
AI Risk Hotspots: A Data-Driven Early Warning System

Ismail Khan Mohammad
Independent Researcher

With
Apart Research

Abstract

The discourse on AI risk is often split between tracking abstract capabilities and reacting to anecdotal harms. This project bridges that gap by creating a data-driven **Situational Awareness System** for AI-related incidents. I correlate historical incident data from the AI Incident Database with training compute data from Epoch AI, demonstrating a strong relationship between the growth in AI power and the rise of real-world harm.

My system features three core modules: a "Risk Radar" for historical trend analysis, a "Cause & Effect Correlator" visualizing the macro trend, and a "Dynamic Forecaster" that identifies and projects the trajectory of the fastest-accelerating harm categories. The dashboard identifies '**Malicious Use & Security**' as the top accelerating hotspot. Based on a micro-analysis of this hotspot, I propose concrete mitigation strategies, including the adoption of C2PA watermarking and stricter API access controls for high-risk capabilities. This project offers a novel, empirical methodology for tracking and forecasting the tangible, societal consequences of AI progress.

Keywords: AI safety, risk forecasting, incident analysis, AI governance, situational awareness

1. Introduction

Research Question: Can we empirically track the real-world harms caused by accelerating AI progress and build a system to forecast which categories of harm are emerging most rapidly?

Hypothesis: There is a quantifiable correlation between the exponential growth in AI training compute (the "cause") and the frequency of publicly reported AI incidents (the "effect"). By analyzing the historical trends of different harm categories, we can identify "hotspots" and project their future trajectory, creating an early warning system for policymakers and AI labs.

Contribution: This project contributes to the field of AI forecasting by shifting the focus from abstract capability timelines to the tangible, societal impact of currently deployed systems. While organizations like METR track the emergence of future capabilities, my system tracks the real-world consequences. It provides a quantitative, bird's-eye view of the AI harm landscape, moving beyond anecdotal evidence. This data-driven approach aligns with geopolitical frameworks like Kai-Fu Lee's 'AI Superpowers,' providing empirical evidence for the real-world harms associated with the different 'waves' of AI adoption. Its potential impact is to provide policymakers with a data-driven tool to prioritize regulatory efforts and to help AI labs identify which safety and mitigation techniques are most urgently needed. My threat model considers both accidental misuse (e.g., algorithmic bias) and intentional misuse (e.g., deepfake fraud) of publicly and commercially available AI systems.

2. Methods

My methodology involved a multi-stage data processing and analysis pipeline:

1. **Data Sourcing:** I used two primary datasets: the AI Incident Database (incidentdatabase.ai) for incident reports and the "Notable AI Models" training compute dataset from Epoch AI (<https://epoch.ai/data/ai-models>).
2. **Harm Classification:** I developed a taxonomy of 9 high-level harm categories and used a hybrid classification strategy. An initial "gold standard" set was created by programmatically mapping structured risk tags from the source data. The remaining incidents were classified using a facebook/bart-large-mnli zero-shot classification model.
3. **Time-Series Analysis:** The classified incidents were aggregated by quarter. Training compute data was aggregated by year.
4. **Forecasting Model:** A Simple Linear Regression model was fitted to the most recent three years of data for each harm category to generate a 3-year forecast. This "recency bias" ensures the forecast is responsive to the latest trends.
5. **Dashboard Development:** An interactive dashboard was built using Streamlit and Plotly, running in a reproducible Docker environment.

All code is available at my public GitHub repository: <https://github.com/ismailm/aI-risk-hotspots-early-warning-system>

3. Results

My analysis produced a fully interactive dashboard with three core modules.

Figure 1: The Risk Radar shows the historical frequency of incidents across all 9 harm categories. While isolated incidents date back as far as 1983, the chart reveals a dramatic and sustained acceleration in the frequency and variety of incidents beginning around 2020.

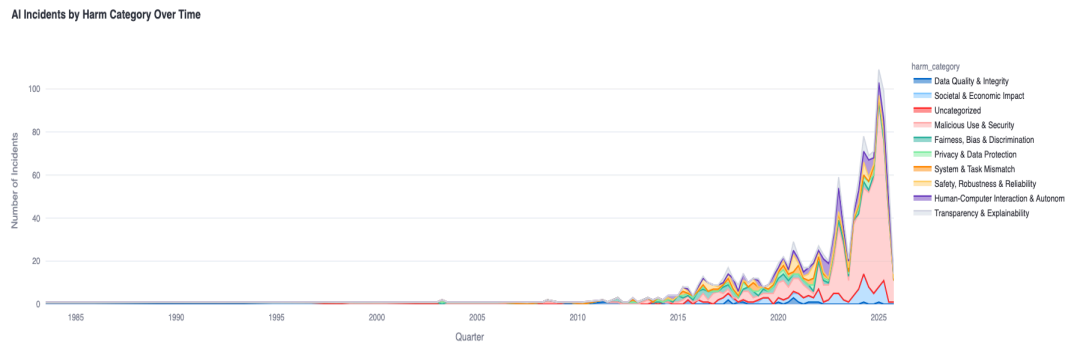


Figure 2: The Cause & Effect Correlator visualizes my core finding, showing a strong correlation between the rise in total training compute and the total number of incidents per year.

