A Pragmatic Classification Framework for AI Incident Monitoring

Anonymous Authors[1]

Abstract

Public AI incident database counts conflate changes in reporting propensity, deployment growth, and shifts in harm frequency per unit of exposure. These issues introduce significant uncertainties, challenging public and corporate policy frameworks centered on realized risks. We propose a simple framework that establishes clear points of inquiry, separately estimates exposure from harm-rate trends, and then classifies into meaningful trajectory categories for governance decisions. The framework combines a structured monitoring question format (SORT) to clarify coverage decisions, a tiered estimation procedure calibrated to available evidence, and LLM-assisted incident matching against public databases. Applied to various monitoring questions, we draw conclusions regarding the monitoring ecosystem more broadly: providing an essential interpretative classification, determining what can and cannot be claimed, and establishing that exposure estimation is required as AI deployments become increasingly common.

1. Introduction

Artificial intelligence (AI) systems are prone to unforeseen failures, security breaches, adverse events, and malicious use (Jeanmaire & Boger, 2025). Public databases now catalog thousands of incidents (OECD, 2024b; McGregor, 2021), providing a public record of harms involving AI systems (OECD, 2024a). Translating these databases into actionable insight, however, requires bridging the gap between what is recorded and what is actually happening (Paeth & McGregor, 2025).

Populated primarily from news reports, public incident databases systematically over-represent acute, dramatic, and Anglophone harms while under-capturing the chronic, diffuse, and systemic harms (Richards et al., 2025; Nixon, 2011; Teo, 2025) (see Appendix A). More fundamentally, simple "raw" incident counts conflate at least three distinct factors: the propensity for incidents to be observed and reported; the scale of system deployment and use ("exposure"); and the frequency of harm per unit exposure (Paeth et al., 2025; Richards et al., 2025). They also conceal the nature and severity of harm in each incident.

	Problem	Framework response
P1	Complexity and lack of standardization confounds analytical consistency	Structured monitoring question (SORT; §2)
P2	More harmful AI systems indistinguishable from more deployed AI systems	Separate exposure and harm estimates (§2.1)
P3	Sparse or missing data treated as absence of harm	Evidence tiers; proxy measures; confidence statements; principled abstention (§2.2)
P4	Absolute values sensitive to reporting bias and uncertainty	Trend estimates robust to stable biases and uncertainties (§4.3)
P5	Point estimates imply undue precision	Order-of-magnitude estimates resolve to simple directional classification (§2.2)

Table 1. **Five problems with raw incident counts,** and corresponding framework responses.

Disaggregating these factors and accounting for, or acknowledging, the inherent biases and limitations, can help decision makers understand and distinguish genuine risk escalation from increased media attention (Paeth & McGregor, 2025). The practical utility extends across sectors: in structured interviews with AI risk and safety practitioners, we found a consistent demand for clarity on which harms are increasing in frequency or severity, and which are relevant to specific operational contexts. Table 1 summarizes five problems with raw incident counts and the corresponding responses proposed in this paper.

Incident monitoring has driven safety improvements in other high-reliability, safety-critical industries (aviation, nuclear power) and population-scale technologies (pharmacovigilance, healthcare). AI stands to benefit similarly (Perrow, 2011) but lacks the necessary institutional infrastructure and interpretive framework (Wei & Heim, 2025).

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

Existing mandatory incident reporting frameworks for AI remain narrow in scope and fragmented across jurisdictions. The EU AI Act (Articles 73 and 55) requires providers and deployers of high-risk AI systems, and providers (but not deployers) of systemic-risk general-purpose AI models, to report serious incidents to national authorities and the EU AI Office respectively. However, reports flow into different parts of the EU system with no mandate to consolidate or publicly share them, and the definition of "serious incident" excludes many harms that any common-sense reading would consider serious. California's Transparency in Frontier Artificial Intelligence Act, TFAIA (SB-53), and New York's Responsible AI Safety and Education (RAISE) Act (from 2027) are narrower still, applying only to a handful of frontier AI developers, and capturing only catastrophic-scale harms, security breaches, and loss-of-control events, with no public disclosure requirement. No other jurisdiction currently mandates AI-specific incident reporting or encourages voluntary reporting, leaving the vast majority of global AI deployments outside any reporting framework.

Despite the OECD developing standardized incident definitions (OECD, 2024b; Wei & Heim, 2025), there is currently no agreement on how AI incident data should be analyzed or compared across time, and no validated methodologies for post-deployment AI monitoring (Rao et al., 2026; Whittlestone & Clark, 2021). This presents an opportunity to establish shared approaches before national regimes fragment and solidify. Despite their limitations, news-sourced public databases are likely to remain the primary records of AI harm for the foreseeable future. Methods to extract a reliable signal from them are urgently needed (Paeth & McGregor, 2025)—both for immediate governance application and to lay the foundations for, and encourage the adoption of, AI incident monitoring more broadly.

In aviation safety reporting, every incident is investigated, causal factors are identified, and the findings are fed back into the system to inform design choices and modifications (McGregor, 2021). The conditions for such analysis (a known fleet size, mandatory and incentivized voluntary reporting, and identifiable operators) do not exist for AI. We often do not know how many people use a given AI system, how many automated decisions are made, or at what scale generative output is produced (Stein et al., 2024; Tanjaya & Pratt, 2025). Epidemiological surveillance was built for this kind of uncertainty. Its operational methods can accommodate incomplete reporting, unknown denominators, and media-driven distortions of apparent incidence (McCarty et al., 1993). Public health principles, such as standardized case definitions, the separation of exposure from harm rate, and tiered approaches to estimation that accept coarse data (Dolin et al., 2025) can be applied meaningfully to AI incident monitoring. A parallel effort applies this lens through a lifecycle-phase model that classifies the latent state of a harm (e.g., rare, expanding, endemic) rather than its directional trend (Abraham et al., 2026).

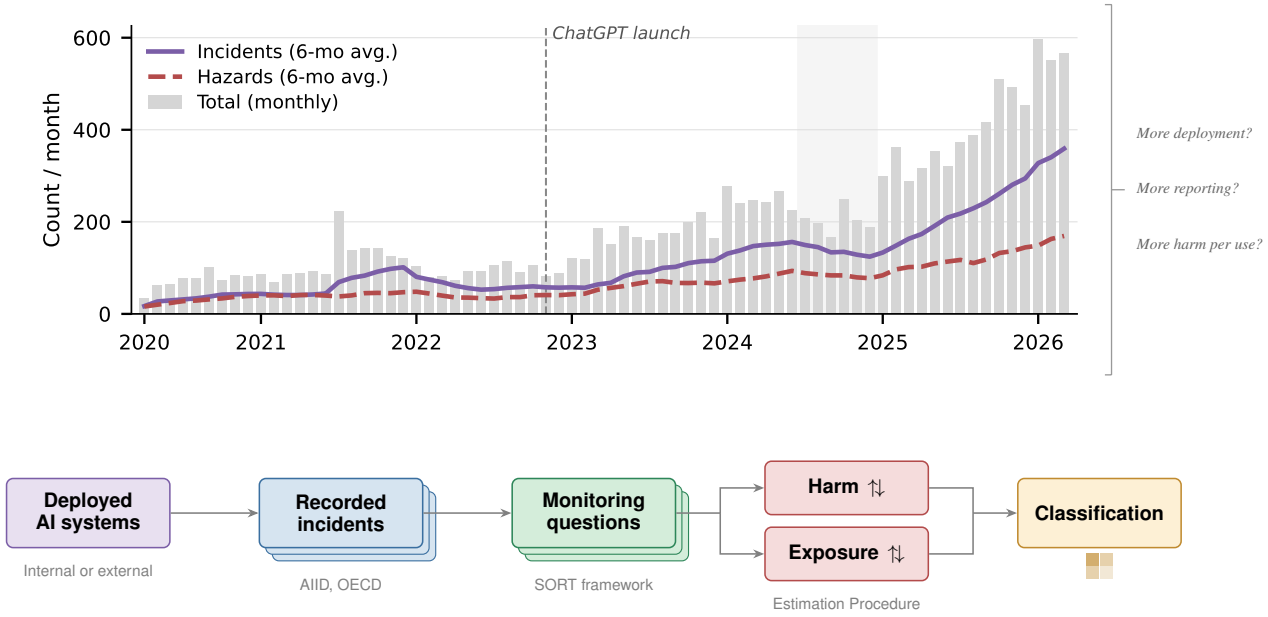

Figure 1. **The interpretive pipeline for AI incidents.** *At top:* rising counts of monthly AI incidents and hazards (defined here as "potential dangers") reported in the OECD AI Incidents and Hazards Monitor (AIM) conflate trends in media attention, deployment growth, and the frequency of harm per use of AI systems (OECD, 2024a). *Below:* our framework takes recorded incidents as inputs to structured monitoring questions, then estimates harm and exposure trends separately to classify each monitoring question into a trajectory category.

Applying a public health lens, we propose a methodology for monitoring AI harms from incident data with the goal of informing governance prioritization. Section 2 introduces the framework: a structured approach to defining monitoring questions, separate estimation procedures for harm and exposure trends, and a four-way trajectory classification. Section 3 applies the framework to two monitoring questions on autonomous vehicles and AI chatbot safety. Section 4 discusses the classification's explanatory value, its limitations, and implications for reporting standards.

2. Framework

The problems in Table 1 arise due to the conflation of three steps: question definition, estimation, and interpretation. Our procedure takes each in turn: first, a precise monitoring question (MQ) is defined; second, trends in exposure and harm are estimated and confidence is calibrated to available evidence (§2.1); and third, the resulting pair of trends are combined and classified into one of four trajectory categories (§2.2). Each step is robust across incident types and evidence constraints and explicit about uncertainty.

Why ask a monitoring question? A precisely defined MQ helps ensure comparability, reproducibility, and internal consistency in the estimation and classification steps.

