

AI Risk Hotspots: A Data-Driven Early Warning System

Ismail Khan Mohammad

Independent Researcher

Apart Research AI Forecasting Hackathon 2025

The Problem: We're Tracking the Cause, Not the Effect

THE CAUSE (Inputs)

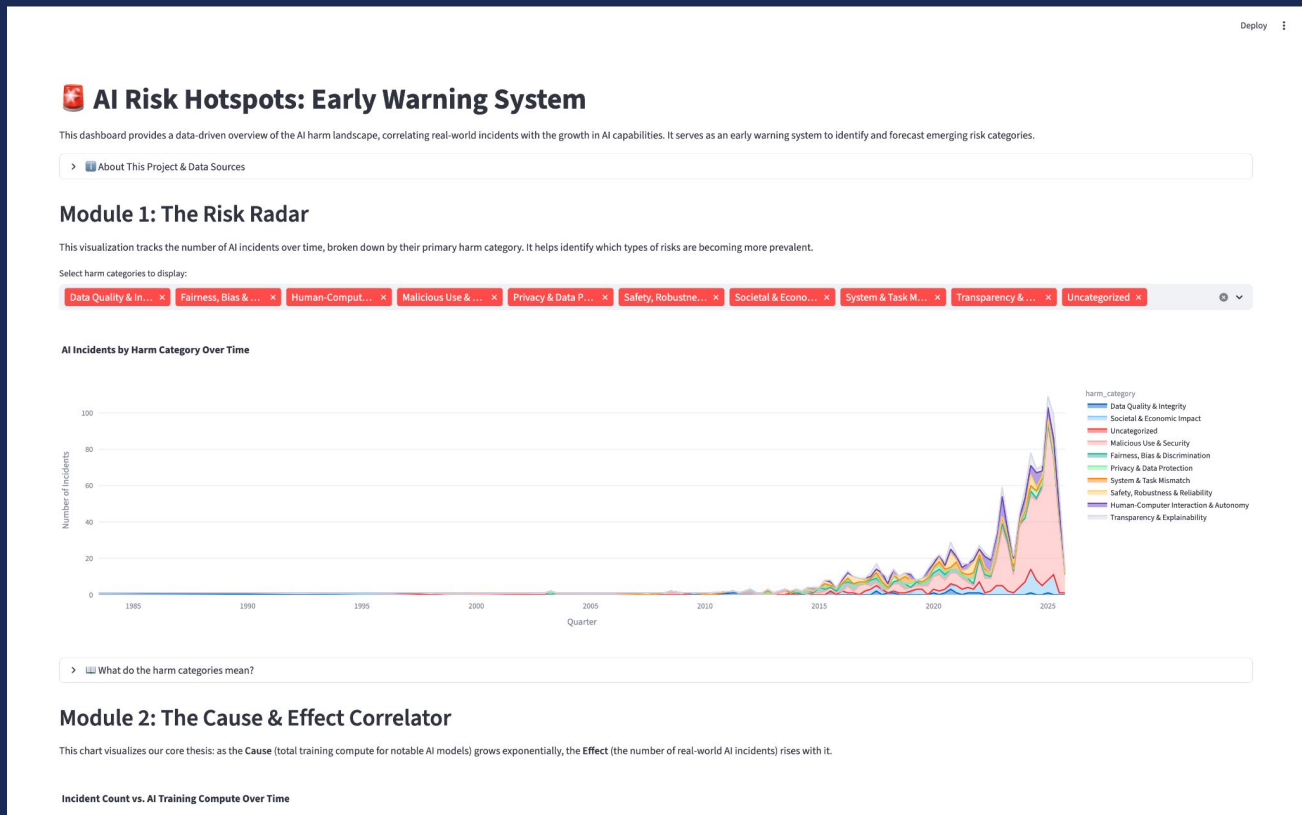
- AI Training Compute (Source: Epoch AI)
- Emerging Capabilities (Source: METR)



THE EFFECT (Outputs)

