EU AI Code of Practice Safety and Security	**Commitment 9 (Measure 9.1-9.4)** Signatories are required to adopt additional measures to track serious incidents, including monitoring external sources such as media reports, research papers, and incident databases, and enabling downstream developers, users, and third parties to report incidents through clear channels. When reporting to relevant authorities, signatories must include details such as the incident timeline, harm caused, affected parties, chain of events, model involvement, corrective actions, and root cause analysis. Reporting must occur promptly—within 2 to 15 days depending on the severity of the incident—followed by updates every 4 weeks until resolution and a final report within 60 days after resolution. All related documentation must be retained for at least five years.
Anthropic	**Serious incident reporting frameworks:** No information found **Red-line Government notifications commitments:** Responsible Scaling Policy contains a broad voluntary commitment on ASL disclosing ASL levels: - "We will notify a relevant U.S. Government entity if a model requires stronger protections than the ASL-2 Standard" **Public transparency reports:** Anthropic has regularly published comprehensive misuse reports which documents real-world cases of actors attempting to exploit Claude for malicious purposes, along with detection methods and enforcement actions taken. - August 2025 - "Threat Intelligence Report: August 2025" - March 2025 – "Misuse Monitoring and Response Report" Other: - Platform Security Transparency Hub provides some enforcement statistics including #banned accounts for Usage Policy violations, number of appeals processed, CSAM reports to NCMEC, and law enforcement requests. **Industry information sharing:** The Frontier Model Forum (FMF) announced an information-sharing agreement signed by member firms (incl. Anthropic, Google, Meta, and OpenAI) to facilitate sharing of threats, vulnerabilities, and capability advances specific to frontier AI. The agreement, narrowly scoped to manage national security and public safety risks (including CBRN and advanced cyber threats), covers three categories: (1) vulnerabilities and exploitable flaws that could compromise AI safety/security, (2) threats involving unauthorized access or manipulation of frontier models, and (3) capabilities of concern with potential for large-scale societal harm. Details on implementation and use are unclear [Frontier Model Forum, 2025].
OpenAI	**Serious incident reporting frameworks:** No information found **Red-line Government notifications commitments:** No information found **Public transparency reports:** Regular reports documenting their disruption of malicious uses of their AI systems. Comprehensive reports detail enforcement actions against state-affiliated threat actors and covert influence operations identify specific threat groups (e.g., Storm-2035, Spamouflage), quantify disruptions (accounts banned, operations terminated), and describe the tactics employed (phishing, malware development, influence campaigns, election interference). - Feb 2024 – "Disrupting Malicious Uses of AI by State-Affiliated Threat Actors" - May 2024 – "Disrupting a Covert Iranian Influence Operation" - Jun 2024 – "Update on Disrupting Deceptive Uses of AI" - Aug, 2024: "Disrupting a covert Iranian influence operation" - Oct 2024 – "Influence and cyber operations: an update" - Feb 2025 - "Disrupting malicious uses of our models" - Jun 2025 - "Disrupting malicious uses of AI" - Oct 2025 - "Disrupting malicious uses of AI" **Industry information sharing:** The Frontier Model Forum (FMF) announced an information-sharing agreement signed by member firms (incl. Anthropic, Google, Meta, and OpenAI) to facilitate sharing of threats, vulnerabilities, and capability advances specific to frontier AI. The agreement, narrowly scoped to manage national security and public safety risks (including CBRN and advanced cyber threats), covers three categories: (1) vulnerabilities and exploitable flaws that could compromise AI safety/security, (2) threats involving unauthorized access or manipulation of frontier models, and (3) capabilities of concern with potential for large-scale societal harm. Details on implementation and use are unclear [Frontier Model Forum, 2025]. Comments on incident response from index survey (Q31) [Response]: "OpenAI has developed and continues to improve incident response programs across key areas of its operations, and is likewise improving and iterating on AI safety incident-specific protocols that are tailored to our operations and technology. Our goal is to respond to incidents in a rapid, coordinated way. [...] Incident Response Capabilities include (1) **Technical Controls for Rapid Mitigation:** We maintain the ability to rapidly roll back model deployments globally and to apply restrictions on model functionalities (such as tool use or capability throttling) in response to emergent risks. The roll back mechanism was successfully utilized within the last year in response to our finding that a GPT-4o model update was overly flattering or agreeable (see Sycophancy in GPT-4o: what happened and what we're doing about it) (2) **Incident Response Planning and Structure:** OpenAI has formal incident response plans for key areas of operations, including AI safety incident-specific protocols. Our response activities include escalation thresholds and mechanisms as well as incident response functions, such as response leads and as on-call rotations across functions to support implementation of response activity. We maintain close coordination across research, engineering, safety, legal, communications and policy teams, and have integrated lessons learned into our formal plans.
