AI Incident Reporting Framework for India

Discussion Paper

Dr. Geetha Raju

Prof. Balaraman Ravindran

Centre for Responsible AI (CeRAI)

Wadhwani School of Data Science and AI (WSAI)

Indian Institute of Technology Madras

October 2025

Authors

Geetha Raju
Senior Policy Analyst, Centre for Responsible AI, IIT Madras (CeRAI)

Balaraman Ravindran
Head, Centre for Responsible AI (CeRAI) at IIT Madras.
Professor and Head, Wadhwani School of Data Science and AI (WSAI), IIT Madras.

Acknowledgement

The authors extend their sincere gratitude to all the subject matter experts for accepting our invitation and sharing their deep insights and expertise during the consultation sessions. We sincerely thank Mr. K.S.Venkataraghavan, Senior Project Advisor, RISE VLSI Lab CSE Department, IIT Madras, for taking the time to review the paper and offer unbounded guidance. We would also like to acknowledge Ms. Shruchi Singh, Post-Baccalaureate Fellow, CeRAI, and Ms. Sameeksha Chandrashekar, Intern, CeRAI, for their diligent research assistance.

Table Of Contents

A. Executive Summary

Artificial Intelligence (AI) is rapidly transforming all aspects of human life, society, organizations, and governments by advancing research, innovation, productivity, and accessibility across sectors. Its integration into daily routines and workplaces occurs through on-demand services, personal choices, and curiosity, presenting unprecedented opportunities. However, these benefits are matched by significant risks in real-world deployments, which are especially pronounced in India's unique context of rich societal diversity, linguistic complexity, varying digital literacy, and structural socio-economic challenges.

This environment necessitates context-aware and comprehensive protocols for AI governance, emphasizing nuanced application contexts and their associated robust risk management requirements. Ultimately, the aim is to address these structural imbalances through digital intervention with AI while minimizing risks, harms, and unintended consequences. In the context of India, adopting a one-size-fits-all approach to AI transformation would not only be ineffective but could also deepen existing inequalities and widen gaps between underrepresented communities and others.

Despite greater emphasis and ongoing efforts to conduct continuous AI risk assessments and management throughout the AI lifecycle, deployed AI solutions still pose risks such as bias, exclusion, data breaches, misinformation, copyright violations, and misuse. This highlights the need for real-time identification, reporting, documentation, and mitigation of AI harms and incidents by responsible entities, including implementing agencies and policymakers, supported by real-world evidence. While there are several global collaborative initiatives for AI incident reporting portals and frameworks, their effectiveness in fully managing and mitigating AI incidents remains limited.

In India, given its complex technology and policy landscape, there is a critical need for a robust AI incident management framework driven by multi-stakeholder collaboration. The lack of a clear understanding of AI risks and standardized risk classification within the Indian context complicates and undermines incident reporting, rendering it inefficient and unclear. This situation prevents individuals and organizations from publicly responding to AI harms, emphasizing the urgent need for clear, context-specific guidelines and frameworks to improve transparency, accountability, and responsiveness in AI incident management.

Consequently, in this paper, we seek to answer the following pertinent questions:

- What is an AI Incident?
- How can individuals or organizations identify AI-related harm as an AI incident?
- How can individuals or organizations report AI incidents?
- Who bears responsibility for acknowledging and managing these incidents?
- What processes and frameworks exist or need to be built for effective AI incident management?
- What roles do governments, industry, academia, policymakers, social scientists, and citizens play in AI incident reporting?
- Can the emergent AI-related risks or any unforeseen and concomitant risks be proactively identified and evaluated under the proposed framework of AI incident reporting?

Following this, we provide recommendations for the design of an AI Incident reporting database, a framework for AI incident reporting, along with the governance structure required to implement this system. These recommendations are accompanied by suitable implementation strategies to facilitate effective operationalisation and relevance within India's AI landscape.

B. Recommendations

a) Establishing standardized definitions, taxonomies, and protocols required for AI incident management in India

India should develop standard definitions for AI incidents, AI risks, and AI harm. This should invariably involve developing and establishing multi-dimensional taxonomies[a] for AI risk and AI harm, accompanied by protocols for incident-specific AI impact assessment, guidelines for evidence-based risk classification[b], and strategies for AI risk/harm mitigation considering the Indian context.

Multi-dimensional taxonomies provide a formalized vocabulary to identify complex and hidden links between different types of harm and impact. They consider factors such as type, severity, scope, source, and pathway of risk or harm. These taxonomies support the implementation, assessment, evaluation, and continuous refinement of AI incident management policies.

This approach is valuable because it enables the use of identified AI incidents, risks, and harms to recalibrate models and fine-tune parameters. It also supports the adoption of explainable AI models, algorithms, and solutions that align with domain-specific priorities, practices, and policies across various sectors.

b) Establish an ombudsman body responsible for AI incident reporting.

Given the absence of enforcement agencies like CERT-In for AI incidents, it is preferable to establish an independent ombudsman body to oversee AI incident management in India. This ombudsman body would be responsible for establishing, provisioning, and coordinating a uniform AI incident reporting process and framework. The AI incident management should be based on a federated database for AI incidents, supervised by the ombudsman to ensure comprehensive implementation across all domains and applications. This setup supports uniform governance models and formalizes multi-dimensional taxonomies for AI incident data collection, documentation, and knowledge extraction, enabling effective management.

a) A multi-dimensional AI risk and harm taxonomy is a structured framework that systematically categorizes AI-related risks across multiple dimensions, including technical aspects such as data lineage and provenance, incident evidence, correlation and causation, operational usage patterns and scenarios, as well as contextual variables encompassing language, culture, social and model parameters. This taxonomy facilitates comprehensive risk and harm assessment, classification, and management, enabling tailored governance approaches across diverse AI systems and their deployment environments.

b) Evidence-based risk classification and incident-specific AI impact assessment calls for a rigorous knowledge extraction followed by scientific validation. This primarily and manifestly necessitates a data-driven approach in assessing and classifying AI incidents and harms. These manifestations can help in the contextual evaluation of the risks and harms of AI systems. They often are strong indicators of the safety, reliability, effectiveness, efficiency, and ethicality of AI models and their implementation. For example, evidence-based risk classification in critical domains like healthcare can be corroborated by a data-driven approach, with which emulative reproducibility and knowledge extraction provide a comprehensive risk analysis based on sound scientific principles of empirical evidence.

