THE ATHENS ROUNDTABLE
ARTIFICIAL INTELLIGENCE AND THE RULE OF LAW

THE FUTURE SOCIETY

Serious AI Incident Prevention & Preparedness

Key insights from the closed-door dialogue session hosted at the **Seventh Edition of The Athens Roundtable on AI and the Rule of Law**

December 4, 2025, in London, UK

FOREWORD

On December 4th 2025, the Seventh Edition of The Athens Roundtable on AI and The Rule of Law convened over 140 global leaders and experts from diverse sectors—including government, industry, civil society, and academia—in London to address urgent challenges in AI governance under the theme "**Facing the Stakes of AI Together: From Shared Concerns to Joint Action**".

The agenda emerged from a multi-month process of policy research, advocacy, and coalition-building led by The Future Society in 2025. In particular, it built upon two preliminary events: the Global Call for AI Red Lines campaign that The Future Society co-presented at the United Nations General Assembly in September 2025, and a **GPAI OECD.AI Expert workshop on AI Incidents** that The Future Society co-delivered in November 2025. You can read the **full post-event synthesis report for the Seventh Edition of the Athens Roundtable** here.

This year's edition was split into two halves. The morning plenary sessions examined current and anticipated frontier AI capabilities alongside emerging policies across jurisdictions, taking a critical look at what is working and what is needed next as existing efforts face real tests of relevance and durability. In the second half of the day, dedicated closed-door dialogue sessions focused on two urgent priorities: **(1) building a shared understanding of unacceptable AI risks** and **exploring viable diplomatic pathways for establishing measurable and enforceable red line thresholds** (building upon the Global Call for AI Red Lines), and **(2) clarifying the institutional capacities and infrastructure needed for serious AI incident prevention and preparedness** (building and expanding upon OECD's work on AI incident monitoring and reporting).

These dialogue sessions brought together curated groups of experts over several hours, and while under Chatham House rule of non-attribution, their contents were transcribed, anonymized and analyzed by The Future Society team with the help of the deliberAIde platform.

These two priorities are **interconnected**: while red lines can define the theoretical thresholds of unacceptable risks posed by specific AI use cases, behaviors or capabilities, incident monitoring and response infrastructure can provide the empirical evidence required to identify, refine, and enforce them.

Thus, concurrently advancing these two priorities can ensure that global safety boundaries remain responsive to real-world harms and emergent AI capabilities, transforming static red line prohibitions into a dynamic, adaptive governance architecture. This document synthesizes the **key insights that surfaced from the dialogue session focused on serious AI incidents**.

Under the auspices of H.E. the President of the Hellenic Republic, Mr. Constantine An. Tassoulas

ORGANIZED BY: THE FUTURE SOCIETY

IN PARTNERSHIP WITH: OECD, unesco, AI IMPACT SUMMIT भारत 2026 INDIA

WITH SUPPORT FROM: Fathom, Arnold&Porter, Cooperative AI Foundation, deliberAIde

TABLE OF CONTENTS

CONTEXT

The increased integration of powerful AI systems into critical infrastructure brings the inevitable risk of serious incidents or accidents, potentially leading to society-wide harm. In the past few years, public concern about the harmful impacts of AI technologies has intensified as evidence of AI-related incidents has mounted. For example, AI companion chatbots have been implicated in teenage suicide cases, engaging vulnerable children in harmful exchanges and causing what has been termed "*AI psychosis*".

Malicious uses of AI have also been increasing exponentially. In late 2025, state-sponsored bad actors began exploiting AI tools to conduct autonomous cyber espionage campaigns. These cases, while among the most visible, represent only a subset of a broader and growing set of incidents that are actively monitored by researchers, companies, and civil society. Frontier AI companies and other experts continue to warn that increasingly capable models enable far more severe risks, yet preparedness for such risks remains inadequate. Therefore, there is an urgent need to build up existing institutional capacities to detect, investigate and learn from such AI incidents before they can cascade into severe, society-wide crises.

Motivated by these concerns, the closed-door dialogue session on AI incidents that took place at the Seventh Edition of the Athens Roundtable convened 36 experts—spanning government, industry, civil society, and academia from across 12 countries, including representatives from multilateral organizations including the OECD and UNESCO, national AI safety and security institutes, and frontier AI companies—to collaboratively assess the current state of risk management practices and incident prevention and preparedness infrastructure for powerful AI systems, identifying what is needed to strengthen them. The session was informed by, and built upon, the GPAI OECD.AI Expert Workshop on AI Incidents held on November 14 2025 in Paris, and focused on unpacking the following critical question: ***What infrastructure do governments, industry and civil society need to monitor and report serious AI incidents, prevent their escalation to potential emergencies or crises, and respond effectively when prevention fails?***

The session was structured around three objectives:

1. **Map the gaps** in existing incident monitoring and reporting mechanisms;
2. **Identify the technical, institutional, and political barriers** that prevent effective incident prevention and preparedness (for risks that may materialize) and response (to major accidents and serious incidents that have already occurred);
3. **Develop actionable recommendations** for different stakeholder groups.

