Designing escalation criteria for international AI incident response: criteria, triggers, and thresholds

Francesca Gomez, Josephine Schwab, Lydia Preston, Matthew Ball, Michael Harré
Arcadia Impact AI Governance Taskforce, in collaboration with Caio Machado (The Future Society)

April 2026

When should an AI incident trigger an international response?

For escalation to be effective in practice, it must occur early enough to enable the benefits of cross-border coordination, such as containment, mitigation, and shared response. We designed a blueprint for an international AI incident escalation framework, intended to serve as a common reference point across jurisdictions, enabling a more aligned approach to escalation while allowing flexibility in how actors respond within their own legal and policy contexts.

AI incident reporting requirements are also emerging in regulation. The EU AI Act defines "serious incidents" that trigger reporting obligations. California's SB 53 introduces "critical safety incidents" for frontier models. The General-Purpose AI Code of Practice operationalises reporting under the EU regime. Through a review of these existing frameworks we also address the question: **can the escalation criteria in these frameworks actually detect the incidents they are designed to catch?**

These are the questions we explore in our paper. We propose an eight-criterion escalation framework for AI incidents, grounded in a review of emerging regulatory regimes and stress-tested against ten real and structured incidents. We found that for several important categories of risk, the framework does not always trigger escalation: not because the criteria are poorly drafted, but because structural design patterns make certain incident types invisible to the escalation architecture.

The full paper is available here

Key takeaways

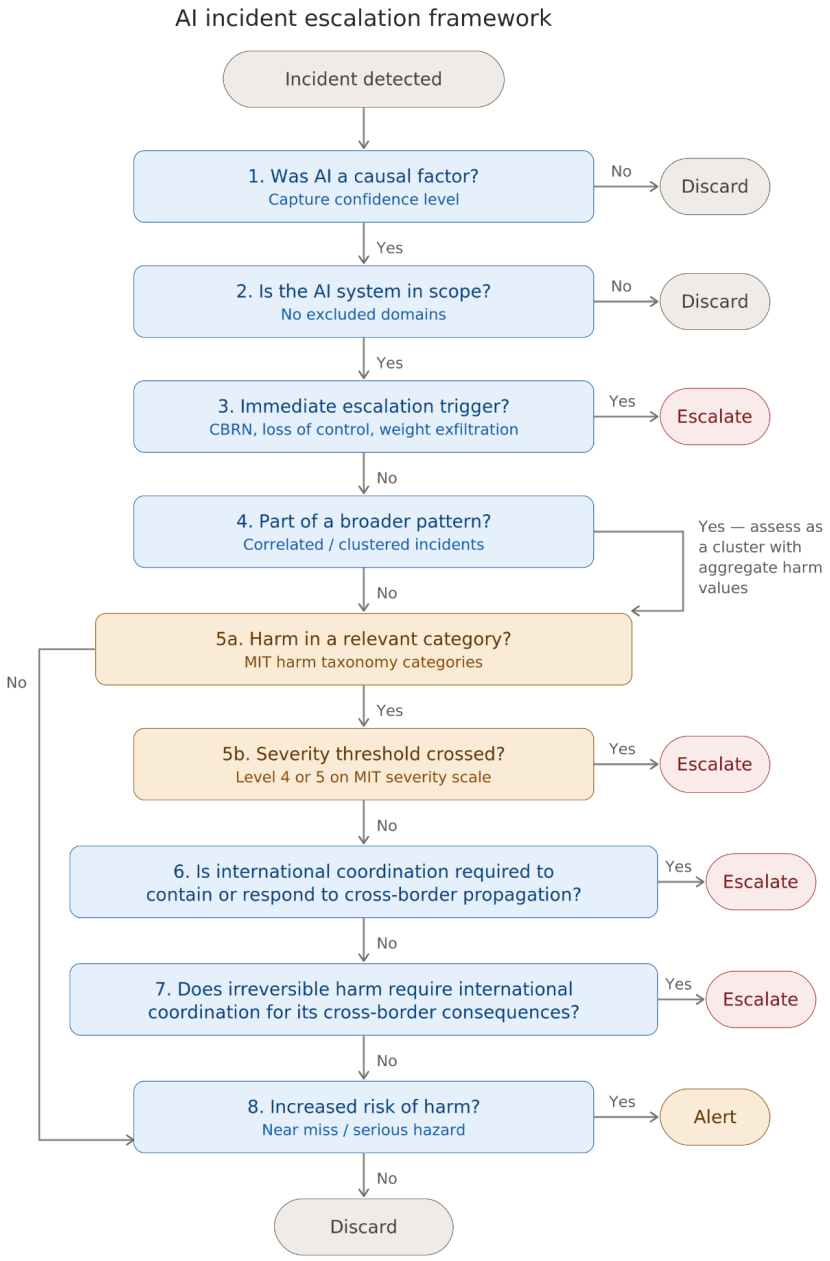
- AI incident reporting requirements are beginning to emerge in regulation and policy, but no operational criteria exist for determining when a detected AI incident should be escalated beyond national handling to international coordination.
- Our paper proposes an AI incident escalation framework intended to serve as a common reference point across jurisdictions, enabling a more aligned approach to escalation while allowing flexibility in how actors respond within their own legal and policy contexts.