Structuring the monitoring question. Our SORT framework structures AI incident monitoring questions around four components: Subject (who or what is at risk), Opportunity (those actually exposed to the harm mechanism), Risk event (the countable harm), and Timeframe (the observation period), producing questions of the form "Among [S] that [O], how many [R] per [T]?" (See figure 2 for a worked example.) SORT is analogous to the PICO framework (Patient/Problem, Intervention, Comparison, and Outcome) in evidence-based medicine (Richardson et al., 1995), and is similarly flexible: [S] and [O] can refer to people, systems, content, conversations or deployments, and can focus on the causes of harm, or victims of harm, allowing the same underlying harm to generate multiple valid MQs depending on who needs the answer, while forcing analytical choices to be explicit rather than buried in unstated assumptions. [T] can be set to accommodate different harm dynamics and data availability. An interactive SORT tool is available to guide users in developing a precise MQ [1].

2.1. Estimation procedures

Answering an MQ requires estimating two variables across consecutive time periods: the total harm associated with [R], and the exposure defined by [S] and [O]. We categorize methods for estimations into four tiers (Table 2) according to the strength of the available evidence and the corresponding confidence we can have in the results.

Table 2. Evidence tiers for harm and exposure estimation.

Tier	Method	Sensitive to	Confidence
1	Direct retrieval	Authoritative Source	High
2	Combine proxies/records	Proxy construction	Medium
3	Expert elicitation	Panel selection	Low
4	Abstain	—	—

At Tier 1, the estimate is read directly from an authoritative source, such as vehicle crash filings, pharmacovigilance registries, or platform transparency reports. At Tier 2, no single source provides complete data, but bounds can be constructed from available partial data sources. Harm estimates have a natural lower bound in the relevant incidents recorded in public databases—the true harm cannot be lower than what has been captured. The upper bound is derived from one or more proxy measures. Exposure has no natural lower bound: both bounds must be constructed from proxies, making lower-confidence estimates more common. At Tier 3, no quantitative sources support even a rough estimate, and domain experts are asked to bound the plausible range. Where the plausible range spans more than two orders of magnitude, or no expert consensus can be reached, the esti-

[1] Available as a Claude artifact (see GitHub).

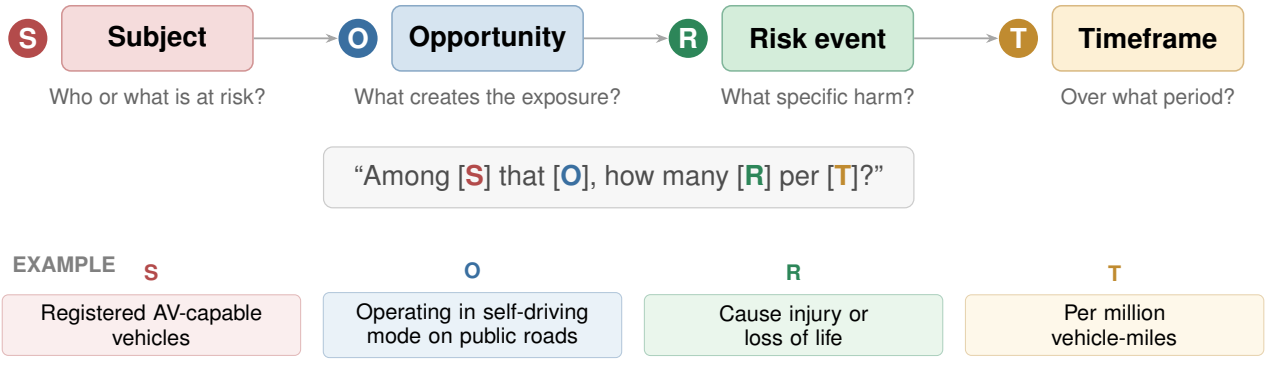

Figure 2. **The SORT framework.** Structured monitoring questions reduce ambiguity in analysis due to ill-defined incident types.

mate is assigned to Tier 4, principled abstention. This is a valid finding in its own right: that current evidence cannot support even an order-of-magnitude estimate.

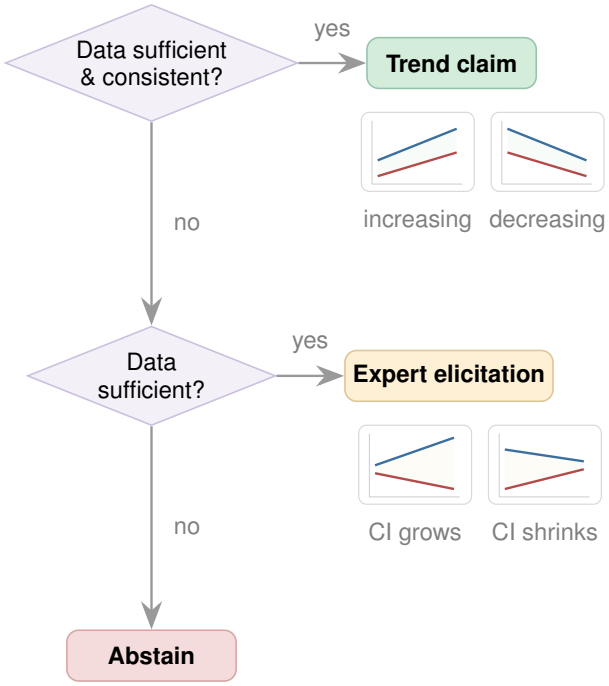

Figure 3. **Estimation procedure.** When lower and upper bounds move in the same direction, a trend claim follows directly. When they diverge, the plausible range either grows (increasing uncertainty) or shrinks (converging bounds but ambiguous direction) and expert elicitation is required. Estimations with only two upper (or two lower) are possible, but with a lower confidence (Tier 3).

Estimating Harm, H. Authoritative single sources rarely exist for AI harms. At Tier 2, news-sourced public incident databases serve as the default for estimating lower bounds on *total harm* associated with [R]. In this paper we draw on two public AI incident databases — the AI Incident Database (AIID) \cite{aiid_website} and the OECD AI Incidents and Hazards Monitor (AIM) \cite{oecd_aim_website} — which share the same basic structure (aggregating news reports on discrete incidents) and broad scope, but differ in editorial approach and sourcing methodology (see Appendix A for more detail). Much of the database search task can be automated. We developed an LLM-powered script that takes an MQ and a database of incident reports as inputs and filters for full and partial matches (see Appendix C.1). A partial match is an incident that matches some but not all SORT components. For example, an MQ specifying [S] as "teenagers" and [R] as "LLM-assisted completed suicides" would yield few full matches but many partial matches: adult cases that share the harm mechanism but fall short on [S], and cases of AI-assisted suicide attempts or planning falling short on [R]. A high ratio of partial to full matches suggests the MQ may be overspecified and can be revised accordingly.

An upper bound on the *total harm* can be estimated using relevant proxy measures, for example by adapting an authoritative source on harm from a broader class containing [S]. In Figure 2, total motor vehicle harm places an upper bound on autonomous vehicle harm, which can be further tightened by comparing statistics from years prior to the introduction of autonomous vehicles. The aim is always the narrowest defensible range.

Estimating Exposure, E We define exposure as *the opportunity for harm to occur*. Exposure helps distinguish between a system becoming more dangerous and more widely deployed. Deployment and use data does not exist or is not publicly available for most AI harm types: we rarely know how many people interact with a particular system, how many decisions are automated, or how many conversations take place. As such, exposure estimation typically relies on Tier 2 methods that combine multiple partial sources.

For example, combining population age data from a census with scam exposure rates from surveys would permit an estimate of the population exposed to AI voice-clone fraud. For an exposure estimate, we provide a best-guess point estimate with a plausible range, and an order-of-magnitude estimate. The plausible range represents the upper and lower limits outside of which the value is unlikely to fall.

Trend Estimation Repeating estimation across consecutive time periods as defined by [T] allows a trend to be established for each variable. Tracking how upper and lower bounds each change across consecutive time periods yields five possible outcomes: (1) both increase, (2) both decrease, (3) they diverge, (4) one is estimable and the other is not, or (5) neither is estimable. Where both bounds move in the same direction, a trend claim (*increasing* or *decreasing*) follows directly. Where they diverge, the trend is ambiguous and expert elicitation is required to resolve it. Where one bound is estimable and the other is not, a trend claim may still be made but confidence is reduced from Medium to Low, and supplementary expert elicitation is recommended. Where neither bound is estimable, or where expert elicitation fails to converge, no trend claim can be made.

2.2. Classification

To classify an MQ, we first derive the harm-per-exposure trend $\hat{H}$ by comparing the rates of change of the harm H trend and exposure E trend from our estimation procedures. If H grows faster than E, $\hat{H}$ is *increasing*, and vice versa. At order-of-magnitude levels of accuracy, this requires only the determination of which trend is dominant, not precise division. MQs for which E or H cannot be determined or for which the trends are not reliably distinguishable are

unclassifiable.

The simplest non-trivial classification is a 2×2-grid taking raw exposure trend E and harm-per-exposure trend $\hat{H}$ as inputs (see Figure 4). This highlights rather than masks our epistemic limitations. The result is a mapping to four trajectories with distinct governance implications as follows:

Escalating $[\hat{H}\uparrow, E\uparrow\rightarrow]$**.**Both the population at risk and the harm per unit exposure are growing. This demands an urgent response: expanded monitoring, active investigation into causal drivers, and possibly regulatory intervention.

Mitigating $[\hat{H}\downarrow, E\uparrow\rightarrow]$**.** More people are exposed, but harm per unit exposure is decreasing, suggesting existing safeguards are working. Continued monitoring is warranted: a failure of current controls could shift the trajectory to escalating.

Concentrating $[\hat{H}\uparrow, E\downarrow\rightarrow]$**.** Fewer people are exposed, but those who are face worse outcomes. This calls for targeted protective measures and investigation into why harm is intensifying.

