The pragmatic case for AI incident monitoring

(or "the egg exists")

Every day, AI systems make decisions that affect billions of people: conversing, driving cars, screening job applications, determining credit scores, diagnosing illness and curating social media feeds. Developers cannot catch every flaw or foresee every real-world application. When things go wrong, AI systems can cause harm.

Other high-reliability industries (aviation, nuclear energy) and population-scale technologies (pharmaceuticals, medical devices) use incident monitoring to learn from failures and improve safety. AI lacks the equivalent institutional infrastructure. EU and US (California) reporting requirements take effect in 2026, but are narrow in scope, excluding most incidents (1, 5).

Certain domains have dedicated monitoring (e.g. the US Department of Transportation maintains a database of autonomous vehicle crashes (3)) and public databases like the AI Incident Database (2) and the OECD's AI Incidents and Hazards Monitor (4) attempt to fill the gap by cataloguing thousands of news reports of AI incidents. However, public sources collectively capture only a fraction of all AI harm.

Furthermore, AI incidents are complex, and data from any source is hard to interpret: a rise in reporting might mean AI is getting more dangerous, or that it's being used by more people, or simply that data collection is improving or journalists are paying more attention. These three things can look identical in the data.

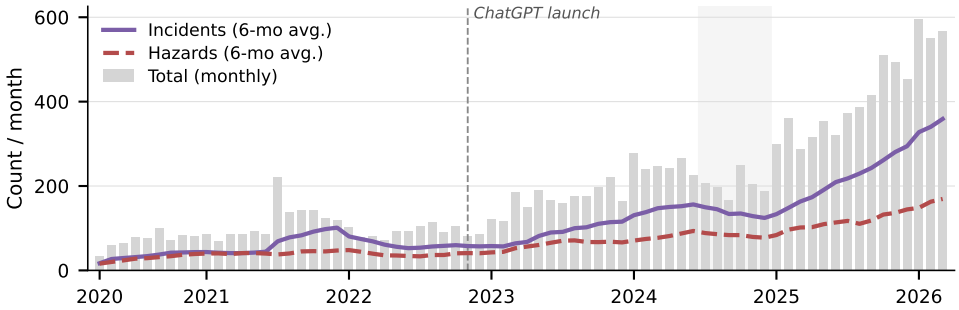

Figure 1: Monthly AI incidents and hazards recorded in the OECD AI Incidents and Hazards Monitor (AIM) (4). Rising raw counts conflate media attention, deployment growth, and the frequency of harm per use.

Fortunately, this is a familiar problem in public health. Epidemiologists working from imperfect data must know how many people are exposed to a pathogen and how dangerous it is per exposure to decide on the best response. Applying the same principles to AI incidents, we developed a methodological framework in three parts:

Definition: a "structured monitoring question" defines the scope of our analysis (what harm are we assessing, to whom, over what time period?).

Estimation: using all available evidence we estimate how many people are exposed to the relevant type of AI system and how frequently harm occurs. By comparing estimates over consecutive time periods, we can cancel out stable reporting biases and estimation uncertainty (tricky data problems), leaving us with crude but comparable trends for harm, exposure, and harm-per-exposure.

