ISSUE BRIEF

FRONTIER MODEL FORUM

Issue Brief: Information Sharing, Incident Reporting, and Incident Response for Frontier AI Risks

DATE

May 12, 2026

INTRODUCTION

Frontier AI model capabilities are advancing at a rapid pace. While model capabilities have the potential to provide powerful benefits to society, they may also introduce or exacerbate large-scale risks to public safety and critical infrastructure. These risks range from misuse in biological and cyber domains to more systemic concerns around autonomous action and loss of control. Managing and understanding these risks effectively requires appropriate information flows across the ecosystem, ensuring timely notification of actionable information to the right sets of actors.

As frontier AI capabilities evolve, robust and mutually reinforcing information channels may play an important role for managing frontier AI risk across the ecosystem. Over time, these channels should facilitate improvement of safety knowledge across the ecosystem and enable timely response to emerging issues and incidents. Industries such as aviation and nuclear energy have grappled with analogous challenges and developed critical insights and frameworks for structuring information flows under conditions of uncertainty and high consequence. The field of frontier AI safety and security is still in the early stages of developing these mechanisms, presenting an opportunity to draw on those lessons and establish an appropriate architecture from the outset.

One of the core challenges to building appropriate mechanisms is conceptual: information sharing, incident reporting, and incident response are often treated or referred to interchangeably. In reality, they are distinct, complementary tools that serve different purposes. Conflating them risks producing mechanisms that are poorly suited to their intended function, whether too broad to be actionable or too

narrow to enable meaningful oversight. The effectiveness of these mechanisms will depend significantly on how they are structured, making it critical that they be precisely scoped and tailored to specific impacts and outcomes.

This issue brief aims to provide a high-level overview of these key concepts, including their differences and intended roles. It also highlights the Frontier Model Forum's (FMF) ongoing information-sharing efforts and offers preliminary recommendations for building robust and effective information channels for frontier AI risk management.

CONCEPTS AND CORE DISTINCTIONS

Establishing effective information channels for frontier AI risks will require clarity about what different mechanisms are designed to do, who they involve, and what obligations they carry. Although these mechanisms are complementary and may operate in parallel, conflating them risks creating systems that are poorly scoped, difficult to enforce, or counterproductive to the collaborative safety culture they are meant to support.

Information Sharing

Information sharing refers to the ongoing exchange of safety-relevant knowledge among frontier AI developers, researchers, governments, and civil society. Information sharing is generally proactive: participants share findings, threat intelligence, and lessons learned in anticipation of future risks rather than in response to a defined triggering event. Information sharing is also typically bidirectional: parties across sectors contribute to and benefit from shared knowledge pools. Because of this reciprocal character, information sharing is heavily trust-dependent. Further, its value erodes quickly if participants fear that sharing will expose them to regulatory penalty, competitive disadvantage, or legal liability. Legal ambiguity around what can and cannot be shared can further complicate or undermine information-sharing efforts. This points to the importance of establishing clearer frameworks and protections to enable and incentivize responsible information sharing.

Information sharing can occur between frontier AI developers, or across sectors, with the goal of pooling collective knowledge on specific risks. Engagement with expertise from adjacent domains (particularly cybersecurity, where threat intelligence is more mature) can add significant value by bringing in external perspectives and insights that individual companies may lack. Importantly, information sharing may also occur on a voluntary basis between frontier AI developers and government. A developer may proactively brief a regulatory agency or interagency body on a novel capability, an emerging threat model, or preliminary findings from internal red-teaming. This exchange may be informal, ongoing, and shaped by mutual agreement about what is useful to share to advance a shared safety goal.

Incident Reporting

Incident reporting refers to a formal notification submitted to a designated government body or authority that a qualifying AI safety or security incident has occurred. Whereas information sharing is typically ongoing and relationship-based, incident reporting is reactive, triggered by a specific event that meets defined criteria. It is also generally unidirectional (from the reporting entity to the receiving authority) and compliance-driven, meaning that the obligation to report arises from law, regulation, or order/agreement rather than voluntary choice (though reporting to regulators and other government authorities can also be voluntary).

Because incident reporting is triggered by event thresholds, the design of any reporting regime depends on how those thresholds are defined. The definition and appropriate scope of what is considered an 'incident' depends on regulatory context, risk categories, and the capabilities of the systems involved. To be effective, incident-reporting frameworks should adopt clearly scoped definitions of reportable events grounded in an evidence-based taxonomy of harms.

It is important to note that not every incident that arises from the use of an AI system will necessitate reporting under incident-reporting mechanisms. One distinction that any incident-reporting regime should address is the boundary between reportable incidents and lower-severity incidents or precursor events. Not every anomaly, near-miss, false positive or concerning signal will meet the threshold for formal incident reporting. These lower-risk incidents or precursor events may nonetheless be safety-relevant, and voluntary information-sharing channels are generally better suited to surfacing and analyzing them. A well-designed information ecosystem on frontier AI risks will create clear pathways for both categories without conflating them or inadvertently discouraging voluntary sharing of early-warning signals through overly broad mandatory reporting triggers.