Receding $[\hat{H}\downarrow\rightarrow, E\downarrow]$**.** Neither dimension is worsening. Additional intervention may not be required. Where specific measures preceded this trajectory, maintaining or extending them to related domains may be worthwhile.

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 4. **Trajectory classification.** Each monitoring question is placed in one of four categories based on the directional trends of harm-rate and exposure. Monitoring questions for which either trend could not be determined during estimation (§2.1) are labelled *unclassifiable* and do not enter the grid. The fill intensity encodes governance urgency.

3. Application and Results

3.1. Case Study: Conversational AI systems and self-harm.

Monitoring question: [S] Among people living in the United States [O] who use conversational AI systems for emotional support, [R] how many receive responses that encourage, or fail to discourage, suicidal ideation or self-harm [T] per calendar year?

Harm estimation, H: No authoritative data source is publicly available for this MQ; we begin at Tier 2. Considering the two most recent calendar years, LLM analysis of the AIAIID found two full matches in 2024 and 17 in 2025. Having two matches is below the threshold for a reliable signal; we can supplement with an alternative database. LLM analysis of the OECD AIM dataset after filtering for US-based incidents involving conversational AI systems resulting in physical or psychological injury yields 8 full matches in 2024, with a harm count range between 9 and 17. 2025 results reflect 55 full matches, with a harm count in the 100k range — an explosive increase. As an upper bound for 2025, OpenAI reported that approximately 0.15% of its weekly active users engage in conversations indicating potential suicidal planning or intent, representing more than one million people per week globally (Zeff, 2025), but no upper bound for 2024 could be identified.

Trend: *Increasing* [$H\ \uparrow$]. OECD AIM results increase over consecutive time periods.

Confidence: Low. The OECD AIM results support the trend of increasing frequency and severity, but the limited AIID matches and upper-bound proxy measures likely reflect limited awareness and detection methods in 2024. In response, companies have begun strengthening safeguards, including OpenAI's efforts to improve model behavior in sensitive conversations in late 2025. Given these shifts in measurement and mitigation, either expert elicitation or close monitoring of 2026 data is necessary before drawing high-confidence conclusions about trend, severity, and the effectiveness of these new interventions.

Exposure estimation, E**:** In the absence of direct survey data on emotional support use, we used Pew Research data on ChatGPT use "to learn new things" and "for entertainment" by age group (2024–2025) as a suitable proxy (Sidoti & McClain, 2025). We extracted both data points and applied them to the US population in the corresponding age groups. For the point estimate, we used the mid-point between the two data points; for the lower bound we used the lower estimate ("for entertainment"), and for the upper bound we used "to learn new things". Additionally, for the point estimate we assumed that ChatGPT holds 80% of the market share of LLM personal use (FATJOE, 2025). For the upper and lower bounds, we assumed a market share of 90% and 70%, respectively. In 2024, we estimate that 64 million people in the US used conversational AI for emotional support (54–73 million, plausible range). In 2025, we estimate 88 million (75–99 million, plausible range). Order-of-magnitude estimate: 10^8.

Trend: Increasing [$E\uparrow$]. Our estimates suggest an approximately 40% increase from 2024 to 2025.

Confidence: Medium. Tier 2 (derived from reasonable

sources).

Classification: *Escalating*. Both the exposure trend $E\uparrow$ and the harm-per exposure trend $\hat{H}\uparrow$ are *increasing* (more people are exposed, per-unit-exposure is more harmful).

3.2. Case Study: Autonomous vehicles and injury/damage.

Monitoring question: [S] Among autonomous vehicles (SAE Levels 3 through 5) [O] on US public roads, [R] how many incidents involving injury or property damage occur per million vehicle-miles [T] per calendar year?

Harm estimation, H: The US National Highway Traffic Safety Administration (NHTSA) requires manufacturers and operators to report certain collisions involving vehicles equipped with automated driving, or SAE Level 2 advanced driver assistance, systems. It can therefore be used as an authoritative source. The number of Automated Driving System (ADS) incidents increased from 526 in 2024 to 975 in 2025, a $\sim$85.4% increase. This is primarily driven by property damage cases rather than injury-related events.

Trend: *Increasing* [$H\ \uparrow$]. NHTSA results increase over consecutive time periods.

Confidence: *High*. Tier 1: Mandatory reporting requirements ensure that the NHTSA provides a comprehensive dataset for analysis.

Exposure estimation, E: From June 2024 to May 2025, the Autonomous Vehicle Industry Association (AVIA) estimated that AVs drove 145 million miles on US public roads, compared to 75 million miles in 2023–2024, suggesting a doubling in exposure in one year (Autonomous Vehicle Industry Association, 2025). Other sources suggest this trend continued rapidly throughout 2025, with Waymo delivering approximately 250,000 paid rides per week in April 2025 and 450,000 by December 2025, an 80% increase over 8 months (CNBC, 2025). For our point estimate, we used the AVIA estimate and assumed the monthly growth rate suggested by Waymo's 2025 reporting. For the lower bound we used the AVIA's May 2024 and May 2025 estimates for the whole of 2024 and 2025, respectively. For the upper bound, we increased the point estimate by 10%, similar to the difference between the lower bound and the point estimate. In 2024, we estimate that AVs drove 78 million miles in the US (75–86 million miles, plausible range). In 2025, we estimate 156 million miles (145–171 million miles, plausible range). Order-of-magnitude estimate: 10^8.

Trend: Increasing [$E\ \uparrow$]. Exposure is estimated to have approximately doubled between 2024 and 2025.

Confidence: Medium. Tier 2 (derived from reasonable sources).

Classification: *Mitigating*. Exposure growth ($\sim$100%) outpaces harm growth ($\sim$85%), yielding a decreasing harm-per-exposure trend [$\hat{H}\downarrow$] against a rising exposure trend [$E\ \uparrow$]. Fewer incidents occur per million vehicle-miles, suggesting that current safeguards are keeping pace with deployment. However, absolute harm might still be rising and warrants continued monitoring.

3.3. Practitioner Demand and Applicability

Between January and April 2026, we conducted semi-structured interviews with 42 AI risk and safety professionals spanning industry/commerce, government/civil society, and research/academia, to inform the design and assess the applicability of our methods. We found cross-sector enthusiasm for identifying clear, measurable incident types and estimating their changing impact over time (insight absent from other news and information sources). Noted applications included government research, evidence-based advocacy, governance investment justification, and updates to the EU AI Act's Code of Practice. The diversity of disciplinary approaches to AI risk and harm informed our design of the SORT monitoring question, enforcing analytical consistency while giving full control over the incident types defined.

To quantify these observations, we administered a Likert-scale questionnaire to 44 respondents (29 also participated in consultations). 90% rated our aims as needed or greatly needed. 80% would apply our methods directly and 95% would use others' results if methodologically sound, suggesting the existence of a latent analytical community. The detailed findings by stakeholder type and survey results are documented in Appendix C.

4. Discussion

The consultations establish the practical demand for insight; the case studies demonstrate the practical difficulties of incident reporting and the clarifying power of the classification process, both in illuminating trends and revealing where data falls short, with implications for the AI governance ecosystem.

4.1. What classification reveals

Trends in raw incident counts can be misleading, and the exposure trend is needed to provide context. Despite their comparable harm trends, Example 1 (conversational AI systems and self-harm) and Example 2 (autonomous vehicles and injury/damage) are classified *escalating* and *mitigating*, due to their differing exposure trends (**P2**).

The classification process reveals the data constraints on both harm and exposure, including which is more limiting. The tiered estimation approach makes explicit the evidential strength of each estimate. *Unclassifiable* is a valid result (**P3**), following abstention on either S and O estimates (**P3**).

In Example 1: No high quality data is available from an authoritative source, so OECD AIM data (*Tier 2; Proxy-based*) was used as the lower bound harm estimate. Two "category adjacent" indirect (survey) proxies provide the basis for a (*Tier 2; Proxy-based Fermi decomposition*) exposure estimate (in the absence of system-provider disclosures of conversational AI use).

In Example 2: NHTSA data are sufficiently relevant and precise for a (*Tier 1; Direct data*) harm estimate, whereas three partially relevant and overlapping "within category" sources combine for a (*Tier 2; Proxy-based*) exposure estimate.

Appendix D applies the framework to four further monitoring questions, which between them exercise every tier of the estimation procedure and surface the kinds of data constraints practitioners are likely to encounter. D.3 (deepfake investment scams) and D.4 (AI voice-clone fraud) show the inverse pattern to Example 2: a near-flat exposure denominator against sharply rising harm, where the *escalating* classification is driven by the harm mechanism rather than deployment growth. D.1 (AI-enabled cyberattacks on financial firms) and D.2 (facial-recognition wrongful arrest) illustrate classifications at the edge of what current data can support. The framework surfaces this as low confidence and flags expert elicitation as needed, communicating uncertainty clearly.

Classification based on estimated trends, rather than absolute values, is more robust to some sources of uncertainty, including systematic reporting biases in incident databases (**P4**). Provided reporting propensity is approximately stable over time, the relationship between absolute harm and its lower and upper bounds likely also remains stable across estimation periods, such that uncertainty in order-of-magnitude absolute levels cancels out in the trend, even for wide confidence intervals. The effect weakens as incident counts drop and small-sample noise grows, and provides no benefit for non-stationary biases (e.g. viral media incidents) or harms that are invisible to and systematically excluded from the database.)

Where confidence in both trends is sufficient, the relative movement of harm and exposure can also inform the choice of intervention: exposure-dominant trajectories point toward deployment, access, or eligibility measures, while harm-per-exposure-dominant trajectories point toward the harm mechanism itself.

In Example 2, exposure is rising faster than harm-per-exposure is falling, suggesting the absolute harm will continue to rise if current trends continue, despite a mitigating *classification*.

