moreover, the early warning evaluations are intended to to flag when a CCL may be reached before the evaluations are run again, however, it is also used for detecting novel risks from the frontier AI systems.	The team follows the general process of (1) Hosting workshops with experts to identify new catastrophic outcomes and/or threat scenarios (2) Designing new assessments if novel outcomes/scenarios are identified.	The RMF has not explicitly designated a process specifically for identifying unknown risks, although it emphasizes the development of naturalistic evaluation environments to assess more realistic, real-world model behaviors.	No safety framework publicly found.	No safety framework publicly found.	No safety framework publicly found.
1.3 Risk Modeling								
Measure 3.3 Signatories will model systemic risks using **at least** state-of-the-art methods, informed by predefined risk scenarios (Measure 2.2) and data collected through prior identification measures (Measure 2.1)	Anthropic has implemented a **multi-layered threat-modeling** strategy spanning three stages: (1) Capability assessment, where it maps plausible catastrophic-risk scenarios—actors, attack pathways, and harms—to determine whether model capabilities approach predefined Capability Thresholds; (2) Deployment safeguards, where it maps out the set of threats and vectors through which an adversary could catastrophically misuse the deployed system; (3) Security safeguards, where it seeks to establish the relationship between the identified threats, sensitive assets, attack vectors and, in doing so, sufficiently capture the resulting risks that must be addressed to protect model weights from theft attempts, using best practices such as the MITRE ATT&CK Framework. It does not mention the specific methodologies involved, lists of risk scenarios, and the complete risk models in the RSP.	The Framework identifies threat modeling as "a causal pathway for a severe harm in the capability area," which is one of the five criteria to meet to categorize a frontier risk to the Tracked Category. It is guided by both (1) the broader risk assessment process, and (2) more specific information that it gathers across OpenAI teams and external experts. The threat models are reviewed and approved by the internal, cross-functional group called Safety Advisory Group (SAG). It does not mention the specific methodologies involved, lists of risk scenarios, and the complete risk models in the Preparedness Framework.	The Framework describes risk modeling as "identifying and analyzing the main foreseeable paths through which a model could cause severe harm," and requires it for both risk assessment and mitigation assessment. The framework does not mention the specific methodologies involved, list of risk scenarios, and the complete risk models.	Meta's risk modeling exercises begin by testing whether the model has the (1) enabling capabilities and (2) could uniquely enable these scenarios to catastrophic outcomes. Inclusion for risk modeling follows a four-layered qualitative criteria, where risks have to be plausible, catastrophic, net new, and irreparable. The risk modeling process is informed by (1) internal assessment; (2) external engagements (governments, external experts, and the wider AI community). The qualitative risk scenarios are included in the risk threshold framework.	**The team adopts threat modeling specifically for Biological and Chemical Weapon risks.** Specifically, it breaks down the 5 critical steps where xAI models are restricted from providing detailed information or substantial assistance. These steps are defined qualitatively, in collaboration with external domain experts from organizations such as SecureBio, NIST, RAND, and EBRC. However, it does not construct specific risk scenarios combining some or all of these critical steps identified.	No safety framework publicly found.	No safety framework publicly found.	No safety framework publicly found.