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			
2.1 Setting a Risk Tolerance								
Measure 4.1 Signatories will establish clear and measurable thresholds for acceptable systemic risk for **each identified systemic risk**, informed by systemic-risk identification (Commitment 2) and analytical evidence from model data, evaluations, modeling, estimation, and post-market monitoring (Commitment 3). They will explain how these thresholds guide risk-acceptance decisions, justify why the approach ensures safety, and apply safety margins to account for uncertainty and potential mitigation failure.	The RSP has defined a **qualitative boundary of acceptable risks**, expressed as capability thresholds of the identified risks of CBRN weapons and autonomous AI R&D. These thresholds (CBRN-3, CBRN-4, AI R&D-4, AI R&D-5, and the autonomy checkpoint) marks the upper bound of risk that Anthropic considers acceptable to manage under existing deployment and security safeguards. Anthropic has not included how it has defined these thresholds and noted that they are "uncertain how to choose a specific threshold," but they "maintain a current list of specific CBRN capabilities of concern for which they would implement stronger mitigations," sharing only with selected organizations such as the AI Safety Institute and Frontier Model Forum.	The Framework establishes threshold levels of capability for when additional safeguards or no deployment apply. High and Critical capability thresholds refer to capabilities that increasing for severe harm in terms of existing and qualitatively new threat vectors respectively. For each risk in the Tracked Category, **capability thresholds qualitatively describe** things an AI system might be able to help someone do or might be able to do on its own that could meaningfully increase risk of severe harm, with corresponding threat models. OpenAI has not included how it has defined these thresholds.	The Framework establishes threshold levels of capabilities (CCLs) for when mitigation plans or suspension of deployment are required until risks are addressed. For each risk identified in the misuse category, **capability thresholds qualitatively describes how an AI system can "uplift" or autonomously carry out actions that will lead to risks of severe harm.** The CCLs are identified through "ongoing analysis" of the risk domains, which are expected by the team to "evolve over time," although the details of which are not included in the Framework. [Version 2.0 and Version 3]	The Framework establishes risk thresholds based on the extent to which a frontier AI model can uniquely enable execution of any of the threat scenarios. The framework introduces **a three-layered capability threshold** of **moderate, high, and critical**, which corresponds to (1) "release" - the model does not provide a significant uplift (2) "do not release" - the model can not yet uniquely enable a catastrophic threat scenario, but provides a significant uplift (3) "stop development" - the model can uniquely enable at least one complete catastrophic threat scenario	The RMF currently has sets **quantitative thresholds for Biological and Chemical risks**, which is to maintain an answer rate of less than 1 out of 20 on restricted queries; and for Loss of Control, which is to maintain a dishonesty rate of less than 1 out of 2 on MASK. It has cited plans to "add additional thresholds tied to other benchmarks." Performance against the Bio & Chem threshold is evaluated using an internal benchmark of benign and restricted biology- and chemistry-related questions developed in collaboration with SecureBio.	No safety framework publicly found.	No safety framework publicly found.	No safety framework publicly found.
2.2 Operationalizing Risk Tolerance								
	For each risk domain, two qualitative Key Risk Indicators (KRIs) are defined (CBRN-3, CBRN-4; AI R&D-4, AI R&D-5) to trigger escalation to ASL-3 or ASL-4 safeguards. **The indicators are primarily qualitative and not directly measurable**, with the exception of AI R&D-5, which specifies a quantitative benchmark based on effective scaling. No clear mapping is provided between these indicators and specific evaluation tests or quantitative thresholds, although Anthropic has noted that they prefer the flexibility of affirmative cases to board-approved evaluations. For each KRI, there are corresponding Key Control Indicators (KCIs) in the required safeguards that would apply upon escalation, including safeguards for deployment and security. These KCIs are defined qualitatively rather than quantitatively. The ASL-3 deployment safeguards "evaluate whether the measures Anthropic has implemented make us robust to persistent attempts to misuse the capability in question," but they do not include numerical thresholds or measurable performance criteria. The ASL-3 security safeguards "evaluate whether the measures Anthropic has implemented make us highly protected against most attackers' attempts at stealing model weights." The ASL-4 safeguards have not yet been defined.	For each risk domain, **two qualitative KRIs are defined**. The indicators are primarily qualitative, with the exception of AI R&D Critical, which specifies a more quantitative baseline. No clear mapping is provided between these indicators and specific evaluation tests or quantitative thresholds. For each KRI, there are corresponding KCIs in the required safeguards would apply upon escalation, including including security controls [High], safeguards against misuse [High], safeguards against misalignment [High], and development halts [Critical].	For **CBRN** and **Cyber risks**, the Framework defines qualitative thresholds for uplift capabilities. For Harmful Manipulation risk, an exploratory threshold is introduced. **ML R&D** risks now include two distinct thresholds: Acceleration Level 1, when models substantially accelerate AI progress beyond historical rates, and Automation Level 1, when models can fully automate the work of an AI research team. For **Misalignment** risks, the Framework retains two Instrumental Reasoning Levels as part of its exploratory approach For each KRI identified in the misuse risk categories, there exist corresponding KCIs as recommended security level (which is mapped to RAND Security Level) with the justifications. For the two instrument reasoning capabilities for misalignment risks, automated monitoring is required for level 1, while the team is still coming up with the approaches for Level 2.	**For each risk (cybersecurity and bio&chem weapons), 3 layers of qualitative catastrophic outcomes are identified.** For each outcome, 1-2 qualitative threat scenarios (Key Risk Indicators) are identified. Correspondingly, the threshold framework includes examples of model enabling capabilities for each threat scenarios. Meta deliberately withholds the detailed breakdown of how each threat scenario could be executed, citing concerns for balancing transparency vs. security. Meta does not include KCIs in accordance with KRIs.	**The quantitative threshold for malicious use risk and loss of control risk is not tied to any specific threat scenarios and is also not related to any specific safeguards accordingly**. While the RMF references safeguards at a high level, such as safety training, system prompts, and input & output filters, it does not specify how these measures are triggered, adjusted, or evaluated against the established thresholds.	No safety framework publicly found.	No safety framework publicly found.	No safety framework publicly found.