The coarseness of the 2×2 classification balances interpretive value with data requirements, asking only which trend is dominant (**P5**).

4.2. Toward routine trajectory classification

Exploring variations of a given MQ indicates that different specifications can affect the classification. Where partial matches substantially outnumber full matches, the MQ may be overspecified. Broadening its specification (by relaxing S or O) can increase the number of full matches, increasing statistical confidence at the cost of some precision. (**P1**) Alternatively (as in Example 1) adjacent datasets can, where available, serve as proxy measures to strengthen confidence in a trend estimate based on sparse data without changing the MQ. (**P1, P3**)

The modular structure (SORT, estimation, classification) encourages reproducibility, transparency and partial recycling of estimation efforts. The interactive SORT tool and methodological guidance lower barriers to use. A public repository of answered MQs (with sources and assumptions) could provide a common analytical resource. Sharing intermediate outputs (attempted S/O specifications; resulting partial-match ratios) would support convergence on the most informative formulations. Growing proxy-data libraries would reduce the marginal cost of each new MQ. At scale, a well-structured repository could reveal a taxonomy of incident types that reflects analysts' framings and priorities.

Applied repeatedly over time, the classification process supports institutional learning: transitions between trajectory categories are arguably more policy-relevant than individual classifications. *Escalating* to *mitigating* may suggest successful intervention, whereas *receding* to *concentrating* signals deepening harm within a subpopulation. Current regulatory regimes are unlikely to produce sufficient incident data for immediate application, but demonstrating governance value may encourage more comprehensive reporting. With broader adoption and shorter reporting timescales, governance insight could progress from broad trajectory identification to detection of phase transitions and rapidly scaling incident types, enabling earlier intervention and more efficient resource allocation.

4.3. Limitations

The framework's view of harm is necessarily constrained. MQs can specify harm in detail, including the nature and severity, but counting incidents collapses this specification to a binary (match or no match), concealing the specifics of each case. (Future work might recover more of this detail through advanced harm characterisation and database filter-

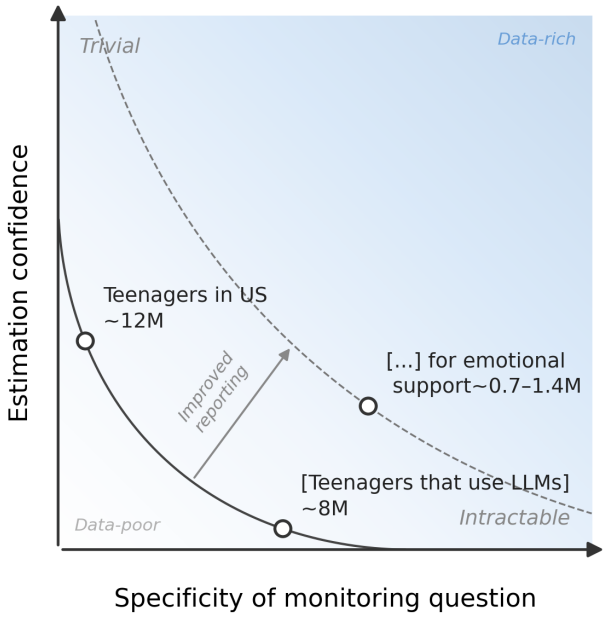

Figure 5. **Specificity–confidence trade-off in exposure estimation.** More precise definitions of the exposed population (S and O) yield more meaningful monitoring questions but in turn become less tractable. Improved data quality shifts the frontier outward. The optimal MQ is one on the most outward frontier.

ing.) Different MQs capture qualitatively different harms that are not easily comparable. The framework supports within-MQ trend reasoning more robustly than cross-MQ prioritisation, which should be approached with particular care. Principled abstention catches thin data but not absent categories. Detecting harms that are absent from incident data requires separate discovery methods.

Classification assumes roughly monotonic trends across the timeframe T: a spike-and-reversal within a single period is invisible and in practice T is typically constrained by the temporal resolution of available proxy data rather than the incident data. The assumption of stable reporting propensity often does not hold in practice due to shifting news cycles and viral media events (which deeper analysis using Google Trends data on fluctuations in public attention can help identify), and other factors, including new reporting regulations or database expansions. These limits on comparability across periods should be acknowledged as a general limitation and noted as a caveat where observed.

The classification process identifies directional trends, but cannot locate absolute levels of harm ("compressed" or otherwise) within the confidence interval. Trends can indicate whether existing interventions are working, but deciding what level of harm warrants action is a matter of judgement. Classifications do not translate directly into priorities: a *mitigating* classification can still imply growing absolute harm if exposure rises faster than harm-per-exposure falls; a *receding* classification may still represent a large absolute burden; a *concentrating* classification at high levels may warrant a more urgent response than an *escalating* one at low levels. At worst, this presents a risk of misuse. *Receding* or *unclassifiable* labels could be invoked to justify inaction, when the first may reflect successful intervention worth sustaining and the second a data gap worth closing.

Incident databases are retrospective. AI harms are, by definition, AI risks that have materialised at least once. They represent a small and lagging subset of the full, expanding landscape of AI risks. The most consequential of these may have yet to materialise, and incident data can do little to help anticipate or directly prevent them. Limited extrapolation beyond observed harms may be possible, for example, by considering how harm might change if the underlying variables (system properties, deployment scale, user population) shifted, but this remains a structured way of reasoning about known harm mechanisms, not a means of surfacing unknown ones.

More broadly, the framework is designed to replace naive interpretations of incident data—the conflation of rising reports with rising harm, or the treatment of absent data as evidence of absent harm—rather than reinforce them. Its value derives from, and depends on, its assumptions and limitations being made explicit.

Impact Statement

The framework replaces the current path from raw counts through naive trend claims to reactive governance with decomposed trends, trajectory classification, and prioritisation grounded in exposure-adjusted harm. Current reporting regimes are narrowly scoped and unlikely to produce sufficient data on their own; the framework is designed to work with the news-sourced databases that remain the primary record of AI harm, and to demonstrate the governance value that could motivate broader reporting infrastructure over time. It makes visible whether rising counts reflect genuine risk escalation due to more harmful or more widely deployed AI systems, or simply increased media interest; where reporting infrastructure is insufficient to support trend claims; and where unclassifiable MQs in expected-harm domains indicate survivorship bias in the reporting funnel.

References

Abraham, S., Chen, T., Chhun, C., Jaramillo-Gutierrez, G., Mylius, S., Raaj, S., Slattery, P., and McGregor, S. AI incident monitoring through a public health lens. *arXiv preprint arXiv:2604.19914*, 2026.

AI Incident Database. Editor's guide, 2025a. URL https://incidentdatabase.ai/editors-guide/. Accessed: 2025-04-24.

AI Incident Database. Ai incident database, 2025b. URL https://incidentdatabase.ai. Accessed: 2025-04-24.

American Civil Liberties Union. More than a dozen wrongful arrests due to police reliance on facial recognition technology. American Civil Liberties Union, 2024. URL https://www.aclu.org/news/privacy-technology/more-than-a-dozen-wrongful-arrests-due-to-police-reliance-on-facial-recognition-technology. Accessed: 2026-04-22.

Autonomous Vehicle Industry Association. 2025 state of AV: Annual report. Technical report, Autonomous Vehicle Industry Association, 2025. Accessed: 2026-04-21.

Chamber of Commerce. Data reveals landline phone decline statistics. Chamber of Commerce, 2024. URL https://www.chamberofcommerce.org/data-reveals-landline-phone-decline-statistics/. Accessed: 2026-04-24.

CNBC. Waymo paid rides surge as robotaxi competition with Tesla heats up. CNBC, December 2025. URL https://www.cnbc.com/2025/12/08/waymo-paid-rides-robotaxi-tesla.html. Accessed: 2026-04-21.

Dolin, P., Li, W., Dasarathy, G., and Berisha, V. Statistically valid post-deployment monitoring should be standard for ai-based digital health. *arXiv preprint arXiv:2506.05701*, 2025.

FATJOE. LLM market share and usage statistics. FATJOE, 2025. URL https://fatjoe.com/. Accessed: 2026-04-21.

Federal Bureau of Investigation. 2022 internet crime report. Technical report, Internet Crime Complaint Center (IC3), 2023. URL https://www.ic3.gov/AnnualReport/Reports/2022_IC3Report.pdf. Accessed: 2026-04-22.

Federal Bureau of Investigation. 2023 internet crime report. Technical report, Internet Crime Complaint Center (IC3), 2024. URL https://www.ic3.gov/AnnualReport/Reports/2023_IC3Report.pdf. Accessed: 2026-04-22.

Federal Bureau of Investigation. 2024 internet crime report. Technical report, Internet Crime Complaint Center (IC3), 2025. URL https://www.ic3.gov/AnnualReport/Reports/2024_IC3Report.pdf. Accessed: 2026-04-22.

Federal Trade Commission. Consumer sentinel network data book 2024. Technical report, Federal Trade Commission, March 2025. URL https://www.ftc.gov/reports/consumer-sentinel-network-data-book-2024. Accessed: 2026-04-22.

Financial Stability Board. 2024 list of global systemically important banks (G-SIBs). https://www.fsb.org/2024/11/2024-list-of-global-systemically-important-banks-g-sibs/, November 2024. Accessed: 2026-04-24.

Forbes. The Global 2000. https://www.forbes.com/lists/global2000/, 2024. 2024 edition released 6 June 2024. Accessed: 2026-04-24.

Garvie, C., Bedoya, A., and Frankle, J. The perpetual line-up: Unregulated police face recognition in america. Technical report, Center on Privacy & Technology at Georgetown Law, 2016. URL https://www.law.georgetown.edu/privacy-technology-center/publications/the-perpetual-line-up/. Accessed: 2026-04-22.