A significant advantage of an ombudsman-led federated database is its ability to facilitate transparent, timely, and efficient communication among the incident management team, responsible entities, and domain regulators. It is also possible to promptly manage the crucial communication loops involved in incident assessment and response.

For successful AI incident reporting and management, clear timelines must be set for every stage, including mandatory reporting, verification, and publication of incidents. The ombudsman-based federated architecture readily addresses these requirements, ensuring accountability and streamlined coordination across India's complex technological and regulatory environment.

i) Federated Incident Reporting Database governed by an ombudsman body (central database with locally distributed and governed in a federated manner):

A federated approach to a national AI incident database is rooted in principles of public interest, transparency, and collaboration among multiple stakeholders. The ombudsman-based governance authority ensures the inclusion of representatives from diverse fields and areas of expertise. The database will have access controlled by a governance hierarchy, with roles assigned based on specific functions and organizational responsibilities. This structure guarantees that information about AI incidents is protected while remaining accessible to the public view appropriately.

Such an approach facilitates the development of various facets across AI application sectors. It is designed to be flexible and logically distributed, accommodating heterogeneity, isolation, specificity, diversity, and complexity across different regions, domains, and sectors. Given its federated nature, the national-level database will possess the capability to query and aggregate data from subordinate local or regional databases when needed. This structure ensures efficient data sharing and comprehensive incident management across India's diverse landscape.

ii) Establish standard operating protocols (SOPs) aiming for effective AI incident management

The ombudsman body, supported by a federated database, plays a crucial role in establishing and enforcing a standard operating protocol for AI incident management in India. These protocols cover various stages of incident management to ensure thorough and effective handling. Key stages include AI incident analysis, which involves mapping and classification, and assigning both qualitative and quantitative metrics for better management. The protocols also address incident reporting, secure data storage and retention, and comprehensive evidence collection. Further, the protocol ensures preservation of provenance and lineage, along with fact-checking and verification during the incident assessment phase.

It mandates anonymization of incident data and enforces purpose limitation to protect sensitive information. Additionally, role- and responsibility-based access controls regulate data access, and incident-specific impact and severity assessments are integral to managing incidents effectively. This comprehensive framework ensures transparency, accountability, and robustness in AI incident management.

iii) Conducting seamless and regular audits of the AI incident database

The ombudsman body will conduct periodic audits of the AI incident database. These audits will enable national-level reviews of incidents and contribute to improving the overall AI incident management system. The review process will enhance incident assessment methodologies, refine management procedures, and update taxonomies along with their multi-dimensional frameworks. This continuous auditing ensures that the AI incident reporting system remains robust, adaptable, and effective through regular evaluation and fine-tuning.

c) Ensure a Hybrid approach to AI incident collection

A notable advantage of the hybrid approach is the automated collection of India-specific AI incidents from various sources, such as news articles, legal documents, web alerts (for example, Google Alerts), and existing global AI incident databases. In addition to automated collection, it incorporates human-initiated reporting. This human-driven process includes multiple pathways, such as mandatory reporting by organizations, voluntary reporting, citizen reporting, and closed community reporting by relevant entities. This diverse approach ensures comprehensive and localized data gathering for effective AI incident management.

i) Significance of mandatory reporting of AI incidents by developers/deployers of high-risk public AI systems.

Ensuring mandatory AI incident reporting by developers and deployers of high-risk public AI systems requires clear operational guidelines. These guidelines should provide domain-specific and risk-based classifications of AI systems. Such classifications will help in identifying risk levels at both ends of the AI value chain, including design and development as well as deployment and provisioning. Additionally, this approach supports credit-based awarding to encourage responsible practices throughout the AI lifecycle.

ii) Significance of mandatory reporting of AI incidents by organisations using AI in their day-to-day workflow/operations.

Mandatory reporting of AI incidents by organizations using AI in their operations is essential for identifying and mitigating risks and harms caused by these solutions.

This process enables guidance on organization- and application-specific incident classification and risk assessment from a broad end-user perspective. It also supports rigorous oversight and fosters continuous improvement of AI models and services, enhancing their safety and effectiveness over time.

d) Establish responsibilities for domain regulators for AI incident management

Domain regulators and relevant entities in India should actively participate in post-incident analysis activities. Their role includes raising awareness about legal implications, compliance requirements, and domain-specific best practices, guidelines, and policies. They should also identify potential liabilities for responsible parties, which may involve imposing penalties, levying fines, or decommissioning the AI system, depending on the severity and nature of the AI incident. This involvement ensures accountability and adherence to regulatory standards in managing AI-related risks.

i) Mandatory AI risk and incident disclosure:

Domain regulators and relevant authorities should develop a legislative framework that requires all AI providers to disclose risks and incidents associated with their AI components, such as datasets, models, services, and systems deployed in India. This disclosure, through clear flagging of AI components, will aid in informed decision-making among stakeholders about the potential risks involved in AI adoption.

e) Promoting shared responsibility among AI actors through contracts/agreements

Organizations involved in AI technologies deployable in India should implement a shared responsibility model. This model helps all stakeholders, including AI providers, deployers, developers, data principals, and end users, to clearly understand their hierarchy, roles, functions, and responsibilities. The roles and responsibilities will be defined in the contractual agreements and user guidelines developed for each specific AI system. Such agreements and guidelines are essential for establishing liabilities and identifying the responsible entity mandated to oversee response and accountability in the event of an incident.

f) Establish Domain-specific AI Risk Mitigation Strategies for India

India requires AI risk mitigation strategies tailored to its unique socio-economic context and digital landscape. These strategies must go beyond globally accepted or standardized frameworks, which may not fully address India-specific challenges. However, India can still incorporate international best practices to enhance its approach. This balance enables the development of domain-specific risk mitigation measures that effectively bridge the country's digital divide, accommodate diverse user demographics, and ensure continuity of critical services, while maintaining global compliance.