This report highlights the main insights that surfaced from those discussions and offers actionable recommendations that can spur and strengthen collective incident prevention, preparedness and response capacity. The discussion insights and recommendations listed throughout do not represent consensus among participants, but are ideas that surfaced and gathered notable support during group discussions.

KEY DISCUSSION INSIGHTS & RECOMMENDATIONS

DEFINITIONS OF "AI INCIDENTS" ARE INCONSISTENT AND UNCLEAR

Participants identified that a key challenge associated with managing AI incidents is rooted in a fundamental methodological and legal difficulty: the definitional boundaries of what constitutes an "incident" are not clear. These boundaries fluctuate significantly based on deep-seated disciplinary and institutional differences regarding semantics, values, and law. This results in a "*moving boundary*" problem where reporting duties can be under-inclusive (missing near misses and critical early warning signals) or over-inclusive (creating excessive noise), depending on how specific statutes, frameworks and standards operationalize the term *[see Annex A for a range of existing definitions]*. The definitional boundaries of what constitutes a 'serious AI incident' matters because it determines what harms are monitored, reported, and responded to. Participants identified three key tensions in defining incidents.

First, definitions tend to be either **damage-centered**, focusing exclusively on tangible, proven harms to health, property, or fundamental rights, or **concern-centered**, which also include near-misses and hazards that signal risk without causing tangible harm.

Second, definitions differ on whether incidents should be limited to **major accidents and technical malfunctions** or also include **deliberate misuse** of systems that function as intended (for example, AI deepfakes used for fraud).

Third, definitions can be either strictly restricted to **discrete incidents**, or can include broader **systemic impacts**; for example, while some argue that rapid, widespread automation-induced job displacement represents normal system functioning, others view it as a reportable societal harm or even crisis, highlighting the challenge of capturing diffuse impacts that lack a single, specific, identifiable causal origin.

RECOMMENDATIONS:

Governments and regulators should define AI incidents clearly, using interoperable, future-proof definitions paired with graduated reporting obligations, while actively shaping international taxonomies and adopting structured incident classification frameworks.

To balance the risks of under-reporting and over-reporting:

- **Policymakers** should **adopt broad, harm-oriented incident definitions** anchored in legislation and codes of practice that are robust to evolving technologies and incentives. Such definitions should be paired with risk-tiered obligations that scale with severity: mandating reporting for severe incidents while enabling voluntary channels for near misses and early warning signals (which can help strengthen incident prevention and preparedness). Governments should also participate actively in international taxonomy development so that definitions are shaped according to domestic priorities rather than simply inherited passively.
- **Regulatory bodies** should operationalize these definitions through **structured classification systems** such as the Goals-Methods-Failures (GMF) Taxonomy, linking system objectives and technical design and deployment choices to observed failures to uncover causal pathways and improve incident interpretability, consistency, and response.

INCIDENTS ARE DETECTED TOO LATE

Current approaches for identifying AI incidents remain largely reactive and fragmented, with incidents surfacing primarily through journalism, social media, and litigation. This has prompted calls for more systematic monitoring infrastructure, including mandatory incident reporting to regulators, independent third-party auditing, and protected channels for whistleblowers to disclose safety concerns. Yet earlier detection is only valuable if it enables meaningful intervention and effectively prevents individual incidents from escalating into widespread crises or emergencies. Reporting requirements that lack enforcement mechanisms document incidents faster without preventing them, while transparency frameworks presume that regulators possess both the technical expertise to interpret what developers disclose and the legal authority to act on it – presumptions that rarely hold in practice.

For many categories of AI harms such as serious AI incidents and major accidents, emergencies or crises, the detection-then-response model may be structurally inadequate regardless of how quickly incidents are identified. Policymakers and other stakeholders are beginning to grapple with these challenges, though critical questions remain unresolved. For instance, while the EU AI Act combines conformity assessments with post-market monitoring obligations and California's SB 53 pairs incident reporting with mandatory safety frameworks and whistleblower protections, neither framework settles the deeper questions at hand:

1. Whether disclosure should trigger **automatic regulatory consequences** or merely contribute to an **evidentiary record**;
2. How **liability should be distributed** across the AI supply chain—among developers, deployers, and end users;
3. Whether **voluntary commitments** from industry can ever **substitute for enforceable legal obligation**.

