MILEVA SECURITY LABS

# ESTIMATING AI SECURITY INCIDENT LIKELIHOOD

USING **OPEN-SOURCE INCIDENT DATA** TO INFORM A **VULNERABILITY-ORIENTED APPROACH** TO RISK ASSESS OF AI COMPONENTS

FORESIGHT INSTITUTE
Australian National University
UNSW SYDNEY

# EXECUTIVE SUMMARY

An AI security incident refers to an event where a threat actor - human or AI - exploits or attacks an AI-enabled system, resulting in loss, harm, or compromise. These incidents may stem from vulnerabilities shared with traditional IT systems, or from AI-specific weaknesses such as adversarial manipulation, data poisoning, or prompt injection. While academic research and emerging government guidance have highlighted these risks, there has been little systematic analysis of how they occur "in the wild."

This study, conducted by Mileva Security Labs in collaboration with the Australian National University (ANU) and the University of New South Wales (UNSW), was generously supported by the Foresight Institute. It builds on an interim report released in December 2024 and represents the final phase of a multi-part investigation into how publicly available data can be used to estimate the likelihood of AI security incidents.

## OUR OBJECTIVES WERE TO:

**Assess** the prevalence of AI security incidents relative to AI safety and general cybersecurity events;

**Identify** the causes and contributing factors behind these incidents;

**Explore** how future likelihood of AI security events might be modelled from available data.

## METHODOLOGY OVERVIEW

**The research combined three complementary analyses:**

**1.**

**Vulnerability Analysis (CVE Data):**

Examined 18,779 software vulnerabilities from MITRE's CVE database (2020-2025) to identify and categorise AI-related entries.

**2.**

**Incident Analysis:**

Analysed and labelled over 1,000 entries across four open-source AI incident databases (AIAAIC, AIID, ATLAS, AIVD), identifying a subset of 96 confirmed AI security cases.

**3.**

**Framework Evaluation:**

Compared five leading AI risk management frameworks-the EU AI Act, NIST AI RMF, Microsoft GSAIP, ISO/IEC 42001, and NCSC AI Guidelines-using a Normalised Coverage-Depth Score (CDS) to assess how comprehensively they address AI security risks.

## KEY **FINDINGS**

- AI-specific vulnerabilities are underrepresented in public databases, with fewer than 1% of CVEs explicitly linked to AI systems. Most entries classify as general IT issues, masking potential AI-related causes.
- AI security incidents do occur, but are often misclassified or conflated with safety or misuse cases. Public datasets emphasise accidental or ethical issues over technical exploitation.
- Framework coverage is uneven. NIST's AI RMF provided the strongest balance of breadth and depth, while other frameworks-such as ISO/IEC 42001 and NCSC guidelines-offered less technical specificity.
- Data limitations persist. Lack of standardised terminology, inconsistent tagging, and limited financial loss data hinder quantitative modelling of AI security risk.

## RECOMMENDATIONS

1. Standardise definitions and reporting: Establish consistent terminology to distinguish AI security from AI safety and general cybersecurity.
2. Integrate AI vulnerabilities into cybersecurity processes: Expand public vulnerability repositories (e.g., CVE) and scoring systems (e.g., CVSS) to include AI-specific descriptors.
3. Strengthen AI supply chain assurance: Secure the datasets, pre-trained models, and third-party components that underpin AI systems.
4. Encourage collaboration: Facilitate joint data-sharing initiatives between government, academia, and industry to improve evidence-based policymaking.
5. Enhance education and capacity: Upskill practitioners in AI security principles and encourage inclusion of AI assurance functions within cybersecurity teams.

## CONCLUSION

This research represents one of the first empirical efforts to quantify and characterise AI security incidents using open-source data. It reveals that while AI security events are still rarely reported, they differ fundamentally from traditional IT vulnerabilities and require distinct mitigation strategies. Building a robust evidence base for AI incident likelihood will be critical to developing trustworthy, resilient AI systems in both government and industry.

# PROJECT INFORMATION

This project was generously funded by the Foresight institute, and carried out as a collaboration led by Mileva Security Labs in partnership with the Australian National University (ANU) and University of New South Wales (UNSW).

## PROJECT PARTNERS

### FORESIGHT INSTITUTE

The Foresight Institute is a non-profit research organisation based in California, USA, that supports the development of technologies for the long-term benefit of humanity. Founded in 1986, it brings together researchers, policymakers, and innovators to explore areas such as artificial intelligence, biotechnology, and molecular nanotechnology.

### MILEVA SECURITY LABS

Mileva Security Labs is an Australian research and advisory company dedicated to the security and governance of artificial intelligence systems. Founded by Harriet Farlow, Mileva bridges academic research and practical application, helping organisations understand, measure, and mitigate risks in AI through research, product development, and training.