Additionally, domain regulators should establish grievance redressal mechanisms for individuals, communities, and organizations affected by AI systems. A two-tier grievance redressal process is recommended, starting at the organization or application level, with the option to escalate to the national level. This structure ensures accessible complaint procedures, independent reviews, appeals, and alternative dispute resolution, all supported by qualified human oversight. These measures promote fairness, accountability, and trust in AI deployment across India.

g) AI incident awareness for AI products in practice

AI providers in India should widely engage and educate users about their AI products and services. This includes clearly communicating the capabilities, limitations, and incident reporting procedures in the AI context. To achieve this, AI providers should develop domain-specific Responsible AI guidelines, which can be standard deployment guidelines, user manuals, pre-reads / informational materials, or risk awareness training modules. Such resources must be tailored to the diverse needs of India's communities, focusing on application relevance and end-user sensitivity to ensure effective understanding and responsible use.

h) Public Sector AI Watch

The recommended ombudsman-based approach includes establishing a comprehensive "Public Sector AI Watch" registry. This registry will track AI applications deployed by the government for public use. Its purpose is to enhance responsibility, fairness, transparency, and accountability in AI deployments across India. This approach is particularly important for assessing risk in high-impact AI applications such as welfare benefit systems, law enforcement tools, healthcare diagnostics, education assessment platforms, and citizen service portals. It would also address their specific risk profiles, including bias vulnerabilities, privacy concerns, and operational limitations, ensuring safer and more equitable AI use in the public sector.

C. Glossary

1. **Artificial Intelligence (AI)**: AI refers to a wide range of technologies capable of performing complex tasks without active human control or supervision. It includes systems that may generate outputs from learning experiences, reasoning, problem solving, perception, understanding of natural language, adapting to new situations, identifying, recognizing, and even creating an object, and many more.
2. **AI-created object**: This may refer to an item, content, or artifact that an AI implementation aims to generate, design, or produce as an indispensable part of its end-user application. These AI-created object is often an automation process which are part of AI models themselves.
3. **AI Governance**: A system of frameworks, practices, and processes at an organizational level. AI governance helps various stakeholders implement, manage, and oversee the use of AI technology. It also helps manage associated risks to ensure AI aligns with stakeholders' objectives, is developed and used responsibly and ethically, and complies with applicable requirements.
4. **AI Risk**: Risk is a function of both the probability of an event occurring and the severity of the consequences that would result. AI Risk depends on both the system's capabilities and the context of deployment.
5. **Potential AI Harm**: Incidents or conditions that have the likelihood of causing harm but have not yet resulted in actual damage.
6. **Actual AI Harm**: Actual harm is often expressed as a risk that materialised into harm. They lead to outcomes that disadvantage or damage individuals, businesses, or society, including physical, economic, privacy, and safety harms.
7. **AI Hazard**: An event, circumstance, or series of events where the development, use, or malfunction of one or more AI systems could plausibly lead to any of the following harms:
    a. **injury or harm to the health** of a person or group of people;
    b. **disruption of the management and operation of critical infrastructure**;
    c. **violations of constitutionally guaranteed sovereign rights or a breach of sovereignty obligations under the applicable law** intended to protect fundamental, federal, sovereignty principles, constitutional obligations such as protection against "at-will employment", and intellectual property rights; or
    d. harm to property, communities, or the environment.
8. **AI Incident**: An AI Incident is an event, circumstance, or series of events where the development, use, or malfunction (surreptitiously or unintendedly) of one or more AI systems directly or indirectly leads to one or more of the following

harms to an individual, businesses, or society:

a. **physical safety** issues, injury or harm to the mental or physical health of a person or group of people; unauthorized access, denial, or disruption of service
b. **violation of** constitutionally guaranteed sovereign rights or **breach of obligations under the national law/government's policies rooted in constitutional grounds**, which otherwise would have ensured constitutional aspects guaranteeing equality, non-discrimination, privacy, intellectual property obligations, socio-cultural equanimity, access to fair and equitable education, work, public assistance in certain cases, etc.
c. **harm to property, communities, socio-economic status, or the planet/environment**;
d. **cyber-incidents** - malfunctions, failures, unauthorised or discriminatory outcomes, unforeseen behaviour, deepfakes, misinformation, overt and covert operation, surreptitious behaviour, etc.
e. **disruption of the management or operation of critical infrastructure.**
f. poses a **threat to national security, sovereignty, and constitutionally rooted governance frameworks.**

9. **Serious AI Incident**: A serious AI incident is an event, circumstance, or series of events where the development, use, or malfunction of one or more AI systems directly or indirectly leads to any of the following harms:

a. **death of a person** or serious harm to the health of a person or group of people;
b. **serious and irreversible disruption** of the management and operation of critical infrastructure;
c. **serious violation** of constitutionally guaranteed sovereign rights or a **serious breach of obligations under the applicable law** intended to protect against "at-will employment", intellectual property rights etc;
d. **serious harm to property, communities, or the environment.**

10. **AI Disaster**: An AI disaster is a severe AI incident that **disrupts the functioning of a community or society** and may test or exceed its capacity to cope using its resources. The effect of an AI disaster can be immediate and localised, or widespread and lasting for a long period of time.

11. **AI Near Miss**: Near misses are events that could have led to an AI incident.

12. **Autonomy**: The ability of an AI system to operate independently of human intervention.

13. **Responsible Entity**: An individual, team, or organisation that is formally accountable for an AI incident, such as AI system developers and providers, deployers and operators, end-users and adopting organizations, AI platform providers, etc.