Jeanmaire, C. and Boger, S. AI Incidents Are Rising. It's Time for the United States to Build Playbooks for When AI Fails. The Future Society, November 2025. URL https://thefuturesociety.org/us-ai-incident-response/. Accessed: 2026-04-15.

Johnson, T. L., Johnson, N. N., Topalli, V., McCurdy, D., and Wallace, A. Police facial recognition applications and violent crime control in us cities. *Cities*, 155:105472, 2024.

McCarty, D. J., Tull, E. S., Moy, C. S., Kwoh, C. K., and LaPorte, R. E. Ascertainment corrected rates: Applications of capture-recapture methods. *International Journal of Epidemiology*, 22(3):559–565, 1993. doi: 10.1093/ije/22.3.559.

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

Nixon, R. *Slow Violence and the Environmentalism of the Poor*. Harvard University Press, 2011.

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

OECD. Defining AI incidents and related terms. OECD Artificial Intelligence Papers 16, Organisation for Economic Co-operation and Development, Paris, 2024b.

OECD. Oecd ai incidents and hazards monitor, 2025. URL https://oecd.ai/en/incidents. Accessed: 2025-04-24.

Paeth, K. and McGregor, S. AI risk, safety, and incident reporting. In Xu, W. (ed.), *Handbook of Human-Centered Artificial Intelligence*. Springer, Singapore, 2025. doi: 10.1007/978-981-97-8440-0_89-1.

Paeth, K., Atherton, D., Pittaras, N., Frase, H., and McGregor, S. Lessons for editors of AI incidents from the AI incident database. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 28946–28953, 2025. doi: 10.1609/aaai.v39i28.35163.

Perrow, C. Normal accidents: Living with high risk technologies-updated edition. *Princeton university press*, 2011.

Pew Research Center. Demographics of internet and home broadband usage in the United States. Pew Research Center, 2024. URL https://www.pewresearch.org/internet/fact-sheet/internet-broadband/. Accessed: 2026-04-24.

Rao, A., Keller, D., Kalra, N., Steed, R., Kwegyir-Aggrey, K., Klyman, K., Staheli, D., and Bergman, S. Challenges to the monitoring of deployed ai systems: Center for ai standards and innovation. *NIST*, 2026. doi: 10.6028/NIST.AI.800-4.

Richards, I., Benn, C., and Zilka, M. From incidents to insights: Patterns of responsibility following AI harms. *Proceedings of the 5th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, 2025. doi: 10.1145/3757887.3763018.

Richardson, W. S., Wilson, M. C., Nishikawa, J., and Hayward, R. S. The well-built clinical question: A key to evidence-based decisions. *ACP Journal Club*, 123(3): A12–A13, 1995.

Sidoti, O. and McClain, C. 34% of U.S. adults have used ChatGPT, about double the share in 2023. Pew Research Center, June 2025. URL https://www.pewresearch.org/short-reads/2025/06/25/34-of-us-adults-have-used-chatgpt-about-double-the-share-in-2023/. Accessed: 2026-04-21.

Stein, M., Bernardi, J., and Dunlop, C. The role of governments in increasing interconnected post-deployment monitoring of ai. *arXiv preprint arXiv:2410.04931*, 2024.

Tanjaya, A. and Pratt, J. Documenting the impacts of foundation models. *Partnership on AI*, 2025.

Teo, S. A. Artificial intelligence and its 'slow violence'to human rights. *AI and Ethics*, 5(3):2265–2280, 2025.

The Record. Clearview AI police searches doubled in 2023. The Record by Recorded Future, 2024. URL https://therecord.media/clearview-ai-police-searches-doubled-2023. Accessed: 2026-04-24.

U.S. Government Accountability Office. Facial recognition services: Federal law enforcement agencies should take actions to implement training, and policies for civil liberties. Technical Report GAO-23-105607, U.S. Government Accountability Office, 2023. URL https://www.gao.gov/products/gao-23-105607. Accessed: 2026-04-22.

U.S. Government Accountability Office. Facial recognition technology: Federal law enforcement agency efforts related to civil rights and training. Technical Report GAO-24-107372, U.S. Government Accountability Office, 2024. URL https://www.gao.gov/products/gao-24-107372. Accessed: 2026-04-22.

Verizon. 2024 data breach investigations report. Verizon, 2024. URL https://www.verizon.com/dbir. Accessed: 2026-04-23.

Verizon. 2025 data breach investigations report. Verizon, 2025. URL https://www.verizon.com/dbir. Accessed: 2026-04-23.

Wei, K. and Heim, L. Designing incident reporting systems for harms from general-purpose AI. *arXiv preprint arXiv:2511.05914*, 2025. doi: 10.48550/arXiv.2511.05914.

Whittlestone, J. and Clark, J. Why and how governments should monitor ai development. *arXiv preprint arXiv:2108.12427*, 2021.

Zeff, M. Openai says over a million people talk to ChatGPT about suicide weekly. TechCrunch, October 2025. URL https://techcrunch.com/2025/10/27/openai-says-over-a-million-people-talk-to-chatgpt-about-suicide-weekly/. Accessed: 2026-04-23.

A. Appendix

The pathway to inclusion in a public incident database can be thought of as a funnel with sequential filters through which an incident must pass:

- Detection: the harm occurs and is detectable
- Attribution: the harm is recognised as an AI failure
- Recording: a record of the harm is created and survives
- Disclosure and reporting: the record reaches someone who reports it
- Capture: the database picks up the report
- Within scope: the report fits the database's definitional scope

A.1. Systematic biases in incident reporting

A range of biases (including visibility, detection, disclosure incentive, media salience, geography/language and victim voice) operate across the stages and compound multiplicatively. This makes certain types of incidents more likely to be included than others:

High inclusion probability: Acute, dramatic, novel, unambiguously AI-related, consumer-facing harm to well-resourced, individual (rather than systemic), English-language, vocal victim(s), in regulated sectors and fashionable domains. E.g. a US-based LLM chatbot produces a shocking output, goes viral on social media, and the company issues a public statement; an autonomous vehicle is involved in a fatal collision investigated by a regulator.

Moderate inclusion probability: Incidents that are detectable and recorded but face friction at the attribution or disclosure filter. E.g. algorithmic bias cases pursued through litigation, incidents surfaced by academic surveys rather than the media, terrorism-related incidents which are suppressed in the interests of national security.

Low inclusion probability: Slow, subtle, chronic, diffuse, systemic harm, harder to attribute, affecting marginalised or geographically remote populations, in unregulated domains, with non-disclosure incentives and no or limited institutional record. E.g. a credit-scoring model in a non-English-speaking country penalises residents of certain postcodes; harm is real but diffuse, no individual claimant has the resources to pursue it, no journalist reports it, and the system is eventually replaced without the pattern ever being formally named.

A.2. Incident databases in practice: the AIID and OECD AIM

The AI Incident Database (AIID) (AI Incident Database, 2025b) and the OECD AI Incidents and Hazards Monitor (AIM) (OECD, 2025) are among the most well known and broadly populated public AI incident databases. They have similar basic structures (one overarching incident title aggregates content—predominantly news articles—from multiple outlets on the same incident) but diverge somewhat in their editorial approach and sourcing methodology, an understanding of which is helpful in contextualising their use in harm classification.

The *AIID Editor's Guide* (AI Incident Database, 2025a) defines an AI incident as: "an alleged harm or near harm event to people, property, or the environment where an AI system is implicated."

The *AIM Overview and Methodology* (OECD, 2024a) defines an AI incident as "an event, circumstance or series of events where the development, use or malfunction of one or more AI systems directly or indirectly leads to [a specific set of] harms." The methodology also defines an AI hazard separately as "an event, circumstance or series of events where the development, use or malfunction of one or more AI systems could plausibly lead to an AI incident." (The database can be filtered to show either or both of these.)

At the time of writing, the AIID recorded 1,460 incidents, and the OECD AIM recorded 9,218 incidents and 5,312 hazards (14,530 incidents and hazards).

The AIM sources its data in the form of clusters of articles reporting on the same AI-related event (not pre-filtered for harm or hazard) from a news intelligence platform. It uses LLMs to classify and filter events as incidents, hazards, or unrelated

content. It uses a rolling four-day processing window, such that related articles appearing more than four days apart risk misclassification as separate events. In practice, the filters also capture some reports that do not meet the strict definition of a discrete incident, including composite narratives, pattern-explanation stories, and accounts of potential vulnerabilities.

The AIID sources its data through paid news subscriptions, keyword-based alerts and searches, with some LLM and Google Translate-supported coverage of non-English sources. The editorial process is manual, with an ethos of comprehensive collection and more precise filtering for discrete incidents (although a few incident titles are used to cluster alike incidents). Reports are predominantly news articles, though legal filings and peer-reviewed articles are occasionally included.

The two databases likely draw on substantially overlapping, though not identical, sources, and are subject to similar reporting biases inherent to news-based sourcing. Both databases are valuable resources in their own right, and together offer complementary coverage of how and to whom AI systems cause harm.

Researchers using either database to count incidents should be attentive to how each database is constructed, in particular:

- The two databases define incidents slightly differently, though the SORT Framework's precise specification of the incident type of interest should help navigate this.
- In practice, news articles about the same incident are often published days, months, and even years after the event, as initial reports give way to analytical pieces and trend stories, meaning the AIM's four-day processing window likely has a real impact on its incident counts, with single incidents fragmented across multiple entries.
- The AIM's tendency to capture composite narratives alongside discrete incidents, and the AIID's occasional use of cluster-level incident titles, introduce further fuzziness into any count.

That said, given the current systemic underreporting of AI-related harms, even a database that overcounts some incidents is likely to represent a reasonable lower bound for most forms of harm.