### AUSTRALIAN NATIONAL UNIVERSITY

The Australian National University (ANU) is a world-leading centre for research and education based in Canberra, Australia. Established by the Australian Government in 1946, ANU is renowned for its contributions to science, policy, and national security research. Its interdisciplinary approach and close collaboration with government and industry make it a key institution in advancing the responsible development and governance of emerging technologies, including artificial intelligence.

### UNIVERSITY OF NEW SOUTH WALES

The University of New South Wales (UNSW) is a leading Australian research university based in Sydney, recognised for its strong focus on innovation, technology, and industry collaboration. Founded in 1949, UNSW is home to world-class research in engineering, computer science, and artificial intelligence, with a mission to translate academic excellence into practical societal impact.

# PROJECT **TEAM**

## HARRIET FARLOW
### PROJECT LEAD

Harriet Farlow is the founder of Mileva Security Labs and her PhD is in adversarial machine learning. She has worked at the intersection of AI and security for a decade, in consulting, tech start-ups, and the Australia Government. Harriet's work bridges technical research and policy, and she aims to empower organisations to navigate the complex landscape of AI security. She has spoken at DEF CON and other leading forums, advocating for practical and scalable AI risk solutions.

## TANIA SADHANI

Tania Sadhani is an AI security researcher at Mileva Security Labs and an Honours student in Machine Learning at ANU. With a strong focus on adversarial threats and AI misuse, Tania contributes to cutting-edge research on AI risk and security frameworks. She is passionate about advancing methodologies that ensure AI systems are both safe and resilient.

## DELIA SCHULTZ

Delia Schultz is an AI Cybersecurity Intern at Mileva Security Labs, where she contributes to research on adversarial machine learning and the security of AI systems. She is currently completing a Bachelor's degree in Data Science and Statistics at the University of California, Santa Barbara, where her studies focus on computational modelling and data-driven risk analysis. Delia has also volunteered as a Peer Counselor at Stanford University, developing skills in communication and problem-solving, and as a Precinct Captain with the ACLU, promoting civic engagement.

## SETH LAZAR

Professor Seth Lazar is a philosopher and researcher based at the Australian National University (ANU), where he leads work on the ethics and governance of artificial intelligence. His research explores how AI systems should be designed and deployed to align with democratic values, accountability, and human rights. He is the founder of the Machine Intelligence and Normative Theory (MINT) Lab and co-leads international initiatives on AI governance and responsible innovation.

## TIM LYNAR

Tim Lynar is a cybersecurity researcher and practitioner specialising in the governance, assurance, and resilience of emerging technologies. He is a Research Fellow at the University of New South Wales, where his work focuses on the security implications of artificial intelligence and its integration into critical systems. Tim brings extensive experience from both academic and operational cybersecurity environments, contributing to projects that bridge policy, risk management, and technical assurance.

# RELATED WORK

Artificial intelligence (AI) is now integrated across most sectors - from financial services and healthcare to national security and public administration - yet its security risks remain poorly defined and inconsistently managed. While cybersecurity traditionally focuses on protecting networks, systems, and data, AI security concerns the protection of AI models and components themselves: ensuring they behave as intended and cannot be manipulated, poisoned, or deceived.

Despite growing awareness, there is still limited evidence on how often AI security incidents occur and what causes them. Most existing studies focus on hypothetical or laboratory-based scenarios rather than real-world data. When incidents do happen, they are often misclassified or grouped with unrelated IT vulnerabilities, leaving organisations without a clear picture of the likelihood or impact of AI-specific threats.

Both academic and industry communities have begun to recognise the distinct nature of AI security. Researchers have proposed taxonomies for adversarial attacks - such as data poisoning, model evasion, and model inversion - and have explored how vulnerabilities arise at each stage of the AI lifecycle. Work by MITRE, Microsoft, and OWASP has shown that many of these threats exploit features unique to AI systems, such as dependency on training data and adaptive model behaviour. The OWASP Top 10 for LLM Applications, for instance, highlights threats such as poisoned training data, tampered pre-trained models, and insecure plugin ecosystems, while MITRE's Adversarial Threat Landscape for AI Systems (ATLAS) documents adversarial tactics that target AI behaviour directly. Academic research has also started to explore how to structure these risks. Frameworks such as the 3D model (Disrupt, Deceive, Disclose) map AI threats to the traditional security principles of availability, integrity, and confidentiality. Others propose merging AI-specific and traditional risk taxonomies using established terms such as vulnerability, exploit, exposure, and hazard to build consistent definitions.

SECURITY RISKS REMAIN **POORLY DEFINED** AND **INCONSISTENTLY MANAGED**

This report addresses that gap by **analysing publicly available incident databases** to better estimate the likelihood of AI security incidents, and by identifying where current frameworks, taxonomies, and reporting practices fall short.

MANY OF THESE THREATS EXPLOIT FEATURES UNIQUE TO AI SYSTEMS, SUCH AS **DEPENDENCY ON TRAINING DATA** AND **ADAPTIVE MODEL BEHAVIOUR**.

