

EU AI Code of Practice Safety and Security	Anthropic	OpenAI	Google DeepMind	Meta	xAI	DeepSeek	Z.ai	Alibaba Cloud	Mistral
	Responsible Scaling Policy (RSP, V3.2) Last updated: April 29, 2026 Frontier Compliance Framework (FCF) Last Updated: March 2026	Preparedness Framework (V2) Last Updated: April 15, 2025	Frontier Safety Framework (3.1) Last updated: April 17, 2026	Frontier AI Framework (1.1) Last updated: July 14, 2025 Advanced AI Scaling Framework (Version 2) Last updated: April 7, 2026	xAI Risk Management Framework Last updated: August 20, 2025	No Public Safety Framework. Signed on to the following commitment framework: Artificial Intelligence Security and Safety Commitments Framework Last updated: July 2025			No Public Safety Framework. Signed on to the following commitment framework: Frontier AI Safety Commitments, AI Seoul Summit 2024 Last updated: Feb 2026
EU AI Code of Practice (Safety and Security) 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.	The FCF defines systemic risks with quantitative thresholds, as they include “foreseeable and material risks of large-scale harm from the most advanced (i.e. state-of-the-art) models at any given point in time, including but not limited to >50 fatalities arising from a single incident, or 1 billion dollars of financial damages.” Such risks include: (a) Cyber offense; (b) CBRN threats; (c) Harmful manipulation (new) (4) Sabotage and loss of control. The RSP explicitly identifies and maps capability thresholds to the following risk categories, excluding cyber offense: (a) non-novel CBRN weapons production; (b) novel CBRN production (including moderately resourced state-backed teams); (c) high-stakes sabotage opportunities (loss of control / autonomous goal-directed behavior), and; (d) automated R&D in key domains (AI R&D acceleration). (Source: FCF pp.4-7, RSP 3.2 pp.6-10)	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. (Source: pp.4-8)	The FSF includes “potential risks that could stem from [Google’s] models and analyze their characteristics to determine which of the identified risks could be significant or severe risks.” These risks are broadly categorized into two groups: (1) Misuse; (2) ML R&D and Alignment. Misuse risks are further broken down into: (a) CBRN; (b) Cyber; (c) Harmful manipulation. The risk identification process starts from “consider [ing] a wide range of risks as part of [Google’s] ongoing research, taking into account the characteristics, capabilities, propensities, and affordances of their models and other sources of information, such as internal risk taxonomies, internal expertise and relevant external research.” (Source: pp.4)	The Framework covers detailed descriptions of catastrophic outcomes, operational thresholds, and mitigation strategies in the following categories: (1) Cybersecurity, (2) Chemical & biological risks, and, (3) Loss of control The Framework has also identified emerging fields for evaluations and outcomes, including: (1) Radiological and nuclear; (2) Physical autonomy (3) Loss of control (recognized as a nascent field that requires continuous research, including concerns for threats to human oversight capacity and AI containment measure) The Framework has identified the following procedure to locate “an estimated ‘reference class’ of comparable models that we use throughout development to track how our model is performing and anticipate associated risks, required assessments, and assess the applicability of available mitigation strategies.” Specifically, for each frontier AI system, Meta “outline anticipated capabilities, planned deployment (i.e., internal deployment, limited deployment, controlled deployment, closed release, or open release), supported modalities, intended uses and anticipated benefits of the model, and expectations for compute requirements.” The process generally involves (1) identification of catastrophic outcomes; (2) threat modeling; (3) identification of key risk factors (e.g. model capabilities and propensities) (Source: pp.4, pp.12, pp.18–24, pp.36–38)	xAI focuses on two overarching systemic risks: (1) malicious use and (2) 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. (Source: pp.1-3)	“DeepSeek, Z.ai , and Alibaba has signed the Artificial Intelligence Safety Commitments published by China Academy of Information and Communications Technology (CAICT), a public research institute directly subordinate to China’s Ministry of Industry and Information Technology (MIIT) in 2024. Commitment II of the Framework asks the companies to “prioritize safety and reliability evaluations focusing on general understanding, reasoning, and decision-making capabilities, as well as performance in critical domains such as industry, education, healthcare, finance, and law” especially for LLMs. Commitment VI of the Framework asks the companies to “strengthen the assessment of risks related to the abuse of AI systems in frontier fields, and prevent potential risks of their abuse in high-risk scenarios.” <i>Note: Companies have signed onto the commitment but have not published their own frameworks. These commitments are just for reference.</i>	Outcome 1 defines that “organisations [should] effectively identify, assess and manage risks when developing and deploying their frontier AI models and systems.” <i>Note: Mistral has signed onto the commitment but have not published their own frameworks. These commitments are just for reference.</i>		