RECOMMENDATIONS:

Governments, industry, multilateral institutions, and the public should strengthen and normalize comprehensive AI incident reporting across the AI lifecycle.

Governance of powerful AI systems depends on the **timely identification, documentation, and sharing** of AI incidents and near-misses across jurisdictions and sectors. As a result:

- **Governments** should **systematically document incidents** to justify proportionate, evidence-based intervention and expanded oversight. This could be achieved, for example, via the establishment of 'National AI Incident Databases' (as recently announced by the Government of India) that record, classify, and analyze risks such as safety failures, biased outcomes, security breaches, and malicious use. By mandating reports from public bodies, private entities, researchers, and civil society, such national databases can facilitate rigorous post-deployment monitoring for the comprehensive detection of incidents and systemic trends, and can thereby enable targeted regulatory interventions that promote stronger incident prevention and preparedness practices and thereby prevent incidents from escalating into more serious emergencies or crises.
- **Multilateral institutions** and fora should develop or encourage **shared incident repositories** to enable cross-jurisdictional monitoring, collective learning, and early detection of emerging patterns of harm before they cascade.
- **Legislators** should mandate **comprehensive post-deployment monitoring** by establishing legal requirements for the continuous oversight of high-risk AI systems throughout their operational lifecycle. Because frontier models can exhibit emergent behaviors or "*model drift*" after release, static pre-deployment audits are insufficient. Mandates should include compulsory real-world performance tracking, automated incident reporting channels, and periodic third-party 'health checks' to ensure systems remain within safety parameters long after they are externally deployed.
- **Legislators** should **strengthen** whistleblower protections to explicitly cover AI hazards (i.e. anticipated or foreseen AI harms) and clarify that non-disclosure agreements (NDAs) are legally void when used to prevent the reporting of such hazards to competent authorities.
- **Industry** actors should **complement public oversight** by establishing protected escalation mechanisms outside normal management hierarchies, ensuring that staff can raise concerns when internal command structures are unresponsive.
- Members of the **public** should be **encouraged to report AI harms** through accessible channels, including regulators, consumer protection agencies, social media platforms, and publicly available incident databases. Routine, multi-level incident reporting can shift AI governance from ad hoc, crisis-driven responses toward a more anticipatory and evidence-based model, with lessons over time informing more formalized reporting obligations and coordination mechanisms.

ACCOUNTABILITY IS FRAGMENTED ACROSS THE VALUE CHAIN

The AI value chain involves multiple actors, assigned different responsibilities, each of whom shapes how a system behaves and each of whom may contribute to downstream harm. When AI incidents or major accidents occur, distribution of agency creates a situation where responsibility diffuses across so many actors that no single party can be held obviously accountable, and victims of harm struggle to identify whom to seek redress from. The problem is compounded by information asymmetries that run in both directions. Upstream developers possess technical knowledge about model behavior that downstream deployers lack, while deployers observe real-world impacts that developers never see. Contractual confidentiality provisions typically prevent parties from accessing each other's data, meaning that when harm occurs, no actor can trace whether the problem originated in the base model, the fine-tuning process, the integration layer, or the deployment context. Each party can plausibly claim the harm originated elsewhere.

Existing frameworks attempt to address this fragmentation, though with limited success. For instance, the **EU AI Act's Article 25** establishes that any party who substantially modifies a high-risk AI system, or places their name on it, assumes provider obligations – a mechanism designed to prevent responsibility from being contracted away. However, the Act's definitions turn on functional criteria that remain contested in practice, and downstream actors often lack the technical capacity to assess whether their modifications trigger provider status. The European Commission's proposed AI Liability Directive, which would have eased the evidentiary burden on claimants by creating a presumption of causation, was withdrawn in early 2025, leaving the underlying problem unresolved. The deeper challenge is that fault-based liability – which is the default in most jurisdictions – requires claimants to prove that a specific actor acted negligently. This evidentiary burden becomes nearly insurmountable when harm emerges from the interaction of decisions made by multiple actors across the value chain, none of whom may have acted carelessly in isolation. What remains clear is that **accountability frameworks must move beyond voluntary disclosure toward enforceable obligations** that specify minimum responsibilities for each role in the value chain and establish clear consequences when those responsibilities are not met.