At the policy and governance level, several major frameworks attempt to systematise AI risk management. The EU AI Act represents the first comprehensive legislative framework for managing AI risks, classifying systems by risk level. In the United States, the NIST AI Risk Management Framework (AI RMF) provides voluntary guidance based on four pillars - Govern, Map, Measure, and Manage - while the international ISO/IEC 42001 standard defines structured controls for AI management systems. In Australia and the UK, agencies such as the NCSC and the Department of Industry, Science and Resources have published emerging AI security guidance focused on national security and critical infrastructure.

SEVERAL MAJOR FRAMEWORKS ATTEMPT TO **SYSTEMATISE AI RISK MANAGEMENT.**

However, most of these frameworks remain qualitative and governance-oriented, rather than technical or quantitative. They outline processes for responsible AI but do not assess the likelihood or frequency of specific AI security incidents. Similarly, vulnerability databases such as the Common Vulnerabilities and Exposures (CVE) list include over 289,000 entries but do not classify or tag AI-specific incidents.

MOST OF THESE FRAMEWORKS REMAIN QUALITATIVE AND GOVERNANCE-ORIENTED, RATHER THAN **TECHNICAL OR QUANTITATIVE.**

Recent academic studies have called for AI and ML vulnerabilities to be formally integrated into these public databases, identifying two main categories of weaknesses: algorithmic flaws, which stem from weaknesses in the underlying model architecture, and model-specific flaws, which can lead to data leakage or adversarial manipulation. Both demonstrate that AI models themselves can be the root cause of a security failure - something not captured in current cybersecurity taxonomies.

This report builds on that emerging body of work and contributes new analysis by applying a **vulnerability-oriented, data-driven approach** to estimate the likelihood of AI security incidents. By consolidating open-source data from AI incident repositories and vulnerability databases, it aims to **clarify how these incidents are currently categorised, where reporting gaps exist, and how quantitative methods can be used to strengthen AI risk management in both industry and government contexts.**

# METHODOLOGY

This research used a three-part approach to understand and estimate the likelihood of AI security incidents. Each part focused on a distinct layer of the AI security landscape - vulnerabilities, incidents, and frameworks - allowing us to connect real-world data with practical risk management guidance.

**1.**

**PART 1**
**CVE Analysis:**
Examined existing software vulnerability data to understand how AI-related flaws are currently recorded and categorised.

**2.**

**PART 2**
**AI Security Incident Analysis:**
Analysed documented AI incidents to identify the types of security failures that occur in practice and the factors contributing to them.

**3.**

**PART 3**
**AI Security Framework Analysis:**
Evaluated whether leading governance frameworks adequately address these risks, using a structured coverage and depth assessment.

Together, these parts form a continuous line of inquiry: from how vulnerabilities are recorded, to how incidents manifest, to how existing frameworks respond - providing a foundation for future quantitative models of AI security likelihood.

## PART 1: CVE ANALYSIS

**Objective**
To determine how AI-related vulnerabilities are currently represented within traditional cybersecurity data and to create a foundation for AI-specific vulnerability categorisation.

**Data Collection and Processing**
Data was drawn from the MITRE CVE repository, which contains over 289,000 entries (1999-2025). For relevance, the period 2020-2025 was selected, yielding 18,779 records. Each record includes a CVE identifier, description, severity score, and publication date.
Data was processed in Visual Studio Code, cleaned, standardised, and converted into a searchable format. Natural language preprocessing removed formatting inconsistencies and converted text to lowercase for consistency.

**Model Design and Classification**
**The CVE analysis used a hybrid model combining semantic vector search and rule-based decision logic:**

**STEP 1 - Embedding & Semantic Search**

75 labelled examples per category were embedded using OpenAI's embedding model and stored in ChromaDB. A semantic search retrieved similar descriptions based on conceptual similarity, not just keywords.

**STEP 2 - Binary Decision Logic**

Each CVE was then passed through a binary decision tree that applied structured rules.
The model first determined if the vulnerability referenced AI components (e.g., models, datasets, LLMs). If so, it assessed whether the flaw originated in the model logic itself (classified as AI Security), the supporting tools or datasets (AI Supply Chain), or broader IT dependencies (AI IT Cybersecurity).

**The output produced four standardised categories:**

1. **AI SECURITY -** vulnerabilities directly rooted in model behaviour or adversarial manipulation.
2. **AI SUPPLY CHAIN -** weaknesses in datasets, pre-trained models, or third-party components.
3. **AI IT CYBERSECURITY -** traditional IT vulnerabilities within AI systems.
4. **GENERAL CYBERSECURITY -** unrelated vulnerabilities.

The resulting dataset provided the foundation for cross-comparison with incident and framework data in later stages.

## PART 2: AI SECURITY INCIDENT ANALYS

**Objective**
To identify, classify, and analyse publicly reported AI security incidents and use them to infer qualitative insights about frequency, type, and contributing factors.

