Phase Modeling for AI Incident Emergence: Adapting Epidemiological Methods to Post-Deployment Governance

By Sophia Abraham, Taiye Chen, Cyril Chhun, Giovanna Jaramillo-Gutierrez, Simon Mylius, Sayash Raaj

Full paper: forthcoming

Abstract

Artificial intelligence systems are now deployed at scale across sectors, accompanied by a growing number of real-world incidents ranging from misinformation and cybercrime to autonomous-system failures. Public repositories of AI incidents provide visibility into these events, but their data cannot be directly interpreted as measures of "risk" without integration of contextual information, such as the prevalence of a risk-associated system in the world. As a result, policymakers, companies, and the general public lack a means of prioritizing responses to AI risk. Inspired by public-health processes, we identify six phases of incident emergence from "No Evidenced Occurrence" to "Rare Mitigated," and outline criteria for distinguishing them, aided by reporting-delay correction, exposure normalization, changepoint detection, and latent-state models. These statistical tools aid phase transition identification even when reporting data is incomplete or distorted. We demonstrate the framework through a detailed case study of autonomous vehicles, showing phase-based interpretation yields a more stable and governance-relevant picture of risk development. We also introduce a meta-reporting card prototype designed to translate phases into clear and actionable policy recommendations for regulators.

Executive Summary

Artificial intelligence (AI) systems are now widely deployed across consumer platforms, public services, and physical infrastructure. As deployment expands, reports of AI-related harms, from deepfake-enabled fraud to autonomous-vehicle collisions, have grown in frequency and prominence. Yet today's incident repositories, including the AI Incident Database (AIID) and the OECD AI Incidents Monitor, cannot be interpreted at face value as indicators of underlying risk. Their counts are shaped by reporting delays, inconsistent observability, media-driven attention cycles, and the absence of exposure denominators. As a result, policymakers and regulators lack a coherent way to assess whether a given harm type is emerging, accelerating, stabilizing, or being brought under control.

This paper adapts epidemiological surveillance concepts to the problem of AI incident interpretation. We develop a phase-based framework that treats incident reports not as direct risk measurements, but as noisy signals from which latent risk trajectories can be inferred. The framework defines six qualitative phases in the lifecycle of a type of AI harm: No Evidenced Occurrence, Rare Occurrence, Rapid Expansion, Endemic Unmitigated, Endemic Mitigated, and Rare Mitigated, and links each phase to corresponding governance levers. We operationalize this framework by combining several statistical methods that correct for or are robust to reporting biases.

We demonstrate the framework through a detailed case study of an incident type that has mandatory government reporting in some regulatory contexts: autonomous vehicles. The analysis reveals that, after adjusting for exposure and reporting distortions, autonomous-vehicle incidents exhibit oscillatory crisis–recovery dynamics rather than sustained acceleration. Periods of elevated risk are closely associated with deployment expansions, system failures, and governance shocks, and are often followed by regression toward lower-risk phases. An exploratory analysis of deepfake-enabled harms, presented in the appendix, illustrates how the same framework can capture more persistent, monotonic transitions in domains shaped by rapid capability diffusion and media amplification.

Across domains, raw incident counts substantially misrepresent underlying risk: delay-adjustment, exposure denominators, and attention effects materially alter inferred trends. By contrast, phase classification provides an interpretable summary of both risk level and trajectory, enabling clearer differentiation between emerging, stabilizing, and effectively mitigated harms.

The phase model offers a governance-relevant lens for interpreting these dynamics. By assigning probabilistic phase labels to each incident type and updating them as new reports arrive, the framework provides a living map of where different types of harm sit in their lifecycle. This enables governments, regulators, developers, civil-society organizations, and insurers to focus attention where it is most warranted, for example, escalating post-market surveillance during rapid expansion, or evaluating mitigation durability during endemic phases.

Finally, we introduce a prototype meta-reporting card that translates phase assessments into practical signals for decision-makers. Rather than relying on raw incident counts, the card summarizes phase status, uncertainty, historical transitions, and governance-relevant triggers, supporting proportionate and timely intervention.

In sum, this work demonstrates how phase-based reasoning, long used in public health, can be adapted to AI governance. By situating incident patterns within a lifecycle framework, policymakers gain a more interpretable, calibrated, and actionable view of how AI harms emerge and evolve, even under imperfect and incomplete surveillance.