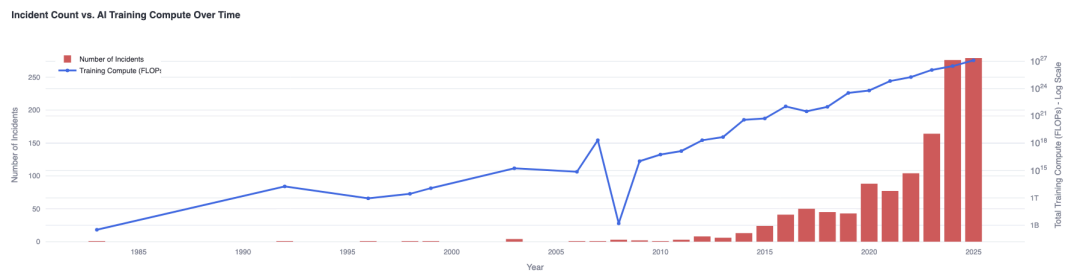
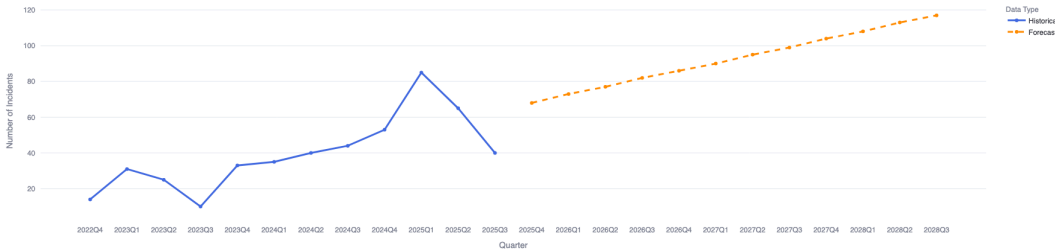


Figure 3: The Dynamic Forecaster identifies 'Malicious Use & Security' as the top accelerating hotspot. Based on the strong upward trend in the most recent three years of data, the model projects a significant increase in these incidents over the next three years. The forecast (dashed orange line) begins after the last complete historical quarter, a methodological choice to ensure the trend is not skewed by partial, recent data.



4. Discussion and Conclusion

My results confirm the primary hypothesis: the growth in AI power is strongly correlated with a rise in real-world harm. The dashboard serves as an effective "Situational Awareness System," providing a much-needed empirical overview of the risk landscape.

The identification of 'Malicious Use & Security' as the top hotspot is a critical, actionable insight. A micro-analysis of recent incidents in this category revealed that the trend is largely driven by a few specific use cases, namely generative impersonation and automated social engineering.

Based on this data-driven finding, I propose the following concrete mitigation strategies, summarized in the table below:

Identified Theme	Specific Harms (Use Cases)	Proposed Mitigation
Generative Impersonation	Voice/video cloning for fraud, deepfake misinformation, non-consensual synthetic media.	C2PA Watermarking: Embed cryptographic provenance data in all generated media to enable automated detection.
Automated Social Engineering	LLMs creating personalized phishing attacks, text-based scams, and fraudulent content at scale.	Targeted Refusal Training: Specifically train models to detect and refuse prompts aimed at

		creating deceptive or fraudulent content.
High-Risk Capability Access	Anonymous actors using powerful APIs (e.g., voice cloning) for malicious purposes.	Stricter API Access Controls: Implement "Know Your Customer" (KYC) protocols for high-risk AI capabilities.

These mitigation strategies are essential first steps. The Adoption of **C2PA Watermarking** would allow platforms to automatically flag synthetic media, giving users a crucial layer of defense. **Stricter API Access Controls (KYC)** for high-risk capabilities like voice cloning would move the ecosystem away from anonymous, untraceable access, making it harder for malicious actors to operate. Finally, **Targeted Refusal Training** represents a necessary evolution in model safety, focusing on the intent behind a prompt, not just its literal content.

Furthermore, the dynamic forecaster reveals that not all risks behave alike. While **'Malicious Use & Security'** is a clear, accelerating hotspot, other categories like **'System & Task Mismatch'** represent more chronic, persistent problems. This suggests that a one-size-fits-all approach to AI safety is insufficient; we need both urgent responses for hotspots and long-term, systemic strategies for chronic risks.

This project demonstrates that it is not only possible but necessary to systematically track and forecast the tangible harms of AI, providing a critical tool for effective governance and preparedness.

5. References

- AI Incident Database. (2025). CSET: AI Incident Database. Retrieved from <https://incidentdatabase.ai/>
- Bhatia, K., & Deshpande, A. (2024). AI governance and explainable AI to navigate the pitfalls of generative AI. Springer Link. <https://link.springer.com/article/10.1007/s00146-024-02130-8>
- Epoch AI. (2025). Data on Notable AI Models. Retrieved from <https://epoch.ai/data/ai-models>
- Mökander, J., & Schuett, J. (2025). AI and misinformation: A systematic review. Springer Link. <https://link.springer.com/article/10.1007/s00146-025-02620-3>
- Stanford University Ethics, Society, and Technology Hub. (n.d.). Towards a Framework for Risk Mitigation of Potential Misuse of Artificial Intelligence in Biomedical Research. Retrieved from <https://esr.stanford.edu/document/towards-a-framework-for-risk-mitigation-of-potential-misuse-of-artificial-intelligence-in-biomedical-research/>
- Vanta. (n.d.). AI security best practices. Retrieved from

6. Appendix

Security Considerations

A primary limitation of my system is its dependence on the AI Incident Database. As a public, crowd-sourced database, it may suffer from reporting bias (some incidents are more likely to be reported than others) and a natural time lag between an event and its publication.

My forecasting model is a simple linear regression and does not account for seasonality or complex non-linear trends. A future improvement would be to implement more sophisticated time-series models (e.g., ARIMA or Prophet).

Finally, the classification system, while robust, is not perfect. A future improvement would be to implement a formal validation script to calculate a precise accuracy score for the LLM classifier against my "gold standard" set and to explore few-shot prompting to further increase its accuracy.