- Real-World Harms
- Societal Impact
- Reported Incidents

Our Solution: A Situational Awareness System for AI Risk

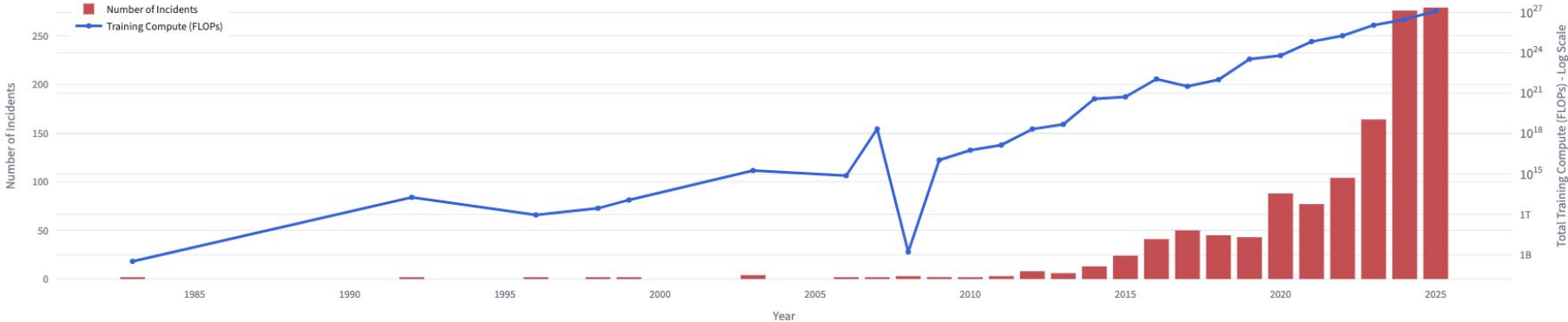


Finding 1: As AI Power Grows, So Does AI Harm

Module 2: The Cause & Effect Correlator

This chart visualizes our core thesis: as the Cause (total training compute for notable AI models) grows exponentially, the Effect (the number of real-world AI incidents) rises with it.

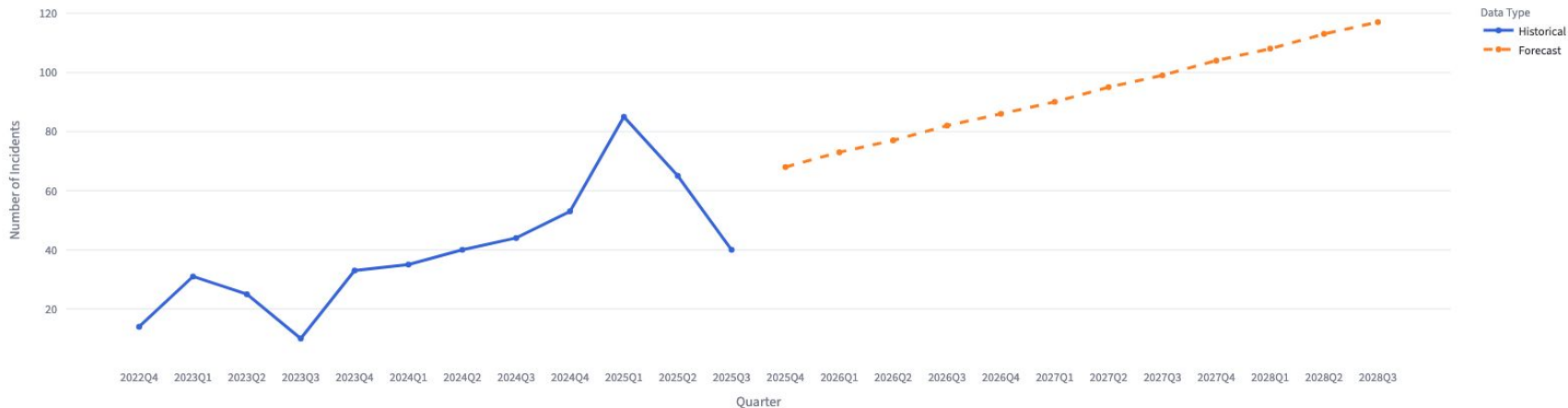
Incident Count vs. AI Training Compute Over Time



Finding 2: 'Malicious Use & Security' is the #1 Risk Hotspot

Incident Trajectory for: Malicious Use & Security

Definition: Incidents where AI was intentionally used for harmful purposes (e.g., scams, deepfakes) or where an AI system was compromised by a security vulnerability.



Feature: From "What" to "Why" with 'On the Radar'

▼ On the Radar: See Latest Events for 'Malicious Use & Security'

Found 219 relevant articles. Displaying the Top 9.

AI makes phishing 4.5x more effective, Microsoft says

Published: 16-10-2025

destruction criminal crime

Scammers now impersonate coworkers, steal email threads in convincing phishing attacks

Published: 08-10-2025

danger

Former Google CEO warns AI systems can be hacked to become extremely dangerous weapons

Published: 16-10-2025

danger death

AI girlfriend apps leak millions of private chats

Published: 20-10-2025

harassment

Beware fake credit card account restriction scams

Published: 27-10-2025

criminal

Don't fall for fake settlement sites that steal your data

Published: 13-10-2025

danger criminal

Scammers target retirees as major 401(k) rule changes loom for 2026 tax year ahead nationwide

Published: 28-10-2025

criminal

How to spot and stop AI phishing scams

Published: 13-10-2025

Jeep and Chrysler parent Stellantis confirms data breach

Published: 06-10-2025

criminal

Next Steps:

1. **Review the articles above.** If you find a relevant research paper mentioned, note its title.
2. **Find the primary source** by searching for the paper's title on [arXiv.org](https://arxiv.org).

News sourced from the AI Incident Database News Digest. [Click here to view the full, live feed.](#)

Analysis: What's Driving the 'Malicious Use & Security' Hotspot?



Deepfakes & Synthetic Media Impersonation

AI-generated videos, voices, and images used to impersonate individuals for scams, promotions, or misinformation.



AI-Enabled Fraud & Phishing

AI-generated text and websites used to steal credentials, manipulate individuals, and create fraudulent campaigns at scale.



Unsafe Exposure & Algorithmic Harm

AI systems acting in unintended ways, such as promoting harmful content, making biased decisions, or producing flawed risk scores.