**Data Sources**
Incidents were consolidated from four open databases:

- The AI Incident Database (AIID): 443
- MITRE's Adversarial Threat Landscape for AI Systems (ATLAS): 26
- The AI Vulnerability Database (AIVD): 48
- The AI, Algorithmic and Automation Incidents and Controversies (AIAAIC) repository

**Our process was twofold:**

**STEP 1 - Developing a smaller, specialised database for case study analysis.**

AIID, ATLAS, and AIVD were merged and manually reviewed in June 2024, producing a subset of 96 AI security incidents. We defined a strict screening criteria to limit the number of incidents to analyse to enable further initial cyber risk modelling and qualitative analysis of particular incidents of interest, and initial quantitative analysis that involved manually imputing and applying cybersecurity frameworks.

**STEP 2 - Analysing the broader data sources for cyber risk statistical modelling**

Then, additional analysis of the AIAAIC (1964 incidents) and AIID (1046 incidents) was performed in March 2025.

**Stage 1: Building the case study database**

**The process of merging theses databases involved:**

1. Aligning the taxonomies of these databases.
2. Undergoing a screening phase where incidents were included based on a criterion of:
   a. Threat actor present - a security incident rather than a safety incident
   b. Involved an AI failure mode - a security incident where the harmed party were users of the AI system rather than an actor using AI to cause harm.
3. Designing additional features to impute, informed by our literature analysis, then using the reports to manually enrich each incident, alongside confidence levels.
4. Quality control by randomly sampling and reviewing incidents with 2 other SMEs.

We imputed the following features and recorded our confidence levels. While some features could not be directly verified due to inconsistencies or gaps in the source data, we filled in missing information based on our best knowledge to enable further analysis, acknowledging this as a limitation of the lack of data on public AI security incidents and reports in general:

- As different sources included different data, missing inputs for general incident data and source information were imputed.
  - Importantly, this included attributes of the target and actor such as size and sector.
- Information about the targeted AI system, including EU AI Risk Category, Sheridan's Human-in-the-Loop Level and Features from the Taxonomy for the European AI Ecosystem
- The attack was categorised using the NIST's Adversarial Machine Learning publication, MITRE ATLAS, CVSS, and additional categories we hypothesised would be interesting such as ML Lifecycle, Compute requirements.
- Upper and lower confidence bounds on financial impact (loss of operational cost, fines and/or recovery cost) and loss on confidentiality.

### Stage 2: Analysing the broader database

In June 2025, we examined 3,010 incidents recorded in AIAAIC and AIID. Recognising potential biases in these data sources, we began with data exploration using existing metadata and basic natural language processing to address two questions: (1) what types of incidents were being captured (e.g., security vs. safety, types of impacts), and (2) whether these incidents could inform cyber risk analysis (e.g., updating priors on AI security likelihood or applying standard Monte Carlo modelling used in cybersecurity). To support these analyses, we also tried using Anthropic's 3.5 model to automatically label incident types and infer additional contextual attributes.

#### Classification and Analysis

In addition to the categories imputed during the manual labelling, incidents were categorised by:

- **Threat Direction:** whether an active threat actor was present.
- **AI Involvement:** Whether the incident involved the misuse of AI (where AI was used to cause the harm), a failure of AI (where harm occurred because an AI component failed), or a mixture of both (there were cases where .
- **Actor Operational Capability:** The sophistication level of the threat actor following the categorisation offered by the RAND report.

To validate the automatic labelling, 100 random incidents were sampled along with looking at incidents from categorisations of interest. The accuracy of the chosen model (Haiku 3.5) was quite low; through the validation we found that Actor Operational Capability was consistently over-estimated even after further prompt injection. As a result, even though the AI labelling provided a useful way to foreground specific incidents, we refer back to the manually imputed set of incidents for further analysis.

#### Limitations

Public incident data is inherently biased toward visible or media-worthy events, creating sampling and reporting bias. Given these limitations, the analysis focused on building a qualitative evidence base to map incident characteristics, causes, and gaps in reporting - forming the groundwork for future probabilistic models of incident likelihood.

# PART 2: AI SECURITY FRAMEWORK ANALYSIS

**Objective**
To assess how well leading AI governance and risk frameworks address AI-specific security risks identified in Parts 1 and 2.

**Frameworks Analysed**
Five major frameworks were selected:

- EU AI Act
- NIST AI Risk Management Framework (AI RMF)
- Microsoft Guide for Securing the AI-Powered Enterprise (GSAIP)
- ISO/IEC 42001
- UK NCSC AI Security Guidelines

**Process**
Each framework's controls were mapped against a dataset of 150 identified risks and 118 standardised mitigations, drawn from sources such as MITRE ATLAS and industry literature. These risks included technical vulnerabilities (e.g., prompt injection, data poisoning) as well as broader misuse and safety issues. Risks were categorised using standard security terminology - attack, exploit, hazard, vulnerability - and mapped to the 3D model (Disrupt, Deceive, Disclose) for security issues, or to Safety for internal or unintentional AI harms.