- We review EU AI Act, California's SB 53, the GPAI Code of Practice, and explore analogous design elements from existing incident-response frameworks in health, finance, cyber, nuclear safety, aviation, and disaster risk reduction. From this, we derive eight proposed criteria for international escalation and translated them into an illustrative, sequential flowchart with gated decision points and threshold checks. We then stress-test our framework against ten documented AI incidents and structured variants to identify where escalation criteria under-detect or misclassify incidents in practice.

- We find that effective escalation depends not only on the design of triggers and thresholds themselves, but also on two upstream dependencies: the definitions of harm and clustering methods against which thresholds are set, and the data infrastructure through which triggers are evaluated. Where these dependencies are underdeveloped, as they are for large-scale manipulation, psychological harm, and loss-of-control risks, even well-designed triggers cannot function as intended.

Our AI incident escalation framework

Our framework is designed to be walked through as a series of questions, starting at the entry point of ‘Incident detection’ in the flowchart. Each question in the flowchart corresponds to an assessment criterion, for which a detailed explanation and set of questions is provided. The responses determine the pathway through the flowchart, which ends in a classification decision: escalate, discard or alert, which corresponds to when international cooperation is required to investigate or respond effectively, even where the incident itself is geographically local.



Criterion	Name	Purpose
1	Was AI a causal factor?	Establishes whether an AI system was involved in causing the incident and captures the confidence level in AI's causal contribution (high, medium, or low) before any further assessment proceeds.
2	Is the AI system in scope?	Determines whether any domain-based exclusions (e.g. military, defence, national security) apply. The paper proposes no excluded domains, to maintain visibility of all incidents with severe harm potential.

3	Has an immediate escalation condition been met?	Identifies a narrow set of incident types (CBRN weapon assistance, model weight exfiltration, and loss of developer control) where escalation should be triggered immediately, regardless of whether confirmed harm has yet materialised.
4	Is the incident part of a broader pattern?	Detects whether the incident shares a root cause (technical, capability, or contextual) with prior incidents or near misses, so that individually sub-threshold incidents can be assessed collectively as a composite cluster.
5a	Has harm occurred in a relevant category?	Determines whether the harm falls within a recognised category using the MIT harm taxonomy (physical harm, financial loss, environmental damage, privacy, democratic norms, etc.), ensuring novel or non-traditional harm types are not excluded.
5b	Has harm crossed a relevant severity threshold?	Assesses whether harm in any single category has reached Level 4 (Severe) or Level 5 (Catastrophic) on the MIT severity scale, providing a graduated threshold between the EU AI Act's qualitative "serious" standard and SB 53's high quantitative floors.
6	Is international coordination required to contain or respond to cross-border propagation?	Determines whether the incident's spread dynamics, via supply-chain, capability, or emergent propagation pathways, create an operational need for international coordination to contain the incident or warn exposed jurisdictions.
7	Does irreversible harm require international coordination for its cross-border consequences?	Captures contained incidents where the permanent, non-recoverable portion of the harm has downstream effects on other jurisdictions by virtue of interconnection, dependency, or precedent, even though the incident itself is not actively spreading.
8	Has a near miss or hazard indicated inadequately mitigated harm?	Captures events where harm was closely averted but the revealed failure mode may affect AI systems deployed by other developers or jurisdictions, warranting an alert even though no harm materialised.