Incident Response

In the context of AI safety and security, incident response refers to the coordinated process of containing, mitigating, and recovering from an incident once one has been identified. It is real-time and operational in character, often unfolding under significant time pressure and uncertainty. While incident response is primarily led by the affected developer's technical, legal, and security teams, it may involve a range of external parties as well. External support in this context may include outside legal counsel, forensic and technical consultants assisting with root-cause analysis, cybersecurity contractors engaged to assess whether an incident involves adversarial exploitation, and, in serious cases, government agencies or sector-specific bodies the developer coordinates with under incident reporting requirements. The specific composition of the response team will vary with the nature and severity of the incident. As a general matter, incident response processes can also cover other categories of risk.

Incident response is resource-intensive and benefits substantially from advance preparation and designated internal teams. Organizations that have developed response playbooks, tested escalation procedures, and pre-established relationships with relevant external parties are better positioned to respond effectively than those that build these capacities under pressure. Incident response approaches may also benefit from pairing automated monitoring with manual reviews, and may inform the calibration of mitigations.

Table 1: Key Dimensions of Comparison

Dimension	Information Sharing	Incident Reporting	Incident Response
Timing	Ongoing	Post-incident, within defined window	Real-time after incident identification
Direction	Multi-directional	Uni-directional, typically industry → authority	Depends on context
Purpose	Collective learning and prevention	Awareness and oversight of ecosystem safety and security	Containment & remediation
Actors	Industry, researchers, government	Regulators, oversight bodies	Internal teams, affected parties

FMF's Approach to Information Sharing

The Frontier Model Forum was established with an explicit mandate to build trusted, secure information-sharing mechanisms across industry, government, and civil society. In March 2025, FMF member firms signed a first-of-its-kind voluntary agreement to share safety-relevant information about frontier AI risks.

Operationalizing this agreement required building the institutional infrastructure to make meaningful sharing possible. FMF developed both legal and technical architecture designed to enable candid exchange while protecting intellectual property and ensuring antitrust compliance. As a first step, sharing has been piloted among FMF's member companies: Anthropic, Amazon, Google DeepMind, Meta, Microsoft, and OpenAI. These are the organizations at the frontier of AI development and the ones most directly engaged in building the technical solutions needed to address the risks covered by the agreement.

The scope of this agreement is deliberately narrow and reflects the principles outlined in this brief. Rather than facilitating broad sharing across the AI landscape, it

targets specific categories of safety- and security-relevant information tied to frontier AI's most serious and distinctive risks, the kinds that are difficult to address through existing channels designed for more general AI. By anchoring the agreement to concrete, well-defined categories, FMF ensures that what gets shared is actionable and that participation remains sustainable. Shared information must relate to one of three defined risk categories (see Table 2).

This narrow scoping reflects a deliberate theory of how information-sharing generates value. By focusing on a well-defined set of risks, FMF is able to bring together the organizational teams directly working to build technical solutions to those risks and facilitate the exchange of information they actually need. Deliberate scoping makes sharing more efficient, more actionable, and more likely to sustain the participation that gives the mechanism its value.

Table 2: Categories of FMF Information Sharing

Type of Information	Description
Vulnerabilities, weaknesses, and exploitable flaws	Vulnerabilities, weaknesses or exploitable flaws may compromise the safety, security, or intended functionality of frontier AI models. Examples may include jailbreaks, adversarial inputs, data poisoning, or other attempts to bypass model safeguards.
Threats	Threats to frontier AI comprise threats directed to the unauthorized access or manipulation of frontier AI models. Examples may include potential threat actors, attack vectors, or cyber-threat indicators.
Capabilities of Concern	Capabilities of concern refer to frontier AI capabilities that have the potential to cause large-scale harm to society. Examples may include capabilities related to the development of chemical, biological, radiological, and nuclear (CBRN) threats, offensive cybersecurity attacks, and model autonomy.

The agreement is equally deliberate about how shared information can be used. Recipients are permitted to apply it only within defined categories, all explicitly oriented toward making frontier AI more safe and secure.

The results from our first year are promising. Information shared through this initiative has led to concrete safety and security improvements across member firms, and has helped develop practical resources for addressing misuse risks.