B. Appendix

Consultation Methodology

Between January and April 2026, we conducted semi-structured consultations (31 video interviews, 11 written responses) with 42 AI risk and safety professionals across 9 sectors (see Table 3). Participants were given an overview of our research aims and asked: whether and how they currently monitor AI risks and incidents; what they found most difficult about staying informed; how useful they would find our proposed classification methods; and what they would need for the outputs to be applicable to their work. These consultations informed the design of our classification methods and helped assess their potential utility and applicability across a range of professional contexts and disciplinary perspectives on AI risk and harm.

Following the consultation phase, we administered a structured five-point Likert-scale questionnaire to 44 respondents (29 of whom also participated in the consultations), covering current practices and experiences of staying informed about AI risk and harm, the perceived need for our project aims and willingness and ability to apply our methods. (Our project aims were described as: identifying clear, measurable incident types and estimating their frequency, severity, and change over time.)

Survey results were grouped by sector to improve statistical significance and reflect broadly similar conceptualisations of AI risk and harm:

- through operational controls and compliance ("industry and commercial", $n = 19$, combining financial services, consulting and technology);
- through empirical research and capability testing ("research and academia", $n = 14$, combining academia and non-governmental AI technical research and testing);
- through policy, advocacy and regulation ("policy and civil society", $n = 11$, combining civil society organisations and government policy, research and regulation).

Table 3. Consultation and survey participant role, sector and count

Survey count	Sector	Interviewee/respondent role	Interview count
9	Academia	AI policy researcher; AI safety doctoral student; AI safety professor; cybersecurity postgraduate; security professor; undergraduate AI research student	6
5	AI technical research/testing (non-governmental)	AI safety researcher (2); AI security founder	3
7	Civil society, NGO, think tank	AI governance specialist; AI policy advisor; AI risk manager; AI safety advocate; AI safety founder; chief AI officer; research director	7
9	Consulting/advisory/independent	AI governance consultant (3); AI governance manager; AI risk advisor; AI risk consultant; AI security consultant; data privacy consultant; organisational psychologist; risk consultant; risk/audit consultant	10
9	Financial services (banking, insurance, payments)	AI governance auditor; AI governance lead; AI security lead; head of analytics; incident manager; risk manager; security officer	7
4	Government policy, research or regulation	AI governance advisor; AI risk advisor (2); AI risk researcher; AI testing coordinator; regulatory researcher	5
0	Healthcare	AI security architect; AI strategist	2
0	Media and communications	AI safety journalist	1
1	Technology	AI project lead	1
44	9 sectors	40 interviewee/respondent roles	42

Table 4. Aggregated Likert responses by sector group (n = 44).

Group	1	2	3	4	5
How much do you need to stay up to date with developments in AI risk and harm to do your job? (1 = not at all, 5 = very much)					
Industry/commercial (n=19)	0 (0%)	1 (5%)	2 (11%)	4 (21%)	12 (63%)
Research/academic (n=14)	0 (0%)	0 (0%)	0 (0%)	3 (21%)	11 (79%)
Policy/civil society (n=11)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	11 (100%)
How hard is it to know which AI risks/harms have the greatest impact? (1 = very hard, 5 = very easy)					
Industry/commercial (n=19)	1 (5%)	5 (26%)	3 (16%)	7 (37%)	3 (16%)
Research/academic (n=14)	2 (14%)	2 (14%)	6 (43%)	4 (29%)	0 (0%)
Policy/civil society (n=11)	2 (18%)	1 (9%)	4 (36%)	4 (36%)	0 (0%)
How hard is it to know which AI risks/harms are increasing in frequency or severity? (1 = very hard, 5 = very easy)					
Industry/commercial (n=19)	1 (5%)	4 (21%)	4 (21%)	6 (32%)	4 (21%)
Research/academic (n=14)	2 (14%)	2 (14%)	3 (21%)	4 (29%)	3 (21%)
Policy/civil society (n=11)	1 (9%)	3 (27%)	1 (9%)	5 (45%)	1 (9%)
How needed are better methods for estimating the severity/impact of different types of AI incident? (1 = not needed, 5 = greatly needed)					
Industry/commercial (n=19)	0 (0%)	0 (0%)	1 (5%)	6 (32%)	12 (63%)
Research/academic (n=14)	0 (0%)	0 (0%)	2 (14%)	4 (29%)	8 (57%)
Policy/civil society (n=11)	0 (0%)	0 (0%)	0 (0%)	3 (27%)	8 (73%)
How needed are better methods for estimating how different types of AI incident are changing over time? (1 = not needed, 5 = greatly needed)					
Industry/commercial (n=19)	0 (0%)	0 (0%)	1 (5%)	4 (21%)	14 (74%)
Research/academic (n=14)	0 (0%)	0 (0%)	2 (14%)	4 (29%)	8 (57%)
Policy/civil society (n=11)	0 (0%)	0 (0%)	1 (9%)	3 (27%)	7 (64%)

Overall Results

Enthusiasm for the project aims was broad across all sectors and role types in consultations. Survey respondents rated all aims highly, with a majority giving the maximum score "greatly needed", and no aim rated as "not needed".

Interviewees across sectors described current reliance on news alerts, personal networks and ad hoc searches, and noted that reliably establishing whether a specific incident type is increasing or decreasing in frequency or harm is difficult by these means. The two tasks survey respondents found hardest—knowing which risks are increasing and decreasing—precisely match our project aims.

Outside of the specific project aims, access to more incident data, and the ability to filter and sort it according to requirements, was the most frequently requested feature. Despite the existence of various industry and cross-sector standards and frameworks there is no universal taxonomy or common articulation of incident types. AI risk and harm are so domain-specific and sociotechnically complex that practitioners' ways of thinking about them are as varied as their professional contexts. This informed our design choice of the SORT monitoring question: it provides analysts with a consistent structure for meaningful searching while giving them full control over the incident types they define.

Despite being informed that a single monitoring question takes one to two hours to evaluate, 35 respondents (79.5%) said they would have time and 37 (84.1%) the confidence to implement the methods themselves, with 42 (95.5%) saying they would also use results on a dashboard, if they could be trusted. Trust would require: transparent and reproducible methodology, clearly defined inclusion criteria, explicit acknowledgement of reporting bias and data gaps, and consistency over time.

Limitations and Caveats

Interviewees for whom the methods were unlikely to be directly practical were either working upstream of incidents in technical threat modelling or AI safety testing, where decisions and priorities are driven by model capabilities and deployment timing, or in roles sufficiently structured by existing frameworks, policies and controls that incident tracking was a secondary concern.

Several interviewees raised concerns about the scope of what incident databases can capture—points generally reflected in our limitations—showing the importance of communicating clearly to prospective users to avoid misinterpretation of what incident trend data can and cannot show.

By Sector

Financial Services

Interviewees were consistently enthusiastic, the most focused on current (rather than prospective) harms, and the most acutely aware of the monitoring gap. A lack of reliable incident tracking or cross-organisational classification consistency resulted in compliance teams reliant on informally shared news links. Where available, sector-specific trend data strengthens the case for security reviews and governance investment and can inform insurance underwriting. Financial services survey respondents rated tracking change over time highest of any sector, while reporting the greatest difficulty with current monitoring.

Consultants, Auditors and Governance Advisors

Enthusiasm centred on justifying governance investment and standardising classification. Linking trends to established frameworks—ISO 42001, NIST RMF, and IEEE standards—would make action explicit, though evaluation periods may be longer than practitioners need given current data availability.

Civil Society, Advocacy and NGOs

Interviewees were among the most enthusiastic for support to evidence-based advocacy, but also the most capacity constrained. Quantifying harm is valuable for planning, prioritisation and communication. Demonstrating increasing harm would make advocacy substantially more persuasive to the small number of well-placed decision-makers who can meaningfully reduce AI harms. Granularity in trend data dictates the resolution at which advocacy can be targeted, so any improvement could have second order effects. An EU policy advisor observed that better tools are needed for monitoring

how risks change over time, and that a centralised source of such changes would directly support updates to the EU AI Act's Code of Practice.

GOVERNMENT POLICY, RESEARCH AND REGULATION

Enthusiasm was strong. Given government interest in the whole spectrum of AI risks, even a lower bound on incident prevalence would be a widely useful tool for risk assessments and inter-departmental communications. Estimating the prevalence of specific issues is also a low cost way to scope new research.

ACADEMIC AND INDEPENDENT RESEARCHERS

Researchers and independent practitioners are the most willing and able to implement the methods themselves. Enthusiasm centred on structure and pattern recognition: the potential to transform fragmented observations into coherent, evidence-based narratives. A technical AI safety researcher noted the potential for reproducible classifications to be cited in legal filings.

HEALTHCARE

Incident trend data would support regulatory submissions requiring consideration of the magnitude of potential harm, and supplement frameworks that provide risk data but do not rank by prevalence or trajectory.

C. Appendix

C.1. LLM Assessor Pipeline

Algorithm 1 LLM-based harm assessment (lower-bound estimate)

Input: Monitoring question (SORT); an incident database; two time periods
Output: A lower-bound harm trend; per-incident assessments

foreach *incident in the database* **do**
  Assess subject and risk-event match as TRUE, FALSE, or INDETERMINATE
  Extract harm quantity
Separate full matches from partial matches (both TRUE or any INDETERMINATE)
Compute harm totals for each period
if *fewer than 3 full matches in either period* **then**
  Abstain (insufficient evidence)
else
  Report increasing or decreasing accordingly
if *partial matches greatly outnumber full matches* **then**
  Flag: monitoring question may be overspecified

Algorithm 1 is implemented as three decoupled stages—loader, assessor, aggregator—driven by a single MQ config containing the SORT tuple (S, R) and either explicit comparison periods or a reporting frequency $\in \{\text{monthly, quarterly, yearly}\}$.

Loader. Supports the AIID snapshot (1,405 incidents, 6,787 reports) and OECD AIM exports. When a frequency is given, periods are the two most recent complete periods at that cadence, offset by a 3-month reporting buffer from the run date. By default the assessor receives only the incident's title, description, and alleged deployer/developer/harmed parties; optionally, the first linked report with non-empty text can be attached, truncated to 4000 characters.