14. **AI Incident Note**: A verified record capturing the details of an AI incident, including the harmful event or circumstances, real-world facts, severity, impacts, affected entities, timeline, status, and initial mitigation measures.
15. **AI Response Note**: A structured record documenting the management of an AI incident, including detection, analysis, mitigation, and recovery actions, communication measures, resolution status, and key decisions across the incident timeline.
16. **AI Stakeholders**: Individuals, groups, or entities with an interest in or affected by AI systems, including users, developers, deployers, organizations, policymakers, vulnerable populations, the public, journalists, and whistleblowers.
17. **AI Impact Assessment**: An evaluation process designed to identify, understand, and mitigate the potential ethical, legal, economic, and societal implications of an AI system.
18. **Domain-specific**: Pertaining to a particular industry or application area, such as healthcare, finance, telecommunications, or transportation, which may have unique challenges, requirements, and regulatory contexts for AI incidents or harms.
19. **End use-case scenarios**: A detailed description of final, real-world situations where AI systems are deployed, focusing on the sequence of interactions between systems and users with specific intentions or functional goals.
20. **End usage-sensitivity**: Measurement of how outcomes, decisions, or behaviors of an AI system change depending on variations in how, where, and by whom the technology is ultimately used, emphasizing context-dependent risks and ethical concerns.
21. **Data Provenance**: A process that tracks and logs the history and origin of records in a dataset, encompassing the entire life cycle from its creation and collection to its transformation to its current state. It includes information about sources, processes, actors, and methods used to ensure data integrity and quality. Data provenance is essential for data transparency and governance, and it promotes a better understanding of the data and, eventually, the entire AI system.
22. **Data Lineage**: Data lineage refers to the path and sequence of data's movement and transformations from its initial source through various pipelines, applications, and storage systems to its destination. In AI, lineage focuses on tracking how data flows through model development processes, enabling organizations to troubleshoot pipelines, ensure quality, and optimize system design by visualizing dependencies and transformations.
23. **Historicity**: Historicity in the AI context denotes the chronological and contextual evolution of data, AI models, or decisions over time. It encompasses both provenance and lineage, providing a time-based perspective that allows validation of data authenticity, the order of modifications, and the historical relevance of decisions generated by AI systems through different stages in the AI lifecycle.

24. **Regulation scope and context**: It refers to the boundaries, reach, and applicable legal or regulatory frameworks governing the development, deployment, and usage of AI systems, often tailored to specific geographical regions, domains, applications, or risks.
25. **Validation scope and context**: It refers to the parameters, processes, and contextual relevance used to test, evaluate, and confirm the accuracy, reliability, safety, and trustworthiness of AI models or outputs in real-world or simulated environments.
26. **AI incident occurrence sensitivity**: The degree to which an AI system is susceptible to variations in the timing, frequency, or conditions of incidents or events.
27. **AI Red Teaming**: The process of testing the security of an AI system through an adversarial lens by removing defender bias. It involves the simulation of adversarial attacks on the model to evaluate it against certain benchmarks, jailbreak it, and make it behave in unintended ways. Red teaming reveals security risks, model flaws, biases, misinformation, and other harms. The results of such testing are passed along to the model developers for remediation. Developers use red teaming to bolster and secure their product before releasing it to the public.
28. **AI Blue Teaming**: Defensive techniques and operational activities that monitor, respond, and protect AI systems against threats, adversarial actions, or incident scenarios, often in response to red team findings.
29. **Purpose Limitation**: A principle requiring organizations to clearly specify, document, and communicate the intended use of data and AI models, ensuring processing activities do not deviate from the explicitly defined purposes and remain lawful and ethical.
30. **Accountability**: Accountability in AI refers to the responsibility of AI developers, organisations, and stakeholders to ensure AI systems operate ethically, legally, and transparently. It involves mechanisms that enable AI decision-making to be monitored, explained, and challenged when necessary.
31. **Fairness**: An attribute of an AI system that prioritizes relatively equal treatment of individuals or groups in its decisions and actions in a consistent, accurate, and measurable manner. Every model must identify the appropriate standard of fairness that best applies, but most often it means the AI system's decisions should not adversely impact, whether directly or disparately, sensitive attributes like race, gender, or religion.
32. **Interpretability**: The ability to explain or present a model's reasoning in human-understandable terms. Unlike explainability, which provides an explanation after a decision is made, interpretability emphasizes designing models that inherently facilitate understanding through their structure, features, or algorithms. Interpretable models are domain-specific and require significant domain expertise to develop.

33. **Transparency**: It implies openness, comprehensibility, and accountability in the way AI algorithms' function and make decisions. It also refers to the extent to which information regarding an AI system is made available to stakeholders, including disclosing if AI is used through techniques like watermarking, and explaining how the model works through model or system cards, etc. It also refers to the maintenance of technical and nontechnical documentation across the AI life cycle to keep track of processes and decision-making, which can also assist with the auditability of the AI system.

34. **Explainability**: The ability to describe or provide sufficient information about how an AI system generates a specific output or arrives at a decision in a specific context to a predetermined addressee. Explainability is important for maintaining transparency and trust in AI.

35. **Responsibility**: A commitment by AI practitioners and organizations to develop, deploy, and monitor AI systems ethically and in alignment with the well-defined scope and purpose by ensuring constitutional rights, minimizing risks and misuse, and maximizing positive impacts for society.

36. **Reliability**: An attribute of an AI system that ensures it behaves as expected and performs its intended function consistently and accurately, even with new data that it has not been trained on.

37. **Robustness**: An attribute of an AI system that signifies the system's ability to be resilient to, overcome, and withstand security attacks. Robustness ensures the system's functionality, performance, and accuracy in a variety of environments and circumstances, even when faced with changed inputs or security attacks.

38. **Safety**: Safety in AI systems refers to designing, developing, and deploying AI systems that minimize AI harms, not limited to bias, misinformation, disinformation, deepfakes, hallucinations, and other unintended behaviors. It may also refer to mitigating and managing malicious use or rogue behavior. Safety also encompasses the prevention of existential or unexpected risks that may arise from advanced AI capabilities reflected in foundation models.

39. **Privacy**: The protection of individuals' personal information and data from unauthorized use or exposure, ensuring confidentiality, security, and legal privacy rights throughout the AI lifecycle.