RECOMMENDATIONS:

Governments should clarify AI value-chain accountability, with public engagement reinforcing the legitimacy and enforcement of these obligations.

Accountability for AI-related harms requires **clear allocation of responsibility across the AI value chain**, spanning developers, integrators, and deployers, to prevent risk from being diffused or deflected between actors. Accordingly:

- **Governments** and legislators should **specify minimum obligations for each role**, including duties related to risk assessment, incident response, and post-deployment monitoring, and consider strict liability regimes for categories of serious AI incidents and major accidents where proving fault is impractical or prohibitively difficult. Such clarity can reduce regulatory uncertainty, strengthen incentives for upstream incident prevention and risk mitigation, and improve access to redress when harms occur.
- Public engagement and **civic advocacy** should **play a complementary role** by sustaining political attention and pressure for robust accountability frameworks. Together, clear legal responsibility and sustained public scrutiny can help ensure that accountability for AI systems is both enforceable and socially legitimate.

DEPLOYED SYSTEMS ARE INADEQUATELY DOCUMENTED

Effective incident response requires knowing what AI systems are deployed, and by whom. In the absence of ***mandatory disclosure***, publicly deployed systems often remain known only to the developers and deployers who manage them. When AI incidents occur, investigators frequently cannot determine which system was involved, how it was modified from its base model, or where else it might be deployed. This information gap prevents the broader ecosystem from learning which architectures, training approaches, or deployment contexts carry elevated risk – knowledge that could inform both technical improvements and regulatory priorities.

Emerging frameworks attempt to address this gap. For example, the EU AI Act requires providers of high-risk AI systems to register in a public database before deployment, and mandates technical documentation covering system design, training data, and evaluation results. In the US, federal agencies are required to establish and annually update **public-facing AI use case inventories** of their AI current and planned AI deployments.

However, these efforts remain partial, with enforcement mechanisms remaining largely untested. More fundamentally, documentation requirements focus on the system as designed rather than as deployed – yet harmful behaviors often emerge from interactions between the model, its fine-tuning, its integration context, and its actual use patterns, none of which static documentation can fully capture.

This challenge extends to **internal AI deployments** within frontier AI companies. Increasingly capable models are deployed internally for research and development purposes before any public release, yet these internal uses are not subject to external documentation requirements. This matters because internal deployment is often where emergent capabilities and dangerous behaviors first manifest.

Such capabilities or behaviors can lead to pre-deployment incidents or near-misses that could serve as valuable early warning signs for harms that could materialize from the public deployment of a particular AI system. AI systems should be **designed for auditability from the outset** by documenting development decisions, maintaining evaluation records, and enabling external reproducibility where feasible - which if absent, oversight remains dependent on what companies choose to disclose.

RECOMMENDATIONS:

Governments and industry should establish robust, standardized documentation and registration requirements for high-stakes AI deployments, while the public should demand transparency from organizations and governments to ensure accountability.

Oversight of AI systems depends on the ability to **trace them from development through deployment** and to understand their intended uses, limitations, and modification histories. Therefore:

- **Governments** and legislators should **mandate the public registration of AI deployments in high-stakes domains**, specifying required information (including developer, deployer, intended use, known limitations, and modification history), retention periods, and conditions for regulator or affected-party access.
- For the **most powerful AI systems**, policymakers should additionally **require documentation of internal deployments and pre-release evaluations**, made available to regulators under confidentiality protections, to enable oversight of emergent capabilities before public release. To support this, standardized documentation practices – such as model cards, system cards, and training-data datasheets – should be adopted and extended to internal research systems and internal deployments, designed as functional records that enable tracing system behavior across the value chain.
- **Industry** actors should **implement these requirements in practice** by documenting public deployments comprehensively and recording internal uses and model characteristics, including differences between internal and public versions, to enable precautionary action by external stakeholders and strengthen incident preparedness.
- Members of the **public** can further reinforce these measures by **actively requesting information from organizations and governments** regarding past AI-related incidents and the use of AI in consequential decision-making processes. Pressuring these entities to release such data is vital for building the robust evidence base required for more effective governance. For instance, in the wake of tragic outcomes, the public can demand the release of specific chat histories or logs in response to cases involving self-harm or suicide, minimizing the chance of corporate or institutional secrecy to obscure pathways to accountability (though it is worth noting that the scope of information the public can access and demand is inherently limited, particularly when it comes to incidents involving national security concerns, where operational secrecy may continue to take precedence over information disclosure).