Recommendations: From Data-Driven Insight to Action

The Problem (Identified Use Case)

Generative Impersonation

Voice/video cloning for fraud and misinformation

Automated Social Engineering

LLMs creating convincing phishing attacks at scale

Lab-Based Mitigation (Technical & Policy)

Adopt C2PA Watermarking

Embed cryptographic, tamper-evident watermarks into all generated media to prove its synthetic origin.

Targeted Refusal Training

Specifically train models to detect and refuse prompts aimed at creating deceptive or socially manipulative content.

Governance Mitigation (Regulatory)

Mandate "Know Your Customer" (KYC)

Require providers of high-risk APIs (e.g., voice cloning) to verify user identity, preventing anonymous abuse.

Clarify Platform Accountability

Establish clear legal liability for platforms that knowingly fail to prevent their services from being used to generate fraudulent content at scale.

Future Work & Next Steps

- **Formal Validation & Accuracy Benchmarking**
 - Implement a script to formally validate our LLM classifier against the "gold standard" set, generating a precise accuracy score and a full classification report.
- **Advanced Forecasting Models**
 - Explore more sophisticated time-series models (e.g., ARIMA or Prophet) to capture more complex, non-linear trends and seasonality in the incident data.
- **Automated Micro-Analysis**
 - Integrate a second, more powerful LLM (e.g., via a GPT or Gemini API) to automate the thematic analysis of hotspots, turning the final "Recommendations" section into a dynamic, real-time feature of the dashboard.

AI Risk Hotspots: Early Warning System

This dashboard provides a data-driven overview of the AI harm landscape, correlating real-world incidents with the growth in AI capabilities. It serves as an early warning system to identify and forecast emerging risk categories.

📄 View This Project & Data Sources

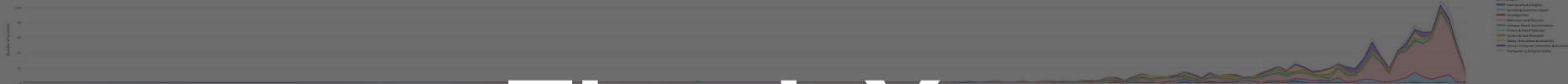
Module 1: The Risk Radar

This visualization tracks the number of AI incidents over time, broken down by their primary harm category. It helps identify which types of risks are becoming more prevalent.

View harm categories in detail:

[AI Incidents & Risks](#) [Autonomous Use & Security](#) [Content Generation](#) [Cybersecurity](#) [Deepfakes](#) [Disinformation](#) [Privacy & Data](#) [Research & Development](#) [Safety & Security](#) [Systemic & Societal](#) [Talent & Work](#) [Transparency & Accountability](#) [Unintended Consequences](#)

AI Incidents by Harm Category Over Time

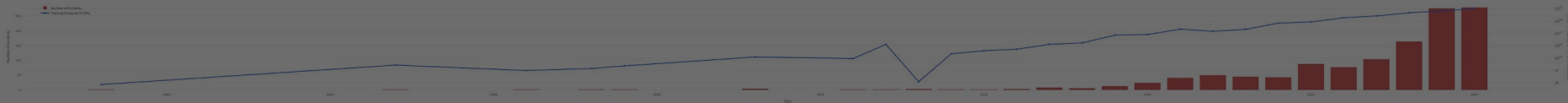


🔍 What do the harm categories represent?

Module 2: The Cause & Effect Correlator

This chart visualizes our core thesis: as the Capex \$Billion training compute for notable AI models grows exponentially, the Effect (the number of real-world AI incidents) rises with it.

Incident Count vs. AI Training Compute Over Time



Module 3: The Dynamic Forecaster

This module provides a 3-year forecast for any selected harm category. Categories are sorted by their recent growth trend, with the top 'hotspot' selected by default.

📄 View Forecasting Methodology

Select harm categories to forecast:

Malicious Use & Security

Incident Trajectory for: Malicious Use & Security

Definition: Incidents where AI was intentionally used for harmful purposes (e.g., scams, fraud, identity theft) or where an AI system was compromised by a security vulnerability.



🔍 Go to the Radar: View Latest Events for 'Malicious Use & Security'

Key Insights & Takeaways

Based on the data and forecasts presented in this dashboard, we can draw several key conclusions:

- The Top Warning:** Our system identifies "Malicious Use & Security" as the AI accelerating risk category. The forecast projects a significant increase in these incidents, indicating an urgent need for prioritized research into targeted mitigation strategies.
- The Means-Ends Chain:** There is a strong visual correlation between the exponential growth in AI training compute (the "Cause") and the overall rising number of real-world AI incidents across all categories (the "Effect").
- Not All Risks Behave Alike:** The forecast reveals that different harm categories have different trajectories. While "Malicious Use & Security" is a clear, accelerating hotspot, other categories like "Human-Computer Interaction & Autonomy" represent more chronic, persistent problems, requiring different safety strategies.

Thank You

Live Demo: <https://ai-risk-hotspots.streamlit.app/>

GitHub Repository: <https://github.com/ismailmk/al-risk-hotspots-early-warning-system>