**Normalised Coverage-Depth Score (CDS)**
To compare frameworks systematically, we developed a Normalised Coverage-Depth Score (CDS):

- **Coverage:** measures the percentage of total risks that a framework addresses.
- **Depth:** measures how comprehensively the framework proposes mitigations for those risks, relative to the most detailed framework in the set.

The final CDS combines these dimensions into a single value, rewarding frameworks that balance breadth of coverage with detailed, actionable guidance.

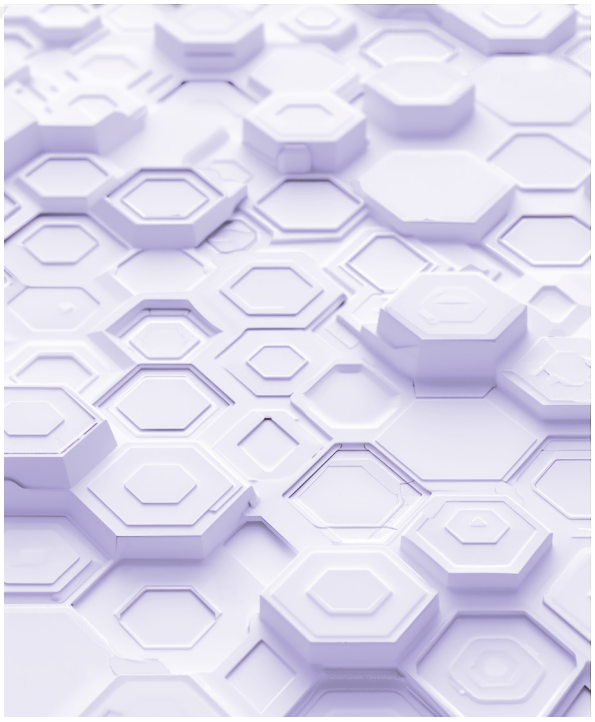
**Qualitative Evaluation**
In addition to the CDS, frameworks were qualitatively assessed for their treatment of two high-priority AI security risks - prompt injection and model evasion - to gauge their practical utility for real-world defence.

## SUMMARY

Across all three components, this methodology connects quantitative data, AI-assisted classification, and comparative framework analysis to answer a single question:

> **How likely are AI security incidents to occur, and how effectively do existing frameworks prepare organisations to manage those risks?**

The combined approach links vulnerability data, real-world incidents, and governance structures to create an evidence base for future quantitative modelling and policy development.



# RESULTS

This section summarises the key findings from the three stages of the study. Together, they provide an integrated view of how AI security incidents are represented in public data, how they manifest in practice, and how well existing frameworks support their mitigation.

## PART 1: CVE ANALYSIS

The first analysis aimed to identify how many vulnerabilities recorded in the Common Vulnerabilities and Exposures (CVE) database were related to artificial intelligence. Out of 18,779 CVEs from 2020-2025, fewer than 1% were classified as AI-related.

| Category | Count | Percentage | Description |
|---|---|---|---|
| **AI Security** | 31 | 0.0016% | Direct vulnerabilities in the AI model itself (e.g., prompt injection, model manipulation). |
| **AI Supply Chain** | 113 | 0.006% | Weaknesses in datasets, pre-trained models, or supporting infrastructure. |
| **AI IT Cybersecurity** | 93 | 0.0049% | Traditional vulnerabilities (e.g., DLL hijack, path traversal) within AI systems. |
| **General Cybersecurity** | 18,542 | 98.6% | Standard IT vulnerabilities unrelated to AI. |

AI vulnerabilities exist but are extremely underrepresented in public vulnerability databases. The vast majority of flaws impacting AI systems are currently classified as general cybersecurity issues, masking their unique causes and implications.

Some examples of each of these vulnerability types include:

- **AI Security:** A prompt injection vulnerability in Blackbox AI v1.3.95 allowed attackers to exfiltrate previous user-AI conversations.
- **AI Supply Chain:** A regex denial-of-service vulnerability in the Giskard AI testing framework (v2.15.5) allowed dataset-based service disruption.
- **AI IT Cybersecurity:** A DLL hijack vulnerability in Lenovo's AI-enabled PC Manager allowed code execution with elevated privileges.

This result supports the hypothesis that AI-specific vulnerabilities remain poorly defined and inconsistently tagged, limiting accurate incident prediction and cross-framework analysis.