RESPONSE CAPACITY EXISTS ON PAPER ONLY

Many organizations have incident response plans, but these often remain untested and disconnected from operational realities. Unlike aviation and nuclear safety, where tabletop exercises and simulated emergency scenarios are mandatory and routine, **AI incident preparedness and response capabilities remain largely inadequate**. Staff may not have been trained on response procedures, or may have received training so long ago that (i) they **no longer remember it** well enough to follow protocols effectively in the event of incidents or (ii) these **protocols become obsolete** as a result of rapid capability advances. Dependencies across teams and organizations may not have been mapped. Communication channels may not have been tested under realistic conditions. And systematic methods for learning from incidents and accidents after they occur are currently either worryingly immature or entirely lacking.

AI incidents can unfold rapidly and at scale. A problematic system deployed across millions of devices can cause widespread harm before anyone recognizes what is happening. Effective response requires practiced capabilities: people who know what to do, strong relationships that enable rapid coordination among various stakeholders, and decision-making processes that can operate under time pressure. These take years to develop, which is problematic because ambiguity about authority during times of crisis wastes critical time.

RECOMMENDATIONS:

Governments, regulators, and industry should institutionalize AI incident preparedness and response capacity alongside reporting requirements, supported by multi-stakeholder testing and governance oversight.

AI governance frameworks often focus solely on incident reporting while **neglecting incident prevention**, **preparedness** and **response capacity**, which limits the ability of organizations to respond effectively when faced with incidents. Addressing this gap could strengthen societal resilience and reduce the likelihood that incidents cascade into broader, more severe emergencies or crises. To do so:

- **Governments** and legislators should **mandate incident preparedness as a core requirement,** including clear plans for response, escalation, coordination, and recovery, to ensure that major AI deployers and developers always demonstrate their incident response capacity prior to publicly deploying any high-risk AI system. This should include the establishment of emergency powers with explicit activation thresholds, defined leadership and coordination roles, and access to necessary resources, so that responses are timely, proportional, and accountable.
- **Regulatory bodies** should support this by **convening multi-stakeholder workshops** and tabletop exercises to test response capabilities, clarify cross-organizational dependencies, and identify gaps before real incidents expose them. Industry actors should build genuine response capacity through regular drills based on realistic scenarios, structured tabletop exercises, and systematic post-incident learning processes drawn from high-hazard industries.
- Incident prevention and preparedness requirements should also be **embedded within corporate governance**, with board-level obligations for significant AI risks so that major accident prevention and incident readiness are treated as operational governance matters rather than merely compliance exercises.

ENFORCEMENT ARCHITECTURE IS LARGELY ABSENT

AI Safety Institutes (AISIs) and their equivalents, established in several jurisdictions over the past two years, have been explicitly designed to provide technical support rather than to serve as supervisory authorities with enforcement powers. Their non-adversarial relationships with frontier AI companies, which grant them voluntary pre-deployment access to frontier models, would be undermined by enforcement functions. This highlights a structural gap, where the greatest technical expertise sits with bodies that cannot enforce tangible change within industry practices, while regulators with enforcement powers lack the technical capacity required to use them effectively. **Few established mechanisms currently connect AISIs to regulatory enforcement mechanisms**, thereby preventing policymakers and regulators from drawing on AISI expertise.

RECOMMENDATIONS:

Governments, regulators, and multistakeholder actors should strengthen the institutional capacity and coordination needed to interpret AI incident reports and translate them into effective oversight, while creating mechanisms that enable shared learning without exposing companies to competitive or legal risk.

To enable the effective enforcement of incident management requirements and best practices:

- **Governments** and legislators should **build multi-stakeholder alliances** comprising public and private sector entities, including civil society, academia, and affected communities, so that domestic voices can strengthen government positions in international discussions concerning incident prevention and management.
- National or regional **regulators** should **expand their technical regulatory capacities** to improve their ability to interpret incident reports and respond effectively, through hiring staff with relevant expertise, engaging external technical advisors, or creating formal communication channels with AISIs to be able to draw more effectively on their technical expertise.
- As recommended by the Trump Administration's AI Action Plan, **trusted public-private information sharing intermediaries**, functioning akin to Information Sharing and Analysis Centers (ISACs) in cybersecurity, should be established, who receive sensitive information and intelligence about AI incidents, accidents, near-misses, and hazards from multiple industry and government sources and share it in anonymized or aggregated form. Such intermediaries could **enable companies to learn from one another's experiences and inform governmental institutions** of problematic AI-driven security threats without exposing companies to competitive disadvantages or legal action.
- **Governments** and regulators should **clarify role delineation for incident monitoring and response** (for example, AISIs provide technical capacity while enforcement sits with regulators who maintain legally mandated relationships with regulated entities), and **establish formal coordination mechanisms** that preserve the distinct value of each function while ensuring effective incident management.