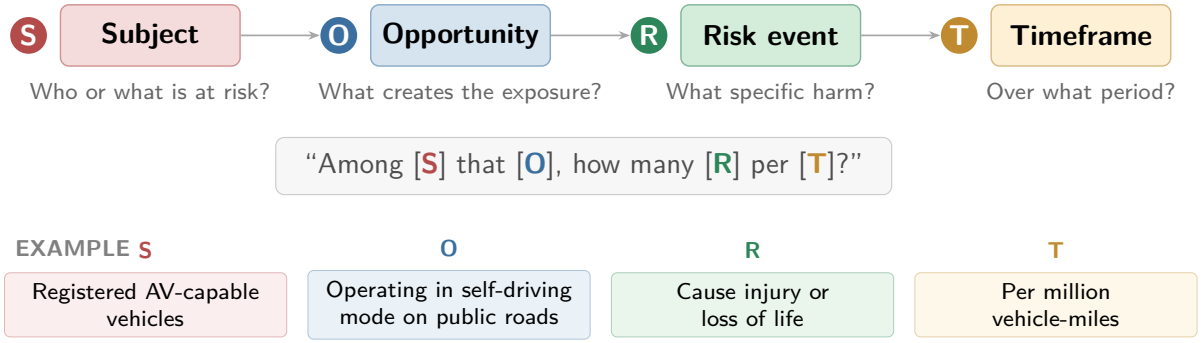

Figure 2: The SORT structure of a monitoring question: **S**ubject, **O**pportunity, **R**isk event, **T**imeframe.

Classification: the combinations of trend directions ("increasing" or "decreasing") for exposure and harm-per-exposure are mapped to four classifications: escalating, mitigating, concentrating, receding. (Where trends cannot be estimated the result is unclassifiable.)

Mitigating *Monitor closely* $\hat{H}\downarrow\rightarrow$ E$\uparrow$	**Escalating** *Urgent attention* $\hat{H}\uparrow$ E$\uparrow\rightarrow$
Receding *Continue strategy* $\hat{H}\downarrow$ E$\downarrow\rightarrow$	**Concentrating** *Targeted measures* $\hat{H}\uparrow\rightarrow$ E$\downarrow$

Figure 3: The classification grid. Trend directions for harm-per-exposure ($\hat{H}$) and exposure (E) map to four trajectory categories, each implying a different governance response.

Classification does not quantify absolute harm or prescribe specific responses, but it can inform governance by revealing what is driving change and whether current efforts are working.

We applied our methods to a range of incident types and found it was possible to generate insight despite real-world data limitations. For example, our case study on conversational AI-related self-harm illustrates the importance of accounting for exposure. Incidents of harm are estimated to increase significantly between T1 and T2, such that a "naive" interpretation might conclude that conversational AI systems are becoming riskier. However, deployment and use is estimated to increase significantly faster, suggesting that, although absolute harm is increasing (and is hence still a problem), conversational AI systems are becoming safer per use.

We also found, repeatedly, that we lacked sufficient data to say anything definitive. This brings us back to our original problem: most AI harm goes unreported. Better monitoring requires better data, but investment in better data requires demonstrating that monitoring is worthwhile. The chicken and the egg. Our project is offered as modest proof that the egg exists: that meaningful analysis is possible even with the imperfect data we have today.

Zooming out, incident monitoring is just one piece of the AI governance puzzle. It deserves more attention, but it has limitations. It works best for discrete, recurring, visible harms, but has little to offer diffuse systemic harms and cannot predict or prevent risks — including catastrophic ones — that have yet to materialise.

References

[1] EU. Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence

(artificial intelligence act), article 73. Official Journal of the European Union, L 2024/1689, 2024. URL https://artificialintelligenceact.eu/article/73/.

[2] Sean McGregor. Preventing repeated real world AI failures by cataloging incidents: The AI incident database. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 15458–15463, 2021. doi: 10.1609/aaai.v35i17.17817.

[3] National Highway Traffic Safety Administration. Standing general order on crash reporting for ADS and level 2 ADAS, apr 2023. URL https://www.nhtsa.gov/laws-regulations/standing-general-order-crash-reporting. Amended April 2023. Accessed: 2026-05-15.

[4] OECD. Overview and methodology of the AI incidents and hazards monitor. OECD.AI Policy Observatory, 2024. URL https://oecd.ai/en/incidents-methodology. Accessed: 2026-04-15.

[5] Scott Wiener. Senate bill 53: Transparency in frontier artificial intelligence act (TFAIA). California Legislature, 2025–2026 Regular Session. Signed September 29, 2025, 2025. URL https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202520260SB53.