	1.1 Classification of Applicable Known Risks								
	1.2 Identification of Unknown Risks								
	Both the FCF and RSP do not specify pre-deployment measures to identify novel risk domains for the frontier model. The risk identification process involves literature reviews and expert consultation, internal safety and alignment research, and insights from monitoring deployed models and investigating serious incidents and critical safety incidents. (Source: FCF, pp.3)	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. (Source: pp.20-21)	The team commits that “it continue to assess whether there are other risk domains where significant or severe risks may arise and will update our approach as appropriate.” (Source: pp.5)	The team “conducts ex-ante threat modeling exercises to help determine whether models with new capabilities may pose novel risks.” (Source: pp.5)	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.	Not applicable		Not applicable	
1.3 Risk Modeling									
EU AI Code of Practice (Safety and Security) 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)	The RSP requires threat modeling, defined as “the specific ways that models might pose threats,” as part of the new risk report. Published independently of model releases and every 3 to 6 months, these reports discuss “the risks of Anthropic’s AI systems and how the company has made determinations about whether to continue AI development and deployment in light of the risks.” The FCF includes threat modeling as part of the risk identification process. (Source: RSP pp.11)	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. (Source: pp.4)	The FSF states that Critical Capability Levels (CCLs) “are determined by identifying and analyzing the main foreseeable paths through which a model could result in severe harm,” although the specific methodologies for risk modeling are not specified. (Source: pp.4)	In the event that [Meta] identify that a Frontier AI is likely to substantially contribute to a threat scenario for a catastrophic outcome, it will conduct a threat modeling exercise. The general process involves: (1) Host workshops with experts, including external subject matter experts where relevant, to identify new catastrophic outcomes and/or threat scenarios; (2) If new catastrophic outcomes and/or threat scenarios are identified, design new assessments to test for them, in consultation with external experts where relevant. Threat modeling exercises are run both internally and externally with domain experts (where necessary), with informed insights from internal experts’ assessment of the catastrophic risks that Frontier AI might pose, as well as engagements with governments, external experts, and the wider AI community. The Framework seeks to consider risks that satisfy all four criteria, including: (1) Plausible , meaning that it must be possible to (a) identify a causal pathway for the catastrophic outcome; (2) define one or more simulatable threat scenarios along that pathway. (2) Catastrophic , referring to outcomes that may have large scale, devastating, and potentially irreversible harmful effects (3) Net-new , referring to outcomes that are currently not realizable as described without access to general purpose AI. (4) Instantaneous or irremediable , meaning that if outcome is realized, the catastrophic impacts are immediately felt, due to a lack of feasible measures to remediate. (Source: pp.4-5, pp.12-14)	The team adopts threat modeling for Biological and Chemical Weapon risks . Building on external research and collaboration with external subject matter experts, the threat modeling identifies 5 qualitative steps where xAI systems are restricted from providing detailed information or substantial assistance. However, threat modeling is not mentioned for other risk domains. (Source: pp.4)	Commitment IV ask companies to “conduct regular and dynamic security penetration tests to simulate potential risk scenarios, identify and report security vulnerabilities in the infrastructure, and assess associated risks.” <i>Note: Companies have signed onto the commitment but have not published their own frameworks. These commitments are just for reference.</i>		Not applicable	