CROSS-SECTORAL LESSONS HAVE NOT BEEN SYSTEMATICALLY APPLIED

Other safety-critical domains – namely chemical manufacturing, aviation, nuclear power, financial services, and public health – have decades of experience building major accident prevention and risk management infrastructure, spanning structured reporting systems, accountability frameworks, cross-organizational learning, and response capabilities tested through regular exercises. Yet these lessons have not been systematically imported into AI governance.

Some risk management mechanisms from other domains could be adapted relatively easily. Public health surveillance systems, for instance, already produce data designed for consistent interpretation and tied to measurable outcomes; adding standardised AI-related codes to existing reporting mechanisms would require modest modifications rather than entirely new or substantial infrastructure. Financial services circuit breakers – predefined triggers that cause markets to temporarily halt all trading during periods of extreme market volatility – are a clear example, demonstrating that automated safeguards can prevent cascading harms. The fragmented nature of current AI incident prevention, preparedness and management efforts, where different actors report and respond individually rather than coordinating, reflects a failure to apply these established approaches.

RECOMMENDATIONS:

Frontier AI companies should adopt formal Major Accident Prevention Policies (MAPP) or Serious Incident Prevention Policies (SIPP), integrate AI incident tracking into existing reporting processes, and deploy operational circuit breakers to reduce the risk of AI incident escalation into severe emergencies or crises.

- **Frontier AI companies** should adopt a formal Major Accident Prevention Policy (MAPP) or Serious Incident Prevention Policy (SIPP), drawing directly from the rigorous safety standards of other high-hazard industries. Rather than vague statements of intent, such policies must explicitly identify catastrophic accident and incident scenarios (such as cascading algorithmic malfunctions or foreseen psychological effects on users) and detail the Safety Management Systems (SMS) designed to prevent them. This could shift AI incident management **from a culture of reactive firefighting to one of anticipatory governance**, where detailed safety reports and independent verifications are required before systems can be externally deployed.
- **Regulatory bodies** should require frontier AI companies and downstream deployers to **embed AI incident tracking systems and reporting requirements**, for instance by adding standardized AI-related codes and data fields to established reporting mechanisms. This approach would minimize administrative burden, leverage existing enforcement and compliance channels, and improve the speed and reliability of incident detection.
- **Frontier AI companies** should complement this by **implementing automated safeguards like operational circuit breakers**: predefined triggers that halt system operation when certain risk thresholds are exceeded, thereby preventing incident escalation and reducing the likelihood of cascading harm.
- Together, **integrated tracking and automated safeguards** can strengthen the ability of regulators and operators to detect, contain, and respond to AI incidents in a timely and proportionate manner.

EFFECTIVELY INCENTIVIZING INCIDENT DISCLOSURE IS DIFFICULT

Organizations currently face **inadequate incentives to disclose AI incidents**, whether internal or external, for disclosure brings reputational damage, legal liability, and regulatory scrutiny. Yet without disclosure, neither organizations themselves nor the broader ecosystem can address, learn from, or prevent future incidents. Participants flagged numerous problematic tensions regarding mechanisms through which disclosures could be effectively incentivized, with no single option appearing adequate by itself. **Punitive regulatory environments discourage disclosure altogether**, and other kinds of purely mandatory incident disclosure systems risk generating low-quality reports that are submitted purely for the sake of compliance rather than for genuine learning. **Purely voluntary systems**, on the other hand, **risk systematic underreporting of serious harms** for legal and reputational reasons, with antitrust concerns further suppressing coordination willingness among firms who might otherwise share safety-relevant information with one another.

RECOMMENDATIONS:

Governments should enable and incentivize industry sharing of safety-critical incident information by creating antitrust-safe harbours, and industry should pair incident reporting with robust corrective action and root-cause learning.

- To ensure that incident reporting leads to meaningful safety improvements, **governments** and legislators should issue **antitrust-safe harbors** (e.g. potentially the Frontier Model Forum) that allow organizations to **share safety-critical information** – such as incident details, near-misses, and effective mitigations – without competitive concerns. This would support collective learning and prevent companies from withholding information due to fear of legal or commercial disadvantage.
- **Industry** actors should complement this by conducting **comprehensive post-incident reviews** that identify specific **root causes** for particular incidents by investigating both model behaviors and company operations. Through such root-cause learnings, they should also start reporting incidents to governments, the public, and/or other companies together with specified corrective actions that demonstrate how similar incidents will be prevented in the future.
- Together, these measures can create a **feedback loop** in which incident data is used to drive **systematic learning and improve safety practices** across the AI ecosystem, thereby strengthening collective incident prevention and preparedness capacities.