Testing our framework

We stress tested our framework against ten real AI incidents, including the Pravda disinformation ecosystem, the Grok deepfake incident, and a CBRN near-miss. These were selected to test across key domains of systemic AI risk, namely, large-scale manipulation and psychological harm; loss of control; chemical, biological, radiological and nuclear (CBRN) threats; and cyber vulnerabilities and offensive capabilities.

This was a useful activity which we would recommend others developing AI incident escalation frameworks to use to surface blind spots and boundary-case ambiguities that only become visible when abstract criteria meet the messiness of real-world incidents.

We found:

- **Certain systemic risks warrant escalation before harm is confirmed, but current regulatory triggers largely require confirmed harm outcomes**. For events such as the provision of expert-level CBRN uplift, the exfiltration of frontier model weights where containment is unconfirmed, or an AI system evading developer control, the intervention window is narrow, the probability of severe irreversible harm is high, and no single actor has visibility over both the risk pathway and the downstream harm.

- **Existing regulatory frameworks may systematically under-classify incidents that present as patterns rather than discrete events**. Both the EU AI Act and SB 53 define incidents in terms of single events; neither has a mechanism for treating a cluster of

individually sub-threshold incidents as a composite warranting escalation. We propose three levels at which AI incident clusters can be identified: technical, capability, and contextual root causes assessed within 30-, 60-, and 90-day rolling windows respectively.

- **Incident escalation frameworks are designed for acute events, but some systemic risks manifest as ongoing, low-severity, high-frequency harms for which no existing framework defines what a reportable incident looks like**. Psychological harm from human-AI interaction, cumulative manipulation through generated content, and gradual information integrity erosion through training data contamination do not follow a discrete-event structure.

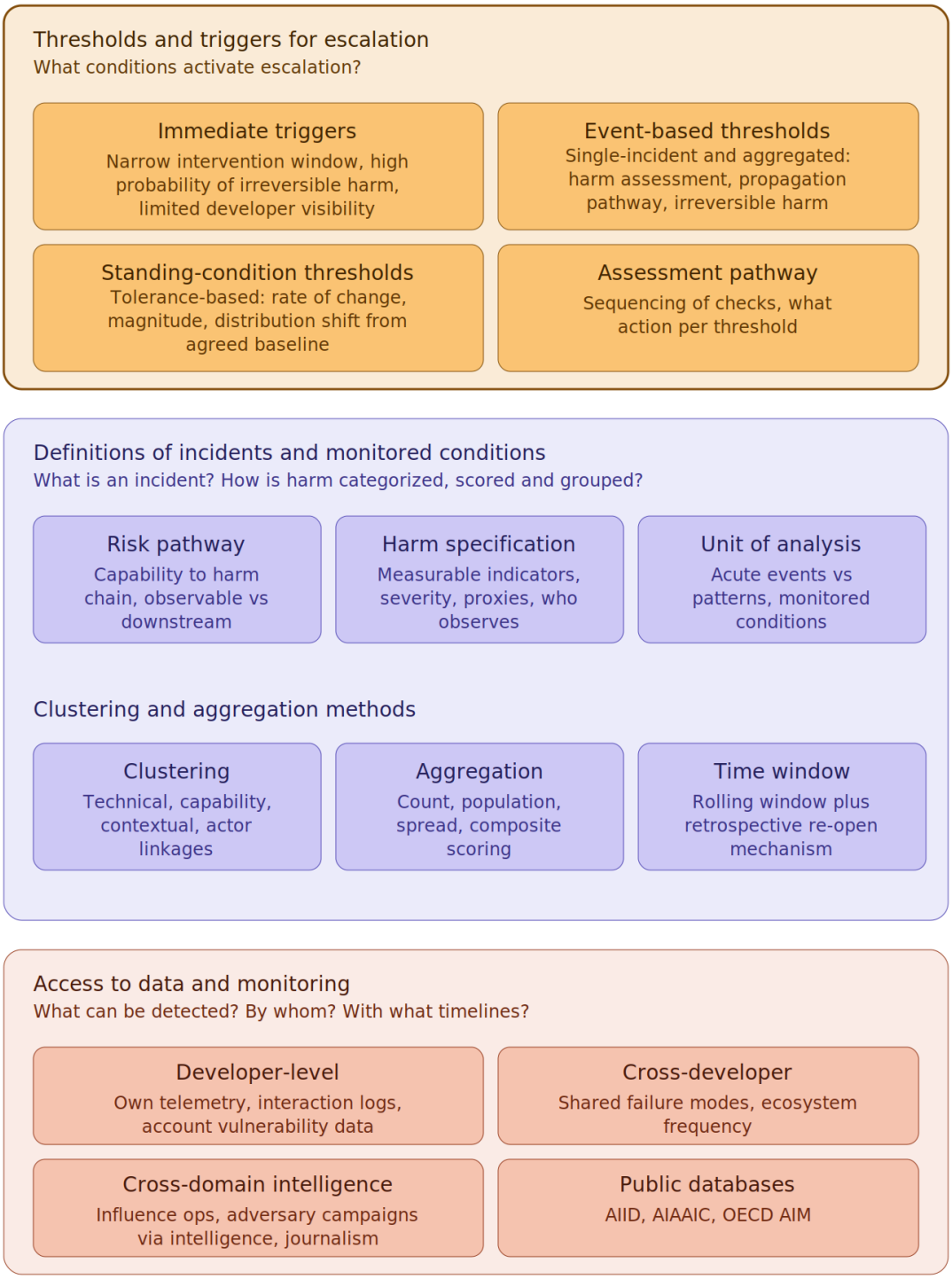
Thresholds don't work without definitions and data