Google DeepMind	**Serious incident reporting frameworks:** No information found **Red-line Government notifications commitments:** Frontier Safety Framework 3.0 states that "If we assess that a model has reached a CCL that poses an unmitigated and material risk to overall public safety, we aim to share relevant information with appropriate government authorities where it will facilitate safety of frontier AI," a commitment it has kept from the last version of the Frontier Safety Framework 2.0. [Google, 2025]. **Public transparency reports:** Relevant publications: - 'Adversarial Misuse of Generative AI' (January 2025) - Detailed how threat actors—from scammers to state-aligned groups—attempt to misuse Google Gemini in deception, persuasion, and cyber operations. Described mitigation strategies and detection tooling [Google 2025]. **Industry information sharing:** The Frontier Model Forum (FMF) announced an information-sharing agreement signed by member firms (incl. Anthropic, Google, Meta, and OpenAI) to facilitate sharing of threats, vulnerabilities, and capability advances specific to frontier AI. The agreement, narrowly scoped to manage national security and public safety risks (including CBRN and advanced cyber threats), covers three categories: (1) vulnerabilities and exploitable flaws that could compromise AI safety/security, (2) threats involving unauthorized access or manipulation of frontier models, and (3) capabilities of concern with potential for large-scale societal harm. Details on implementation and use are unclear [Frontier Model Forum, 2025].
Meta	**Serious incident reporting frameworks:** No information found **Red-line Government notifications commitments:** No information found **Public transparency reports:** Meta consistently issues quarterly integrity reports about its platforms [Meta, 2024], these include reports on disrupting adversarial threat such as influence operations [Meta, 2025]. No reports for frontier AI models available. **Industry information sharing:** The Frontier Model Forum (FMF) announced an information-sharing agreement signed by member firms (incl. Anthropic, Google, Meta, and OpenAI) to facilitate sharing of threats, vulnerabilities, and capability advances specific to frontier AI. The agreement, narrowly scoped to manage national security and public safety risks (including CBRN and advanced cyber threats), covers three categories: (1) vulnerabilities and exploitable flaws that could compromise AI safety/security, (2) threats involving unauthorized access or manipulation of frontier models, and (3) capabilities of concern with potential for large-scale societal harm. Details on implementation and use are unclear [Frontier Model Forum, 2025].
xAI	**Serious incident reporting frameworks:** No information found **Red-line Government notifications commitments:** No information found **Public transparency reports:** xAI mentions in its RMF that it aims for "public transparency" about its risk management policies and intends to publish updates but has not mentioned whether it is going to publish misuse and model misalignment report. **Industry information sharing:** There is no publicly visible evidence that xAI systematically shares incident-data or model-failure information with industry partners.
DeepSeek	Article 14 of the **Interim Measures for the Management of Generative Artificial Intelligence Services (2023)** requires providers to promptly remove or disable unlawful AI-generated content, retrain or adjust their models where necessary, and report both the incident and any user misuse to relevant authorities. While not directly tied to catastrophic or frontier-safety events, it establishes a government-facing incident-reporting system for information-integrity compliance. **Deep-Synthesis Provisions (2023)** regulates that service providers of deep synthesis technology must remove illegal or harmful synthetic content, preserve records and "timely" report the incident to the CAC and other competent departments.
Z.ai	Article 14 of the **Interim Measures for the Management of Generative Artificial Intelligence Services (2023)** requires providers to promptly remove or disable unlawful AI-generated content, retrain or adjust their models where necessary, and report both the incident and any user misuse to relevant authorities. While not directly tied to catastrophic or frontier-safety events, it establishes a government-facing incident-reporting system for information-integrity compliance. **Deep-Synthesis Provisions (2023)** regulates that service providers of deep synthesis technology must remove illegal or harmful synthetic content, preserve records and "timely" report the incident to the CAC and other competent departments Its survey response (Q31) has indicated that the company has implemented the following capability: (1) Maintained and tested technical capability to rapidly roll back a deployed model to a previous version globally (within 12h). (2) Successfully tested rapid full model rollback including internal deployments within the last 12 months. (3) Maintained and tested technical capability to rapidly tighten model safeguards and restrict specific capabilities (e.g. web-browsing) globally. (4) Successfully tested rapid throttling or capability-restriction including internal deployments within the last 12 months. (5) Conducted at least one full live emergency response drill/simulation in the past 12 months. (6) Created a formal, documented emergency response plan for AI safety incidents with threshold for triggering emergency response, a named incident commander, and a 24 × 7 duty roster. (7) Established a risk-domain-specific (e.g. bio, cyber) 24-hour communication protocol and points of contact with relevant government agencies.
Alibaba Cloud	Article 14 of the **Interim Measures for the Management of Generative Artificial Intelligence Services (2023)** requires providers to promptly remove or disable unlawful AI-generated content, retrain or adjust their models where necessary, and report both the incident and any user misuse to relevant authorities. While not directly tied to catastrophic or frontier-safety events, it establishes a government-facing incident-reporting system for information-integrity compliance. **Deep-Synthesis Provisions (2023)** regulates that service providers of deep synthesis technology must remove illegal or harmful synthetic content, preserve records and "timely" report the incident to the CAC and other competent departments.