WHETHER INCIDENT MANAGEMENT EFFORTS SHOULD BE PRECAUTIONARY OR PURELY REACTIVE IS CONTESTED

A fundamental tension was identified by participants between **precautionary approaches**, which call for the front-loading of governance requirements before harms occur and thereby prioritize *incident prevention*, and **reactive approaches**, which call for the incremental building of incident preparedness and response capacity based on evidence of actual incidents and accidents.

Proponents of precautionary approaches, motivated by the potential for AI systems to cause rapid, large-scale harms if not prevented, argued that deployers should be responsible for anticipating plausible incidents and for devising appropriate prevention and response plans prior to system deployment. After all, some harms, once they occur, cannot be undone and can quickly cascade into widespread emergencies or crises, and so waiting for problems to manifest could be severe. Proponents of reactive approaches, however, countered that precautionary requirements can be burdensome, speculative, and politically difficult to sustain, being labelled as "*overreach*".

RECOMMENDATIONS:

Frontier AI companies should shift their regulatory focus ***from reactive oversight to proactive accountability****.*

- **Governments** and legislators should **require frontier AI companies to systematically identify and document plausible incident types** specific to their systems both before and after external deployment. This documentation must be maintained and made readily available to regulators, ensuring that developers are held accountable for anticipating and mitigating potential high-risk failures rather than merely reacting to harms after they manifest.
- **Regulators** should **establish risk-tiered incident reporting and management obligations** that **scale with severity**. For instance, regulatory frameworks should mandate the immediate reporting of severe incidents and major accidents that have already occurred to ensure rapid containment and public safety. Simultaneously, they should establish and promote voluntary channels for reporting near misses and early warning signals, which can strengthen incident prevention and preparedness. By capturing these lower-stakes precursors, regulators and industry actors can identify systemic vulnerabilities and strengthen incident prevention and preparedness strategies before they escalate into catastrophes.

SECURING GLOBAL HARMONIZATION WITH RESPECT TO INCIDENT MANAGEMENT PRACTICES IS CHALLENGING

Most participants were proponents of **global harmonization and the establishment of cross-border infrastructure** with respect to incident prevention, monitoring and management. After all, AI systems are developed and deployed globally by companies operating across many jurisdictions, and so incidents in one country can occur as a result of models developed in another, running on infrastructure in a third. Given the reality that, **without global harmonisation, companies can exploit regulatory arbitrage** and incidents can fall through jurisdictional gaps, some participants argued for the necessity of establishing harmonised incident prevention, tracking and management practices through common definitions, shared reporting standards, and coordinated enforcement protocols. Yet other participants flagged that building cross-border incident prevention, monitoring and management infrastructure through de facto global standards is difficult to achieve in practice, since different jurisdictions tend to have different risk tolerances, institutional capacities, and geopolitical contexts.

RECOMMENDATIONS:

*Governments, multilateral institutions, and AI safety bodies should **coordinate positions**, **harmonize incident prevention, monitoring and management practices**, and build interoperable, secure information-sharing systems with capacity-building support.*

To strengthen collective bargaining power and ensure consistent incident prevention, monitoring and management across jurisdictions:

- **Governments** and legislators should coordinate positions before major forums, particularly through coalitions of like-minded states that include middle powers and Global South governments, to avoid fragmentation and advance shared interests.
- **International institutions** should:
  - **Provide implementation guidance** for universalizable frameworks such as the OECD's Common Reporting Framework for AI Incidents, while also establishing secure cross-border information-sharing channels for incidents that affect multiple jurisdictions.
  - **Coordinate with governments and standard setting bodies**, prioritizing interoperability rather than uniformity, enabling domestic systems to connect across borders without requiring identical regulatory structures.
  - **Support national and international standard-setting bodies** in collaborating with the International Network for Advanced AI Measurement, Evaluation and Science (formerly the International Network of AI Safety Institutes) on the development of globally interoperable incident reporting standards.
  - **Fund capacity-building initiatives** that support domestic and regional bodies in building incident management infrastructure and in enacting corresponding legislation.
- The **International Network for Advanced AI Measurement, Evaluation and Science** should (i) **collaborate with national and international standard-setting bodies** on the development of globally interoperable incident reporting standards, and (ii) **facilitate technical coordination among AISIs** to harmonize incident monitoring infrastructure and response frameworks like AI-IRS, establishing **common protocols for information sharing** to enable effective incident prevention, detection, containment, and recovery across jurisdictions while preserving each institute's operational independence.