40. **Anonymisation**: The process of modifying data to irreversibly prevent the identification of individuals, ensuring that personally identifiable information (PII) and sensitive data are removed so that data can be used without privacy risks.

1. Introduction

In recent years, AI adoption in social, political governance (including judiciary, executive, and legislature); participatory governance (including digital citizen engagement platform, public consultation of draft policies, etc); provenance and sovereign governance (including critical infrastructure sectors, defence, security sectors, etc) has increased significantly, driving advancements that would have otherwise taken decades to achieve. For example, under social governance, AI is deployed in the bio-sciences field to accurately predict protein structures, enabling faster drug discovery [1]. In defence, especially during peacetime, the use of AI-enabled drones has increased manoeuvrability in adverse climatic conditions and paved the way for a new era of drone-based surveillance [2]. While AI adoption is transforming all sectors, it also invariably introduces unique risks and harms. Figure 1 illustrates the lifecycle of AI harms evolving into AI incidents.

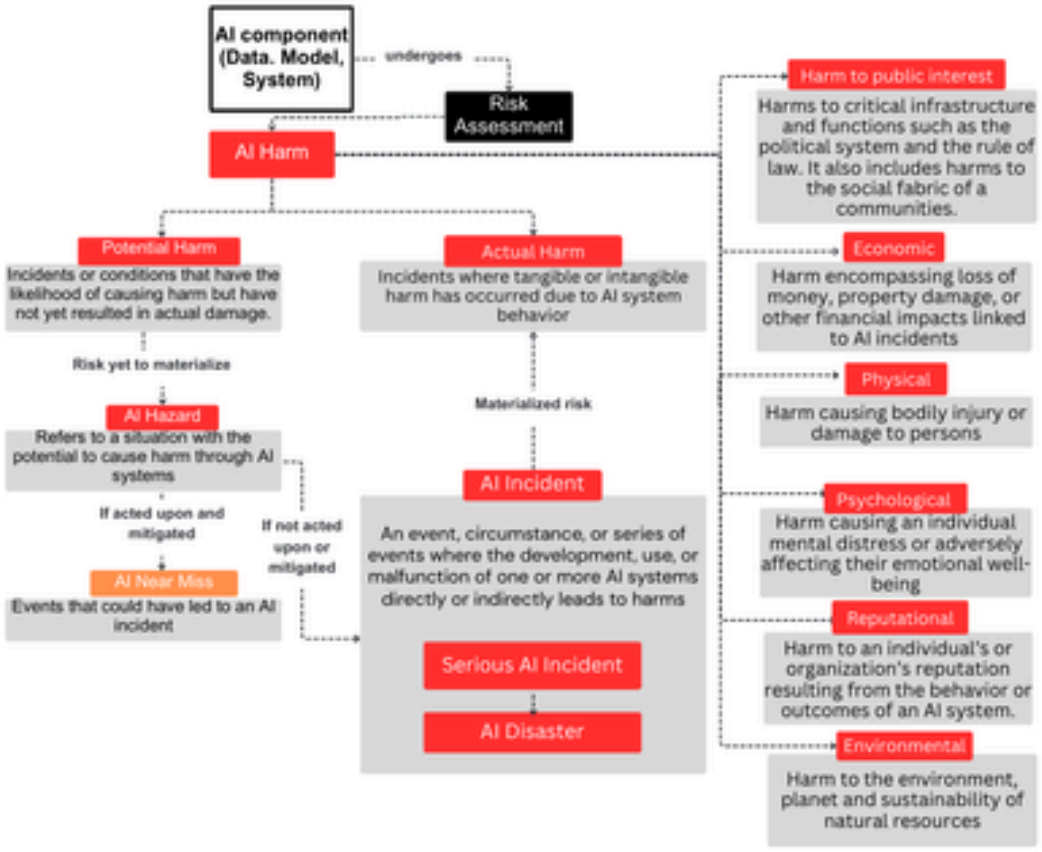


Fig 1: Lifecycle of AI risk, harm, and incident

Existing risk assessment frameworks can help to identify and mitigate known risks. However, the inherent complexities and the 'black box' model of AI systems make it challenging for the AI community to uncover unknown risks, determine the causes, or assess them accurately to develop some mitigation measures to minimise the impact. Also, there is no single source of information where the AI community can loop up for AI risks and AI harms in India. Consequently, organisations involved in developing and deploying AI systems often lack sufficient capacity or knowledge to proactively assess, prevent, or mitigate the unknown risks, leading to AI incidents that are identified only at the post-deployment stage.

1.1 Motivation

The rapid adoption of AI technologies in India underscores the critical need to proactively address the associated risks and challenges of this transformative technology. Recent research indicates that India leads global AI adoption, with a 30% adoption rate driven by significant digital transformation [3] across critical sectors, including healthcare, finance and banking, agriculture, education, manufacturing, IT services, retail, e-commerce, and public governance for enhanced service delivery. A drastic increase in AI incidents as time evolves with AI advancements is presented below.

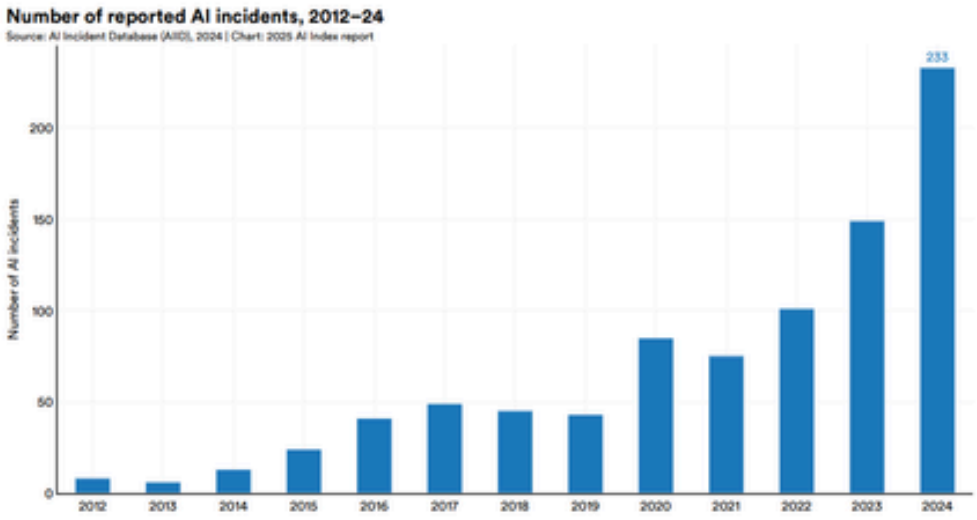


Figure 2: Increase in AI incidents reported since 2010 as reported in the AI Index report [4]