# PART 2: AI SECURITY INCIDENT ANALYSIS

This stage sought to identify, classify, and understand real-world AI security incidents drawn from four major databases: AIID, AIAAIC, MITRE ATLAS, and AIVD. Based on our analysis key takeaways include:

**1.** The AI imputation method attempted was not accurate. Future work should be attempted with more powerful models. In the AIID and AIAAIC datasets, incidents labelled as "threat actor" cases were often misclassified by the AI model used for tagging, highlighting current challenges in automated data labelling.

| Database | Incident Type | No Actor | Threat Actor |
|---|---|---|---|
| AIID | Misuse of AI | 27 | 413 |
| | Failure Mode | 495 | 65 |
| | Combination | 8 | 37 |
| AIAAIC | Misuse of AI | 71 | 719 |
| | Failure Mode | 399 | 181 |
| | Combination | 23 | 137 |

**2.** There is evidence of AI security incidents. Incidents that resulted in harm (rather than a proof of concept or research) often involved systems that were widely used and techniques that were not too complex.

**3.** On the suitability of of these datasets out of the box for cyber risk modelling, there are two main limitations:

**1.** On informing likelihood, as these sources rely on media, it is unclear whether the incidents reflect either true frequency or reporting bias, where certain accidents and misuse receive more media attention.

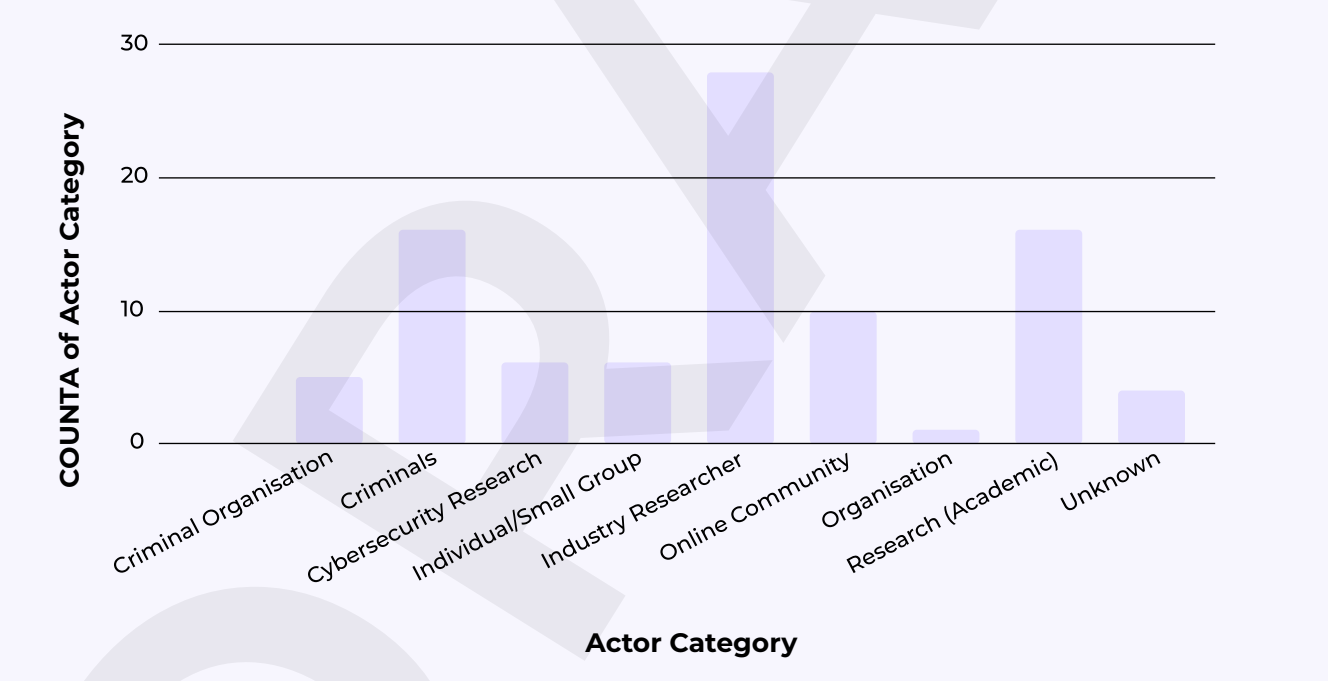
| Category of Primary Source | COUNTA of Event Name |
|---|---|
| Academic Publication | 11 |
| Blog | 24 |
| Database | 1 |
| Database Only | 6 |
| Media Release | 9 |
| News Reporting | 26 |
| Research Report or White Paper | 7 |
| Social Media | 11 |
| **GRAND TOTAL** | **95** |

| Size of Affected Organisation | Number of incidents |
|---|---|
| | 21 |
| Large business: >200 | 64 |
| Medium: 20- 199 | 4 |
| Micro: <5 | 2 |
| Micro: <5, Small: 5-19 | 1 |
| Micro: <5, Small: 5-19, Medium: 20- 199 | 1 |
| N.A | 2 |

**2.** Most incidents were linked to human misuse or AI system failure, rather than targeted adversarial exploitation. These failure modes are still important to address, as they result in real losses, but maybe less helpful if the focus is AI security.

**3.** These databases contained a large proportion of incidents that did not lead to significant losses which would not contribute to cyber risk modelling (see below that incidents involving individual actors made up a large percentage of incidents in the specific dataset.)