Assessor. Per incident we call the OpenRouter Chat Completions API (default `anthropic/claude-3.5-haiku`; any OpenRouter model is substitutable and the exact string is logged per run) at temperature 0.1, JSON response mode, and a 1000-token cap. The prompt embeds the SORT fields and incident record and requires a flat JSON object with: $S\_match, R\_match \in \{\text{TRUE, FALSE, INDETERMINATE}\}$ with short reasoning; integer harm bounds [lower, upper] in the harm unit with reasoning; and a suggested proxy measure. The model is instructed to prefer INDETERMINATE over FALSE

when in doubt. Up to 3 attempts with exponential backoff and longer waits on rate-limit errors; exhausted retries yield both matches INDETERMINATE. Calls can be parallelised, and a single invocation can batch a list of MQs.

Aggregator. An incident is a *full match* iff $S\_match = R\_match =$ TRUE, and a *partial match* if either component is INDETERMINATE. Per-period bounds are unweighted sums over full matches, $H_p^{\text{lower}} = \sum_{i \in \text{Full}(p)} \text{lower}_i$ and analogously for the upper bound including partial matches.

Inter-assessor agreement. We validate pipeline output against human raters on two MQs (AV injury, AI chatbots), each scored once with `claude-3.5-haiku`. For each MQ, the sampler stratifies pipeline output into full/partial/negative and draws a seeded random sample of up to k per stratum; raters label a blinded, shuffled subset using the same schema as the pipeline (`S_match`, `R_match`, severity, harm-quantity bounds). We report Cohen's κ on `S_match` (3-way), `R_match` (3-way), and the binary full-match for all pairs of raters, treating the LLM as one rater among humans. This is a single-model, single-run evaluation: run-to-run stability of the same LLM and cross-model agreement are out of scope and left to future work.

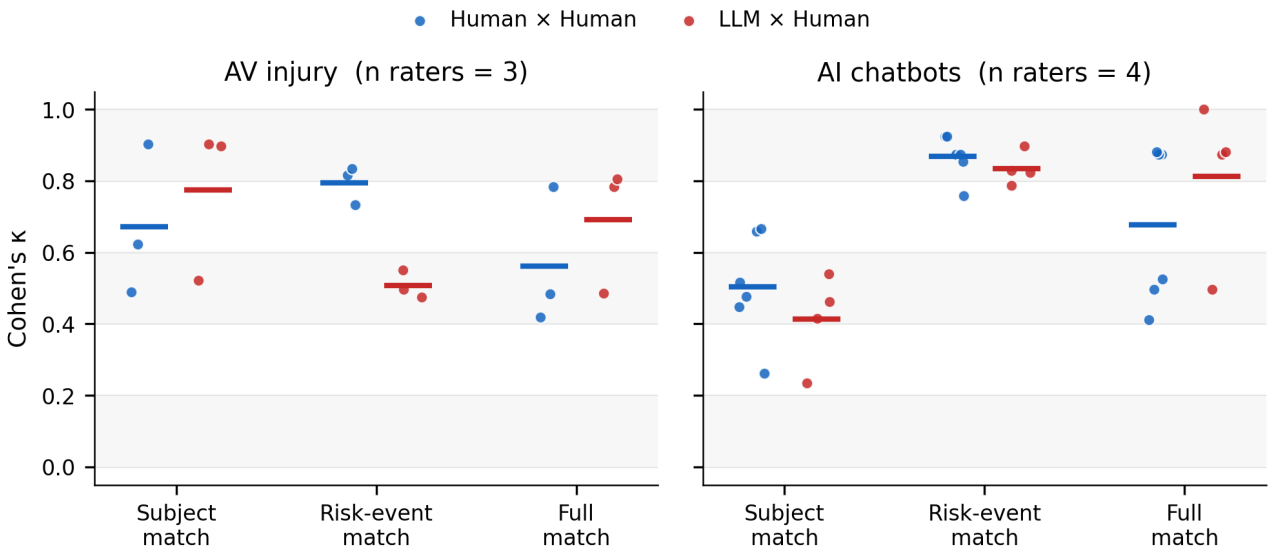

Figure 6. Inter-rater agreement (Cohen's κ) for each topic. Dots show pairwise κ between raters on the 3-valued S_match / R_match labels and the derived binary full-match; horizontal bars mark group means. Human–human pairs define the achievable ceiling; LLM–human pairs within that cluster indicate model agreement at human-panel quality. Shaded bands: Landis & Koch interpretation zones.

Results. Three raters scored AV injury and four scored AI chatbots. Figure 6 plots pairwise κ for every human-human pair (the ceiling implied by human disagreement) and every LLM-human pair. On the composite full-match decision the LLM sits inside the human-human cloud on both topics ($\bar{\kappa}_{\text{LH}} = 0.69$ vs. $\bar{\kappa}_{\text{HH}} = 0.56$ for AV injury; 0.81 vs. 0.68 for AI chatbots). The decomposed labels show topic-specific asymmetries: on AV injury the model matches or beats humans on subject identification (0.77 vs. 0.67) but falls below the ceiling on risk-event identification (0.51 vs. 0.80); on AI chatbots the pattern reverses—the model matches humans on risk-event (0.83 vs. 0.87) and is slightly below on subject (0.41 vs. 0.51).

Validity. With 3-6 human-human and 3-4 LLM–human pairs per topic, this is a validation sanity-check, not a powered test of equivalence: a nonparametric comparison of $\bar{\kappa}_{\text{HH}}$ and $\bar{\kappa}_{\text{LH}}$ from so few points has negligible statistical power. One LLM-human pair on AI chatbots yielded $\kappa = 1.0$ on full-match, a degree of agreement not seen in any other pair and consistent with possible label leakage during rating; we plot it for transparency but do not rely on it. Because we evaluated a single model on a single run, these results speak to agreement with human judgment on two MQs and not to the stability of that agreement across runs or models; a stronger validation would require at least a second independent run and a second model, plus ≥ 5 complete raters per topic with pre-registered bootstrap CIs on $\bar{\kappa}_{\text{LH}} - \bar{\kappa}_{\text{HH}}$.

D. Appendix

D.1. Example 3: AI cyberattacks on financial companies.

Monitoring question: [S] Among financial companies (1000+ employees) [O] operating globally, [R] how many are affected by AI-enabled cyberattacks or malicious exploitation of systems [T] per calendar year?

Harm estimation, H: There is no public, authoritative source for AI-enabled cyberattacks on organizations, so we apply the Tier 2 methodology. LLM analysis of the AIID found 1 full match in 2023, with a harm count of 1, and 3 full matches in 2024, with a harm count of 3. Similar data insufficiency is present in OECD AIM, with no full matches found in 2023 and 2 in 2024. Due to pervasive underreporting by companies, data is insufficient to identify the true incident count and severity. For the upper-bound proxy measure, we use the count of security incidents and data breaches for large financial organizations reported in the annual Verizon Data Breach Investigations Report. In 2023, there were 122 security incidents and 87 data breaches (Verizon, 2024), and in 2024, there were 134 and 117, respectively (Verizon, 2025).

Trend: *Indeterminate* [$H \uparrow$]. The number of security incidents and data breaches increased by $\sim$9.8% and $\sim$34.5%, respectively but it is unknown whether they are AI enabled.

Confidence: *Low*. An upper-bound proxy measure was identified, but analysing the AIID was indeterminate due to data insufficiency. As a result, expert elicitation is recommended.

Exposure estimation, E: We approximate exposure by the global count of financial-services firms with 1,000 or more employees — the population of firms large enough to plausibly be targeted by AI-enabled cyberattacks and to appear in breach-reporting proxies. No authoritative registry exists, so we construct bounds from three overlapping sources. The Financial Stability Board's 29 globally systemically important banks (Financial Stability Board, 2024) provide an inner anchor of firms that unambiguously meet the size threshold. The Forbes Global 2000 classifies roughly 500–600 companies in its Financials sector (Forbes, 2024), of which approximately 400 fall within banking, insurance, and capital-markets sub-industries relevant to [S]; this serves as a mid-range proxy since Global 2000 inclusion correlates with (but does not strictly enforce) the 1,000-employee threshold. Extending beyond publicly traded firms to include large private and mutual financial institutions via industry-database filtering yields an estimated 1,500–2,500 firms globally. For the point estimate we take 2,000; for the lower bound, 1,500 (near the Forbes-derived floor); for the upper bound, 2,500. Firm-count growth is modest and largely offset by M&A-driven consolidation, so the denominator is approximately stable over consecutive calendar years. Order-of-magnitude estimate: 10^3.

Trend: Approximately flat [$E \rightarrow$]. Firm-count changes are dominated by M&A activity rather than net growth; per-firm attack surface is rising sharply with AI-system adoption in financial services, a factor not captured in the firm denominator and flagged for expert elicitation.

Confidence: Low. Tier 2 (derived from reasonable sources, but not for target [S]).

Classification: ***Unclassifiable***. The AIID signal (1 full match in 2023, 3 in 2024; 0 and 2 respectively in OECD AIM) falls below the three-match threshold for a reliable trend claim in either period. The Verizon DBIR proxy captures security incidents and breaches at large financial firms but does not isolate the AI-attributable share, and cannot be relied on as a directional signal for this MQ without expert elicitation. Per §2.1, we abstain: current public evidence does not support an trend estimate for AI-enabled cyberattacks on large financial firms. The resulting gap, high policy demand with no defensible trend claim, is itself the governance-relevant finding, pointing to the need for mandatory AI-incident disclosure in the financial sector.

D.2. Case Study: Facial-recognition misidentification and wrongful arrest.

Monitoring question: [S] Among US individuals arrested by state or local law enforcement [O] in jurisdictions using facial recognition to generate investigative leads, [R] how many are wrongfully arrested due to facial-recognition misidentification [T] per calendar year?