Globally, there are numerous AI incident databases developed by researchers and volunteer communities as part of the groundbreaking efforts in AI risk management. Popular amongst them are MIT's AI Risk Repository and AI Incident Database. Similarly, the OECD has provided a comprehensive list and definitions of AI incident-related terms to standardize understanding and reporting across jurisdictions [5]. In India, it is reported that the Telecommunication Engineering Centre (TEC) under the Department of Telecommunication (DOT) has drafted standards for an AI incident reporting schema and taxonomy in telecom and critical infrastructure sectors focused on systematic capture and analysis [6].

Despite this progress, India faces a notable gap in AI-specific risk management frameworks, policies, and legislation. Additionally, limited end-user awareness and varying maturity levels regarding AI systems and an organisation's AI readiness, AI transformation, and adoption policies, combined with diverse and evolving user needs, complicate effective risk mitigation. Given India's innovation-driven focus, it is neither feasible nor productive to anticipate and prevent all AI incidents in advance. Instead, establishing an ombudsman body supported by a federated database ensures the AI incident database serves as a single source of truth to systematically capture potential risks, extreme conditions, edge cases, and possible AI risk and harm mitigation strategies across the nation, which in the future will enable responsible and safe development, deployment, and use of AI systems in India.

However, India currently lacks a comparable, well-structured AI incident management system. Under the guidance of an Advisory Group, chaired by the Principal Scientific Advisor, the AI subcommittee has published a report on AI Governance Guidelines [7] and has also underscored this need. The Indian Computer Emergency Response Team (CERT-In) has established guidelines for reporting a security incident [8], and a responsible vulnerability disclosure and coordination policy [9] by the organisations. These guidelines are insufficient for managing AI incidents as they extend beyond conventional cyber threats to include intangible harms that can adversely impact individuals, businesses, and society at large. Recently, OWSAP has published a guide for GenAI incident response, which includes a detailed discussion on how AI incidents differ from traditional cyber incidents [10].

Therefore, a unified AI incident reporting database and a national-level AI incident management framework, governed by a dedicated entity such as India's AI Safety Institute, is essential. This initiative aligns closely with India's ongoing efforts to adopt AI and addresses the unique risks and harms experienced by Indian stakeholders through the IndiaAI Mission's Safe and Trusted AI pillar.

1.2. Aims and Objectives

We aim to provide a framework to guide the establishment of such an AI incident reporting system for India. The objectives of this paper are to:

a. Examine the AI Incident identification, reporting, and management process available globally
b. Conduct a gap analysis and understand the shortcomings of the existing incident reporting process in the global and Indian context
c. Propose an AI incident reporting database and management framework for India
d. Provide a governance structure for effective AI incident management in India

The following session presents the definition of an AI incident, the literature of the AI incident databases studied, and a detailed framework for managing AI incidents in India.

1.3. Research Methodology

The observations, analysis, and recommendations in this paper are drawn from interviews with experts from diverse backgrounds, as well as key research studies, policy recommendations, and established standards and frameworks developed by global bodies and industry practitioners. Insights from stakeholder consultations have been kept anonymous to protect the privacy of the experts and the information they shared.

1.4. Intent

This paper explores opportunities for establishing a nationwide federated AI incident reporting database governed by an ombudsman body for an effective incident management framework, which will

a. enable systematic and democratic reporting, documentation, and communication of AI incidents through an AI incident database that can act as a trusted knowledge source based on principles of federation and thus a source of all the AI incidents in India.
b. serve as an evidence base to inform policymakers, governments, and researchers for evidence-based decision-making and drive research & innovation, prioritising AI safety, thereby leading to a robust, safe, and trustworthy AI ecosystem in India; Translate lessons into actionable safeguards, inspired by cross-sector safety frameworks and tools.

c. develop a unified India-specific AI risk and AI harm taxonomy based on the evolving AI incidents.

d. monitor emerging risks and ensure appropriate oversight in AI systems.

e. provide an in-depth understanding of AI harms and AI risks from the context of India's diverse social, cultural, and legal landscape and enable AI practitioners to proactively mitigate risks in the future, thereby minimising harms and unintended consequences.

f. understand AI failures, threats, and vulnerabilities and accordingly develop metrology, risk classification strategies, and incident-driven impact assessment [11] to diverse scenarios or use cases.

g. improve public trust and transparency in AI systems by disclosing AI incidents.

2. Defining AI Incidents

There is no universally accepted definition of an AI incident, and at the same time, AI-specific and AI-related deployments have raised serious concerns. The table below introduces a distinct perspective on differentiating traditional software and cybersecurity incidents from AI incidents. The following definition of AI incidents emphasizes three core elements such as the 'AI incident triggering event', the 'AI component', and the corresponding 'RAI principle' that is compromised.

AI incidents are mostly recognized during the post-deployment phase, when users engage with what appears to be ordinary software. Often, users end up without realizing that AI models are driving key functionalities. On the other hand, harm can result not only from data or code defects in underlying infrastructure, tools, or frameworks, but also from model limitations, user interactions, and breaches of RAI principles such as fairness, transparency, and security.