Looking ahead, FMF is working to expand this model to include other frontier AI companies, explore appropriate information sharing with government partners, and use this infrastructure to develop, host and distribute a broader range of safety tools

Table 3: Scope of Covered Use

Use Categories	Description
Threat Detection and Security	Improving threat detection systems and protocols, and developing threat intelligence capabilities such as indicators of compromise and threat actor profiling, solely for safety and security purposes
Safe Development of Models	Informing safety-specific model development practices, including through synthetic data generation for fine-tuning to address known safety gaps
Incident Response	Strengthening incident response plans by identifying threat-specific procedures and countermeasures in advance, limited to safety and security incidents
Model Evaluations	Informing safety and security evaluations, including automated benchmarks and red-teaming exercises, limited to identifying and mitigating safety- and security-relevant risks
Safeguards	Improving defenses against known attack methods, such as adversarial inputs or data poisoning, for safety and security purposes only
Safety Research	Advancing AI safety research to promote responsible development and enable independent capability and safety evaluations
Safety Guidance Development	Informing internal policies, governance frameworks, and compliance procedures directly related to AI safety and security
Monitoring and Detection Systems	Enhancing systems that detect anomalous behavior, unauthorized access, or potential misuse, for safety and security purposes only
Vulnerability Assessments	Identifying and remediating vulnerabilities or security gaps in an organization's own systems, limited to safety- and security-relevant assessments

and resources across the ecosystem. The aim is to raise the baseline of AI safety knowledge industry-wide and to ensure our information sharing remains interoperable with existing information-sharing initiatives on more general AI and emerging mechanisms for incident reporting and response.

EMERGING PRACTICES

The following recommendations reflect emerging guidance for designing information architecture to best address frontier AI risks. The ecosystem for frontier AI incident response and information sharing continues to be nascent, given the emerging and evolving nature of frontier AI risks and related taxonomies. As the ecosystem evolves, it will be important to identify where existing frameworks and institutions can be adapted and where frontier AI may require distinct solutions.

Design information-sharing institutions deliberately. Effective information sharing requires institutional infrastructure that is purpose-built for the task. This means clearly defining the objectives of any sharing regime at the outset, identifying which stakeholders need access to what kinds of information, and establishing governance structures that sustain participation over time. Vague or open-ended mandates create risks on both ends of the sharing spectrum: they can produce over-sharing that burdens participants with low-value noise (information that arrives without context, prioritization, or clear relevance to the recipient's decision-making) and under-sharing that fails to reach stakeholders who most need it, because no one has been clearly designated to receive or act on a given category of information. Getting the architecture right from the start is critical, before trust has eroded or participation has declined.

Define incident reporting with precision. Incident reporting serves a different function than ongoing information sharing, and should be scoped accordingly. Incident reporting mechanisms should be designed to surface meaningful signals. This requires definitions of reportable incidents that are specific enough to be consistently applied, tied to concrete thresholds or harm categories, and developed with input and insights from those who will be obligated to report. Poorly scoped definitions create several compounding problems. Overly broad definitions risk generating a volume of reports that obscure genuinely actionable signals, making it harder to identify and respond to serious harms; vague or ambiguous definitions will also produce inconsistent reporting across the industry, as individual organizations are left to interpret open-ended criteria differently, undermining comparability and eroding confidence in the data.

Build incident response procedures and capacity carefully. Reporting and response are related but distinct. Having a mechanism to receive reports is not the same as having the capacity to act on them. Institutions should invest in response infrastructure (including designated personnel, clear protocols, and internal coordination mechanisms) commensurate with the volume and severity of incidents they expect to receive. Response capacity should be developed in parallel with

reporting obligations.

Leverage existing institutional strength for incident response. Institutions should leverage personnel, practices, policies and internal structures for incident response in non-AI areas to build and strengthen their incident response infrastructure related to frontier AI risk.

Avoid common conflation pitfalls. Analogous domains have surfaced many lessons learned on how to develop effective information mechanisms along with pitfalls that should be actively guarded against.

- *Avoid designing information-sharing regimes that inadvertently trigger reporting obligations.* If participation in an information-sharing forum exposes contributors to regulatory reporting requirements, organizations will rationally limit what they share, discouraging the open exchange it was meant to foster. Sharing and reporting mechanisms should be carefully scoped to avoid this dynamic.
- *Establish reporting timelines that allow for thorough analysis.* Timelines for incident reports are often set with well-intentioned urgency, but if they implicitly assume that root-cause analysis or full impact assessment has already been conducted, they will either produce premature and unreliable reports or place unrealistic burdens on reporters.
- *Avoid assuming reporting mandates substitute for response capability.* Requiring organizations to report incidents does not automatically generate the capacity to respond to them.

CONCLUSION

Information sharing, incident reporting, and incident response are each important components of a responsible information architecture for addressing frontier AI risks and promoting responsible development. However, the value of these mechanisms depends on how they are designed. Poorly scoped information-sharing regimes, imprecise and broad reporting obligations, and underdeveloped response capacity may risk undermining the ecosystem they are meant to support.

The field is at an early and important juncture. The norms and institutions taking shape now will be difficult to revise if and when a major incident forces reactive decision-making under pressure. Getting the architecture right from the outset, by treating these three tools as distinct and building each with precision, is critical. The FMF aims to share more updates and lessons learned from our information-sharing function over the coming months.