HOW AI INCIDENTS RELATE TO AI RED LINES

Ultimately, the red line domains that surfaced during this year's Athens Roundtable discussions underscore growing global consensus that certain AI use cases, capabilities and behaviors are inherently incompatible with human safety and democratic values. As our Global Call for AI Red Lines campaign aimed to establish, we believe that red lines can serve as critical reference points for establishing a universally shared understanding of what constitutes unacceptable AI threats, and for building common boundaries among different stakeholders around non-negotiable societal protections. They can define the thresholds or conditions under which AI development or deployment becomes intolerable, on grounds that their potential for widespread harm surpasses what any government or organization can responsibly permit.

However, the **effective global governance of severe AI risks and hazards also requires the practical capacity to prevent, detect, and respond to these threats wherever they occur**. **AI incidents and accidents are therefore directly connected to these red lines**. They provide concrete evidence that certain risks have materialized in the real world, rather than being merely hypothetical. When governments observe harmful events or concerning near-misses, this can strengthen the credibility of red line categories, motivate action, and, crucially, inform the establishment of entirely new red lines for previously unconsidered harms that are identified through real-world experience. After all, since the general-purpose nature of and emergent capabilities in powerful AI systems make it impossible to exhaustively anticipate all of the societally unacceptable outcomes they may bring about in advance, incident patterns can reveal where new boundaries must be drawn. Incidents can also function as early warning indicators.

They can reveal when pressure is building towards a particular red line, signaling that unacceptable risks may already be unfolding and thereby creating pressure for governments to monitor, detect, and report such events and to participate in shared incident monitoring and response processes. **Severe AI incidents often transcend borders, so no country can manage them alone**. Cross border monitoring and response infrastructure is therefore essential for collective preparedness, effective prevention, and coordinated action, as it can enable red line conventions to be established, tracked, and enforced globally. This can also support the adoption of robust national and regional frameworks for general-purpose AI governance. If governments both support specific red lines and understand the incident patterns that precede them, they are more likely to put in place the institutional structures necessary to prevent the materialization and escalation of certain kinds of incidents. Such frameworks could thus help countries prevent and manage serious AI incidents and thereby contribute to the stability and security of the global environment.

Taken together, **shared red lines and incident management infrastructure can form a coherent and adaptive AI governance strategy**. **Red lines can set the thresholds** that must never be crossed, while **incident management infrastructure can provide the empirical evidence** needed to refine existing red line thresholds, identify new ones as they emerge, and enable coordinated interventions once thresholds are crossed. Accordingly, focusing on these two areas can enable governments to act early, together, and consistently in the face of severe AI risks.

To find out more about the key takeaways and recommendations that surfaced from the other closed-door dialogue session that took place at this year's Athens Roundtable – focused on **building a shared understanding of unacceptable AI risks and exploring viable diplomatic pathways for establishing measurable and enforceable red line thresholds** – click here.

ANNEX A - DEFINITIONS OF "AI INCIDENTS" ACCORDING TO EXISTING STATUTES AND FRAMEWORKS

Existing definitions of what constitute an "*AI incident*" vary.

1. The OECD/GPAI framework defines an "*AI incident*" as an event or series of events where the development, use, or malfunction of one or more AI systems directly or indirectly causes harm such as injury to health, disruption of critical infrastructure, violations of human rights, or damage to property or the environment. It also defines an "*AI hazard*" as a circumstance in which the development or use of a system could plausibly lead to such an incident, enabling a broad scope that supports upstream intervention and preventive monitoring.
2. The AI Incident Database (AIID) defines an incident as an "*alleged harm or near-harm event*" involving an AI system, explicitly including near misses and unverified claims to prioritise early warning and learning from weak signals.
3. The EU AI Act (Article 3(49)) defines a "*serious incident"* more narrowly as a malfunction or discrete event that causes death or serious health harm, serious and irreversible disruption of critical infrastructure, fundamental rights infringements, or serious damage to property or the environment.
4. California's SB-53 introduces "*critical safety incidents*", covering single episodes where a model causes or enables severe harms, including assisting in the creation of weapons of mass destruction, enabling autonomous criminal conduct without human oversight, or evading developer control.

THE FUTURE SOCIETY

Contact Us

GENERAL: info@thefuturesociety.org
THE ATHENS ROUNDTABLE: aiathens@thefuturesociety.org

www.aiathens.org
www.thefuturesociety.org