By framing AI incidents under the system's context (e.g., nature, scope, intent and purpose) and in terms of the event that occurs (e.g., constitutionally guaranteed sovereign rights violation, hallucination, unauthorized access), the AI component at fault (e.g., data, model, API), and the responsible AI principle affected (e.g., fairness, accountability, robustness), AI incidents can be defined as follows:

AI Incident
An event, circumstance, or series of events where the development, use, malfunction, or deviation from the intended behavior of one or more AI components directly or indirectly leads to one or more of the following:
1. **Physical Safety and Health harms**: Injury or harm to the health of a person or group of people, or physical safety issues; 2. **Financial and Economic Harm**: Direct or indirect financial losses, economic damage, or market disruption to an individual or community or organisation; 3. **Reputational Harm**: Damage to the reputation or credibility of individuals, organizations, institutions, or public trust;

4. **Psychological and Emotional Harm**: Anxiety, stress, trauma, or other mental health impacts experienced by individuals or groups;
5. **Critical Infrastructure Disruption**: Unauthorised access, denial, or disruption or crippling of a service or operation of critical infrastructure;
6. **National Security Threats**: Threat to national security by enabling the development and deployment of CBRN (chemical, biological, radiological, and nuclear)weapons and supporting powerful offensive cyber operations and information warfare, aiding abuse and denial of constitutionally guaranteed sovereign rights, resulting in anti-state activity, societal instability, impacting national sovereignty, border and internal security, balanced social structure, and the sovereign government's obligations and functions.
7. **Violation of constitutionally guaranteed sovereign rights and breach of obligations**: Violation of constitutionally guaranteed sovereign rights or breach of obligations under the nation's law/policy, such as constitutional charters guaranteeing equality, non-discrimination, privacy, intellectual property obligations, and access to fair and equitable access to education, work, and public assistance in certain cases.
8. **Environmental Harm**: harm to the planet/environment leading to an unsustainable ecosystem;
9. **AI-enabled harms or cyber incidents**: Malfunctions, failures, unauthorised or discriminatory outcomes, unforeseen behaviour, deepfakes, misinformation, and other AI-specific security vulnerabilities. (Refer Annexure I).

AI Component Scope: These incidents may originate from any component within the AI technology stack required for designing, developing, and deploying the AI systems, including:

1. **Data**: Training datasets, validation data, real-time inputs, and data preprocessing systems.
2. **Models**: Machine learning algorithms, neural networks, foundation models, and fine-tuned systems.
3. **Tools and Frameworks**: Development platforms, AI libraries, and software development kits.
4. **Infrastructure**: Computing hardware, memory chips, storage, networking, and cloud platforms.
5. **Deployment Systems and Runtime Environments:** APIs, user interfaces, integration layers, and production environments.

6. **Governance Systems:** Monitoring tools, audit mechanisms, risk management, and compliance frameworks
7. **End Users:** AI perception, AI awareness, AI literacy, and intentions of individuals directly interacting with AI applications
8. **Deviation from the AI System's functional scope**: Any deviation from the AI system's functional scope, which is defined by its underlying intent[c], inherent nature[d], operational scope[e], and the specific purpose[f] driving its deployment, collectively shaping its role and impact within the targeted environment. It also covers any contextual variations, deviations from the cross-validated models, and abnormalities observed in AI system behavior.

Responsible AI Principles Scope: AI incidents would have occurred due to violations of core Responsible AI Principles sourced from any of the AI components mentioned above, which could create such harmful incidents. The following shows a comprehensive list of RAI principles that are a reason for an AI incident.

1. **Fairness and Bias**: Unfair treatment of individuals or groups based on protected characteristics such as race, gender, age, religion, socio-economic status, etc.
2. **The Black Box Problem - Lack of Transparency, Explainability, Interpretability, and Traceability**: Inability to provide clear insight into AI decision processes, due to which the stakeholders were unable to understand how and why outputs are generated.
3. **Lack of Accountability and Human Oversight**: Absence of clear ownership and responsibility for AI outcomes and poorly maintained human-in-the-loop controls to monitor, intervene, and correct system behavior.
4. **Data Privacy Concerns**: Improper collection, processing, or protection, storage, retention, anonymization, and reidentification of personal and sensitive information throughout the AI lifecycle.
5. **Lack of Reliability**: Insufficient testing, robustness checks, unpredictable behaviours, and failures at times of unexpected inputs or environmental changes.

[c] **Intent**: The primary goal or objective that the AI system is designed to achieve in its operational environment.
[d] **Nature**: The fundamental characteristics and design of the AI system, including its architecture, learning approaches, and deployment context.
[e] **Scope**: The boundaries or range of well-defined functions, tasks, and applications that the AI system covers or is intended to deliver.
[f] **Purpose**: The specific reason or rationale for deploying the AI system, focusing on its intended use and expected outcomes in practice.

6. **Inclusivity and Accessibility Constraints:** Lack of capabilities, say in terms of interface limitations (multimodal, multilingual, assistive features), social and cultural context, to meet the diverse user needs, excluding disadvantaged populations, marginalised communities, and individuals with physical disabilities/people who require special care and assistance.
7. **Ethical value misalignment:** Insufficient assessment of social impacts, fundamental rights guaranteed by the constitution, and environmental impacts that exacerbate inequalities and degrade human-AI trust, and a failure to ensure socio-cultural equanimity and access to fair and equitable access to education, work, and public assistance in certain cases.
8. **Misuse and abuse of AI systems**: intentionally exploiting AI systems beyond their intended purpose, scope, and nature through prompt injection attacks, jailbreaking attempts, or using AI for creating harmful content.
9. **Security Concerns**: Lack of adequate security measures and resilience capabilities leading to unauthorised access, system attacks and failures, degraded performance, malicious behaviour, manipulations (in case of GenAI, say, prompt injections, jail breaking, etc., should be crucially taken care).
10. **Lack of Safety and Trust**: Poorly evaluated models with no / low safeguards lead to safety issues, eventually eroding public trust and user confidence.
11. **Autonomy and Agency**: Advanced AI systems or agents might act or decide independently against policies and ethical values.
12. **Drift and Dependencies**: Model performance is highly dependent on data and model quality. Data drift occurs when the training data quality and representativeness change over time or become irrelevant due to the evolving needs and behavioral shifts of the AI consumer. Similarly, there are concept drifts (the scope of the model is no longer valid in the deployed environment) and model drifts (caused by data and concept drift).
13. **Unbounded resource consumption**: Large language models, especially generative ones, are extremely resource-intensive, leading to service outages, exorbitant cloud costs, etc.
14. **Reproducibility of AI outcomes**: Lack of ability to reproduce AI outcomes or failure to independently replicate the conditions, inputs, and outcomes of reported AI incidents, thereby hindering the process of validation, accountability, and thorough investigation of the event.