**AI Security Incident Monitoring (Dec 2024) - Actor Type**




**4.** The majority of AI security incidents stem from well-known risks such as privilege escalation, data exposure, or system misconfiguration, rather than entirely new attack classes. However when AI specific attacks (such as Adversarial Machine Learning) were observed, they were often more straightforward which does not align with the academic focus of the field.

| NIST Top 1 | COUNTA of NIST Top 1 |
|---|---|
| | 0 |
| AI Supply Chain Attacks | 3 |
| Black-Box Evasion Attacks | 20 |
| Black-Box Evasion Attacks Score-based attacks | 1 |
| Cyber [not NIST list] | 25 |
| Direct Prompt Injection Attacks | 10 |
| GenAI based attack [not NIST list] | 7 |
| Indirect Prompt Injection Attacks | 4 |
| Leaking sensitive information. | 1 |
| Malicious executable in model files [not NIST list] | 3 |
| Model Extraction | 4 |
| Model Poisoning | 2 |
| N.A | 4 |
| Property Inference | 1 |
| Targeted Poisoning | 1 |
| White-Box Evasion Attacks | 3 |
| White-Box Evasion Attacks Optimization-based methods. | 3 |
| Worm [not NIST list] | 1 |
| **Grand Total** | **93** |

**5.** AI introduces unique amplifiers of risk. This included, large and unpredictable input spaces, low interpretability and probabilistic outputs, dependence on upstream datasets and third-party models, centralised control of high-value data and capabilities. These characteristics both create new vulnerabilities and magnify traditional ones, underscoring the need for integrated risk management rather than siloed “AI-only” approaches.

# PART 3: AI SECURITY FRAMEWORK ANALYSIS

The framework analysis assessed how well five major AI governance and security frameworks address the risks identified in Parts 1 and 2.

| Framework | Risks | Mitigations | Coverage (%) | Depth (%) | CDS Score |
|---|---|---|---|---|---|
| NIST AI RMF | 151 | 95 | 100.0 | 100.0 | 200.00 |
| EU AI Act | 151 | 94 | 100.0 | 98.95 | 198.95 |
| Microsoft GSAIP | 151 | 77 | 100.0 | 81.05 | 181.05 |
| NCSC Guidelines | 150 | 61 | 99.34 | 64.21 | 163.55 |
| ISO/IEC 42001 | 150 | 58 | 99.34 | 61.05 | 160.39 |

**Findings:**

- **NIST AI RMF** achieved the **highest Coverage-Depth Score (CDS)**, addressing all identified risks and offering the **most comprehensive mitigations**, particularly for “disrupt” and “disclose” threats.
- The **EU AI Act** provided **similar coverage but less technical depth**, reflecting its focus on governance and compliance over operational guidance.
- **Microsoft’s GSAIP** performed **strongly as a practical enterprise** tool but is **limited by its commercial scope**.
- **ISO/IEC 42001 and NCSC guidelines** provided **moderate coverage** and **fewer specific mitigations**, with ISO notably omitting adversarial examples.

**Interpretation**
No single framework provides complete coverage across the spectrum of AI security risks. Regulatory frameworks emphasise governance and ethics, while industry frameworks offer operational detail but narrower scope. The results highlight an opportunity to combine these perspectives - linking quantitative vulnerability data (Part 1) and real-world incident trends (Part 2) with practical control frameworks (Part 3) to develop more adaptive AI risk management models.

## SUMMARY

**Across all analyses, the findings indicate that:**

- **AI-specific vulnerabilities exist but are likely underreported and inconsistently categorised.**
- **Most public incidents currently stem from misuse or system failure rather than deliberate attack, though both categories share underlying risk drivers.**
- **Existing frameworks address AI security unevenly, often focusing on governance or ethical risks rather than technical attack mitigation.**

This evidence base provides a foundation for future likelihood modelling and risk quantification, and supports calls for clearer incident reporting, better AI vulnerability classification, and unified frameworks that integrate both cybersecurity and AI-specific assurance.

# LIMITATIONS AND FUTURE WORK

This study faces several inherent limitations that reflect both the nascent stage of the AI security field and the immaturity of available public data. AI-related vulnerabilities and incidents are still infrequently reported and inconsistently categorised, making it difficult to develop statistically robust models of likelihood or risk. The scarcity and uneven quality of publicly documented AI incidents remain a key barrier to comprehensive analysis.

While the dual CVE classifier performed effectively in distinguishing AI-related vulnerabilities from traditional ones, the brevity and vagueness of many CVE descriptions-some containing fewer than ten words-made it challenging to accurately interpret context or intent. Similarly, because parts of the classification and incident-labelling processes relied on AI-assisted semantic models, results are subject to the inherent variability and opacity of large language model outputs. These limitations underscore the need for improved transparency and reproducibility in AI-supported analytical workflows.