Harm estimation, H: LLM filtering of the AIID returned six full matches in the 2022–2023 period (harm count six) and six full matches in the 2024–2025 period (harm count range 15–71). Full-match count is stable across the two periods, but the harm count range expands substantially, indicating rising per-incident severity. Data sufficiency is moderate and the partial-to-full ratio is 5.83. Documented wrongful-arrest cases compiled by the American Civil Liberties Union (American

Civil Liberties Union, 2024) provide a supplementary floor: one documented case in 2022, two in 2023, and two in 2024, accumulating to 14 cases across nine US states through 2024. These figures are lower bounds: they capture only cases that reached litigation or press coverage. A defensible upper bound is difficult to construct. GAO-23-105607 documents approximately 60,000 cumulative facial-recognition searches across seven US federal law-enforcement agencies, starting on different dates between early 2018 and mid-2019, through to April 2023 (U.S. Government Accountability Office, 2023), with GAO-24-107372 extending coverage into 2024 (U.S. Government Accountability Office, 2024). State and local search volumes are not systematically inventoried. The *Perpetual Line-Up* study estimated that more than 117 million US adults were enrolled in police face-recognition networks as of 2016 (Garvie et al., 2016), implying sustained deployment breadth.

Trend: *Increasing* [$H \uparrow$]. Per-incident severity rises between the two AIID periods and documented case surfacing roughly doubled between 2022 and 2023–2024, though both measures are floors rather than direct harm counts.

Confidence: *Low*. Tier 3 (partial expert elicitation required). AIID data sufficiency is moderate but match counts are flat, and no proxy provides a tight upper bound on wrongful-arrest incidence. The federal-search-volume ceiling (60,000 cumulative) and the documented-case floor (single-digit annual) span more than two orders of magnitude.

Exposure estimation, E: We approximate exposure by the share of US jurisdictions whose law-enforcement agencies deploy facial-recognition technology. Jurisdiction-level data on the use of facial recognition are publicly available up to 2020, when 20% (54 of 268 cities sampled) were found to use the technology (Johnson et al., 2024). Deployment is reported to be increasing rapidly, with one commercial application seeing police search volumes double between 2023 and 2024 (The Record, 2024). With no available post-2020 estimates, we apply three trajectories to the 2003–2020 observations. For the lower bound, we fit a linear trend to all observations (conservative, as early slow-growth years attenuate the slope). The point estimate extrapolates the straight line through the 2016 and 2020 observations. For the upper bound, we fit an exponential growth model (+16% per year), reflecting the accelerating adoption rate observed across survey waves. Under these assumptions, facial recognition deployment is estimated at 24% (19–27%) of jurisdictions in 2022, 25% (19–31%) in 2023, 27% (20–41%) in 2024, and 32% (25–56%) in 2025.

Trend: *Increasing* [$E \uparrow$]. Deployment share is rising; the rate of increase is uncertain and spans a linear-to-exponential range.

Confidence: *Low*. Tier 2 (derived from reasonable sources with missing upper bounds; post-2020 estimates are extrapolations).

Classification: *Escalating*. AIID harm count rises sharply between the two periods (six to 15–71) while the AIID match count is flat and jurisdiction-level facial recognition deployment is also rising; per-incident severity appears to be growing faster than exposure, yielding a rising harm-per-exposure trend [$\hat{H} \uparrow$]. Both bounds remain wide: the harm-side upper bound is weak in the absence of mandatory disclosure of facial recognition-use, so this classification should be read as directional rather than precise, and expert elicitation or a targeted disclosure mandate would tighten confidence substantially.

D.3. Case Study: Deepfake-enabled investment scams.

Monitoring question: [S] Among US individuals [O] exposed to online investment solicitations, [R] how many lose money to investment scams involving AI-generated deepfake media (video, image, or voice) [T] per calendar year?

Harm estimation, H: LLM filtering of the AIID returned 13 full matches in 2024 (harm count range 13–1,101, predominantly moderate severity) and 21 full matches in 2025 (harm count range 23–970,166, shifting toward severe severity), with data sufficiency rated high. The partial-to-full ratio is 14.32; inspection of the partial matches shows that most are deepfake incidents used for adjacent harm categories (romance scams, political impersonation, identity fraud) rather than investment scams specifically, reflecting the intentional narrowing of R rather than overspecification of S. For an upper-bound proxy, we use the FBI IC3 investment-fraud category, which by construction contains all deepfake-enabled investment scams: any deepfake investment scam is first an investment scam. IC3 reports investment-fraud losses of $3.31 billion in 2022, $4.57 billion in 2023, and $6.57 billion in 2024 (Federal Bureau of Investigation, 2023; 2024; 2025). Within the 2024 total, $5.8 billion across 41,557 complaints is the "pig butchering" cryptocurrency-investment subcategory that the FBI associates with synthetic-media tooling. Operation Level Up, launched in January 2024, notified 4,300 victims and preserved an estimated $285 million in its first year (Federal Bureau of Investigation, 2025).

Trend: *Increasing* [$H \uparrow$]. AIID full-match count rose 62% from 2024 to 2025 with a shift in severity distribution toward severe outcomes. Directional support is strong: IC3 total investment-fraud losses rose 44% and the pig-butchering subcategory rose 47% between 2023 and 2024.

Confidence: *Medium.* Tier 2 (category-adjacent proxy). The FBI IC3 is a large mandatory-intake complaint corpus with stable methodology; however, the deepfake-specific share within investment fraud is an FBI attribution rather than a directly measured split.

Exposure estimation, E: We approximate exposure by US adults with internet or phone access, the population reachable by AI-generated deepfake investment solicitations across email, messaging, and social-media channels. 96% of US adults use the internet, an estimate stable since 2023 (Pew Research Center, 2024). For the point estimate, we applied 96% to all US adults. For the upper bound, we additionally included adults reachable by telephone, using a telephone access rate of 99% (Chamber of Commerce, 2024). For the lower bound, we retained the 96% internet access assumption for adults aged 18–74, but reduced this to 50% for those aged 80 and older. We estimate 249 million adults are potentially exposed to deepfake investment solicitations (plausible range 241–257 million). Order-of-magnitude estimate: 10^8.

Trend: *Approximately flat* [$E \rightarrow$]. US internet use and phone ownership have been stable over the study period.

Confidence: *Medium.* Tier 2 (derived from reasonable sources).

Classification: *Escalating.* Harm is rising sharply ($H \uparrow$) while the exposed-adult denominator is approximately flat ($E \rightarrow$), yielding a rising harm-per-exposure trend [$\hat{H} \uparrow$]. The high partial-to-full match ratio (14.32) flags possible overspecification of R (§4.2): relaxing "investment scams" to a broader fraud category would capture adjacent deepfake harms such as romance, identity, and political-impersonation fraud.

D.4. Case Study: AI voice-clone impersonation fraud.

Monitoring question: [S] Among US adults [O] who receive phone or voice contact from unknown parties, [R] how many lose money to scams involving an AI-cloned voice impersonating a trusted person [T] per calendar year?

Harm estimation, H: LLM filtering of the AIID returned 11 full matches in the 2022–2023 period (harm count range 334–47,136) and 101 full matches in the 2024–2025 period (harm count range 20,918–4.0×10^6), an order-of-magnitude rise in both full-match count and reported-harm scale. The partial-to-full ratio is 3.93 and data sufficiency is high, indicating that the monitoring question is well specified for this harm type. For an upper-bound proxy, we use the FTC Consumer Sentinel Network's imposter-scam category, which by construction contains all AI voice-cloned scams: any voice-clone scam is first an imposter scam. Imposter-scam losses reported to the FTC totalled $2.6 billion in 2022 and $2.7 billion in 2023 (combined 2022–2023: $5.3 billion), rising to $2.95 billion in 2024 across 845,806 reports (Federal Trade Commission, 2025). The FBI IC3 provides a tighter AI-adjacent sub-bound through its 2025 Annual Report, which for the first time tracked AI as a complaint descriptor: approximately $5 million in reported losses to "distress scams" (voice-cloning impersonation of family members in apparent crisis) in 2025, within a total of $893 million in adjusted losses across all AI-enabled fraud types (Federal Bureau of Investigation, 2025).

Trend: *Increasing* [$H \uparrow$]. AIID full-match count rose roughly tenfold between the two periods, with directional support from the FTC imposter-scam proxy, which rose 9.3% from 2023 to 2024.

Confidence: *Medium.* Tier 2 (category-adjacent proxy). The FTC Consumer Sentinel Network and FBI IC3 are mandatory-intake consumer-fraud reporting streams with stable methodology; however, the AI-voice share within the imposter-scam category is not directly measured.

Exposure estimation, E: We approximate exposure by US adults reachable by phone or voice messaging in 2024. More than 99% of US adults have a mobile or landline telephone, a rate that has been stable for over a decade (Chamber of Commerce, 2024). For the point estimate, we applied this rate to the US adult population. For the upper bound we assumed 99.9% of US adults can be reached by phone or voice messaging; for the lower bound, we retained the 99% rate for adults aged 18–74 but reduced this to 50% for adults aged 75 and older. We estimate 257 million adults receive phone or voice contact in 2024 (plausible range 244–259 million). Order-of-magnitude estimate: 10^8.

Trend: *Approximately flat* [$E \rightarrow$]. US phone ownership has been stable over the study period.

Confidence: *Medium.* Tier 2 (derived from reasonable sources).

Classification: ***Escalating.*** Harm is rising sharply ($H \uparrow$) while the exposure among adults is approximately flat ($E \rightarrow$), yielding a rising harm-per-exposure trend [$\dot{H} \uparrow$]. Urgent response is warranted: expanded monitoring, authentication countermeasures at voice-contact endpoints, and regulatory attention to consumer-facing disclosure and detection.