The four patterns above describe problems in the escalation layer. But the stress testing revealed something more fundamental: escalation criteria sit atop a three-layer dependency chain: triggers depend on definitions, definitions depend on data, and gaps at any layer propagate upward.

The definitions gap. Severity scales are calibrated for tangible outcomes. The MIT AI harm severity scale's Level 4 descriptors work for mass casualties or infrastructure collapse, but the principal harms in the manipulation and psychological harm domain, emotional dependence, dignity violation, distorted perceptions of reality, do not map onto existing descriptors. Both incidents we tested in this domain scored Level 3 across multiple categories simultaneously, but the per-category design provides no mechanism for cross-category aggregation. The severity criterion is essentially non-functional for this risk domain.

A related gap concerns vulnerability. Non-consensual deepfakes of children are legally distinct from those of adults; a chatbot reinforcing delusional thinking in someone with pre-existing psychosis presents qualitatively different risk. Yet severity scales assess harm without adjusting for who is affected, even though developers already operationalise the distinction through age prediction and mental health risk flags.

The data gap. The stress testing revealed two distinct problems. Cross-developer data: pattern-based triggers require visibility across providers, but no sharing arrangements exist. Cross-domain intelligence: the GRU-linked disinformation cluster was identified through national intelligence analysis and investigative journalism, not AI developer telemetry. Even perfect cross-developer visibility would not have detected it. Detecting state-orchestrated campaigns requires pre-specified interfaces with institutions whose mandates already cover adversarial campaign detection.



Implications

The design of escalation criteria, what is assessed, in what order, against what thresholds, using what data, directly determines which incidents governance frameworks can see and which they cannot.

Our proposed framework and identified design patterns are intended to support policymakers and other stakeholders worldwide who are developing or implementing AI incident reporting regimes. This work aims to show how the design and operationalisation of escalation criteria shape which incidents are identified and escalated in practice, particularly under real-world constraints and across common risk pathways. In domains where risk pathways are less well understood, such as manipulation, psychological harm, and loss of control, significant definitional gaps remain – building ongoing monitoring without clarity at the definitions layer risks creating divergence across developers that compounds over time.

Our work highlights that the escalation layer, where most regulatory attention has focused, sits in a dependency chain. To function effectively, it will rely on definitional work that specifies what constitutes an incident and how harm is categorised, as well as data collection and sharing infrastructure that determines what can actually be detected.