Furthermore, biases within public incident databases - such as media-driven sampling bias and regional underreporting - likely distort the observed distribution of AI security versus AI safety events. Until more standardised and comprehensive reporting mechanisms are in place, estimates of AI incident frequency must be treated with caution.

Looking ahead, these limitations present opportunities for targeted advancement. As AI becomes more deeply integrated into critical systems, the number and detail of reported AI incidents will naturally increase, improving data availability for future studies.

## FUTURE WORK

**Future research should focus on:**

- Developing **shared AI incident reporting standards**, potentially in partnership with international standards bodies or regulators.
- **Integrating private sector and proprietary data sources** (e.g., security vendor telemetry) to complement public databases.
- **Expanding the classification model** to include probabilistic reasoning or Bayesian likelihood estimation for emerging attack types.
- **Benchmarking incident analysis against loss and impact data**, enabling transition from descriptive to quantitative AI risk modelling.
- **Exploring temporal trends**, particularly how new model architectures (e.g., multimodal, agentic systems) introduce evolving classes of vulnerability.

These next steps will enhance both the accuracy of likelihood estimation and the usefulness of AI risk frameworks for decision-makers. By addressing current data and methodological gaps, future iterations of this research can move toward evidence-based, predictive AI security risk management.

# RECOMMENDATIONS

The findings indicate that AI security incidents remain a small but distinct category within the broader cybersecurity landscape. Their relative scarcity in public datasets is likely not due to low occurrence but rather underreporting and poor classification. Most AI-related vulnerabilities identified occur in supporting infrastructure and supply chain components, where detection and attribution are easier using conventional security tools.

However, as AI becomes more integrated into critical business and government systems, the nature of security risk is shifting. Vulnerabilities now extend beyond traditional IT flaws to include issues that arise from the behaviour, training data, and decision processes of AI models themselves.

Prompt injection, for example, has rapidly emerged as a dominant category of AI security incident since the public release of large language models. This illustrates how each generation of AI technology introduces new, unique exposures - and how current detection and reporting methods are struggling to keep pace.

Even within the short period between the December 2024 interim analysis and this final report, the threat landscape has evolved markedly. The rapid development and deployment of AI systems have outpaced the adaptation of risk management frameworks, regulatory requirements, and assurance methods.

## KEY **RECOMMENDATIONS**

**1. Establish consistent definitions and reporting for AI security incidents.**

- → Develop unified terminology distinguishing AI security from AI safety and general cybersecurity.
- → Encourage incident reporting schemes that explicitly capture AI-specific causes and impacts (e.g., model manipulation, data poisoning, prompt injection).
- → Align with international standards bodies to ensure comparability across jurisdictions and sectors.

**2. Integrate AI-specific vulnerabilities into existing cybersecurity processes.**

- → Expand public vulnerability repositories (e.g., CVE) to include explicit AI-related tags and descriptors.
- → Support initiatives to embed AI/ML vulnerability categories within vulnerability scoring systems (e.g., CVSS).
- → Train cybersecurity analysts to identify and document AI-related issues using consistent taxonomies.

### 3. Strengthen supply chain and infrastructure assurance for AI systems.

- → Require stronger controls for dependencies, data pipelines, and third-party model integrations that underpin AI systems.
- → Prioritise security reviews for open-source models and libraries, which often serve as attack vectors.
- → Treat AI supply chain vulnerabilities as integral to overall organisational cyber risk.

### 4. Encourage proactive, cross-sector collaboration.

- → Create partnerships between academia, government, and industry to share incident data, indicators of compromise, and mitigation patterns.
- → Establish anonymised or aggregated AI security incident databases to improve evidence-based policymaking.
- → Foster public-private research collaborations to identify and track emerging attack techniques.

### 5. Improve the usability of AI risk management frameworks.

- → Harmonise the practical elements of existing frameworks (e.g., NIST AI RMF, EU AI Act, ISO/IEC 42001) for clearer implementation guidance.
- → Include likelihood and severity scoring mechanisms tailored to AI system contexts.
- → Provide clearer mapping between AI-specific risks and existing cybersecurity controls to reduce redundancy and confusion.

### 6. Invest in education and capacity-building for AI security.

- → Upskill both security and data science professionals in adversarial AI, model robustness, and AI governance.
- → Integrate AI security principles into broader cyber training programs.
- → Encourage organisations to establish dedicated AI assurance roles or teams.

### 7. Promote continuous reassessment as AI evolves.

- → Recognise that AI security risks are dynamic; prompt injection and model evasion are today's challenges, but others will follow as architectures evolve.
- → Encourage ongoing updates to incident databases, frameworks, and regulations to reflect this evolution.

In summary, AI security risk is not static - it evolves with every new system design and deployment trend. Strengthening how the industry reports, categorises, and manages these risks is critical for preventing emerging vulnerabilities from scaling alongside AI adoption. Addressing underreporting, integrating AI security into standard cybersecurity practice, and fostering cross-sector collaboration will be essential to building trustworthy AI systems over the next decade.