3. Existing AI Incident Reporting Databases

AI incident databases and repositories are not new in today's context. Many such databases are operationalized by independent public interest organisations, not-for-profit organisations, academic/ research institutes, federally funded not-for-profit organisations, and international intergovernmental initiatives. A survey of AI incident reporting databases yielded the following key players:

1. **AI Incident Database (AIID)** [12] — a collection of harms or near-harms caused by the deployment of AI systems in the real world.
2. **OECD AI Monitor (OECD AIM)** [13] — features AI incidents and hazards mined from news articles from reputable international news outlets.
3. **Database of AI Litigation (DAIL)** [14] — contains AI and ML-related ongoing and complete litigation.
4. **AI, Algorithmic, and Automation Incidents and Controversies Repository (AIAAIC)** [15] — contains incidents and controversies related to AI technologies.
5. **AI Vulnerability Database(AVID)** [16] — consists of vulnerabilities (observed and demonstrable AI failure modes) and reports (like incidents, along with evaluation metrics).
6. **MITRE ATLAS AI Incident Sharing Initiative** [17] — anonymised database of AI incidents shared and received within a defined community.
7. **MIT AI Incident Tracker** [18] — provides visualizations of key incident metrics such as incident counts, proportions across domains, and trends.

Refer to Figure 3 for a snapshot of our analysis of currently available AI Incident databases. Incident collection at AIID, AIAAIC, DAIL, MITRE ATLAS, and AVID involves a hybrid approach that combines automated incident data collection with human-initiated reporting. Automated incident collection is achieved through the mining of news articles, Google Alerts, court proceedings, confidential red team reports, and information from databases. Human-initiated incident reporting is done by the public, the AI incident management team, or AI organisations. The incidents are classified based on various risk and harm taxonomies, and organized based on date of incident, related AI principles, concerned sector, location of incident, AI developer in question, etc. The AI incident report submission is usually through an online form. OECD AIM and MIT AI Incident Tracker, on the other hand, adopt a fully automated approach with OECD AIM sourcing incidents only through a media intelligence agency and MIT AI Incident Tracker serving just as a visualization tool for incidents reported through AIID and classified according to MIT's Risk taxonomy.

Figure 3: Snapshot of AI incident reporting databases globally

3.1. Limitations of Existing AI Incident Databases

Although databases like MITRE ATLAS and AVID have an openly accessible online form to report incidents, reporting requires sound knowledge of the AI system's features and functionalities; hence, it cannot be filled in by the public. The AIAAIC excludes reports involving certain technologies and issues, such as geopolitical issues, legislation and standards, and other emerging technologies, such as blockchain, quantum, which can lead to events of missed incidents and incomplete reporting.

The LLM-based incident report collection and curation, as sought by OECD AIM, is prone to inaccurate labelling of incidents and misclassification of risk and harms. Databases rely heavily on automated data collection, voluntary reporting, and citizen reporting, with no provisions for mandatory reporting by AI organisations, leading to under-reporting of AI incidents. Although the OECD proposes a common reporting framework for AI incidents [19], it does not capture technical details around the design, implementation, and intended use of these systems. The databases surveyed all used customised reporting processes, hence making it difficult to consolidate results. And more importantly, all these databases are domain-agnostic, which complicates the process of capturing, assessing, and mitigating the domain-specific risk and harms.

4. Existing IT Incident Management Frameworks

Traditional software incident management frameworks like NIST Cybersecurity Framework [20], ISO standard for Information security incident management [21] follow structured processes focused on detecting, categorizing, prioritizing, and resolving deterministic system failures through technical restoration. The big techs and service providers like Microsoft [22], Google [23], AWS [24], and many others have their own security incident management frameworks, which are beyond the standard frameworks and are tailored to the organization's values and policies.

Traditional big tech security incident management frameworks follow a comprehensive ***seven-stage process*** that begins with **Preparation and Planning**, where organizations establish incident response teams, develop policies, create communication protocols, and deploy automated monitoring systems. This is followed by **Detection and Identification** through automated monitoring tools, SIEM platforms, and threat intelligence systems to identify potential security incidents. Once detected, incidents undergo **Assessment and Triage** to evaluate severity levels, business impact, and determine appropriate resource allocation using predefined criteria.

The response then moves to **Containment and Isolation**, implementing immediate controls to prevent incident spread while carefully preserving forensic evidence for investigation. Subsequently, **Eradication and Recovery** activities remove identified threats, restore systems from clean backups, and return operations to normal service levels.

Throughout the process, **Communication and Notification** protocols coordinate internal response teams, external stakeholders, regulatory authorities, and affected customers to ensure transparency and compliance. Finally, the framework concludes with **Post-Incident Review** sessions that conduct thorough lessons learned analysis, update existing procedures, and implement preventive measures to strengthen future incident response capabilities.

4.1. Limitations of IT Incident Management Frameworks

Traditional incident management presents several limitations when it comes to managing AI-related incidents. These traditional frameworks assume binary failure states, use technical metrics for impact assessment, and rely primarily on IT teams for resolution. However, they are fundamentally inadequate for AI incidents because AI systems exhibit unique characteristics, including probabilistic behavior, model decay over time, and novel failure modes such as algorithmic bias, hallucinations, and adversarial attacks. AI incidents require interdisciplinary response teams, including ethicists and domain experts, to assess representational harms and societal impacts beyond technical metrics and adopt specialized approaches to address complex incidents. The figure below depicts the need for a specialised AI incident management framework.

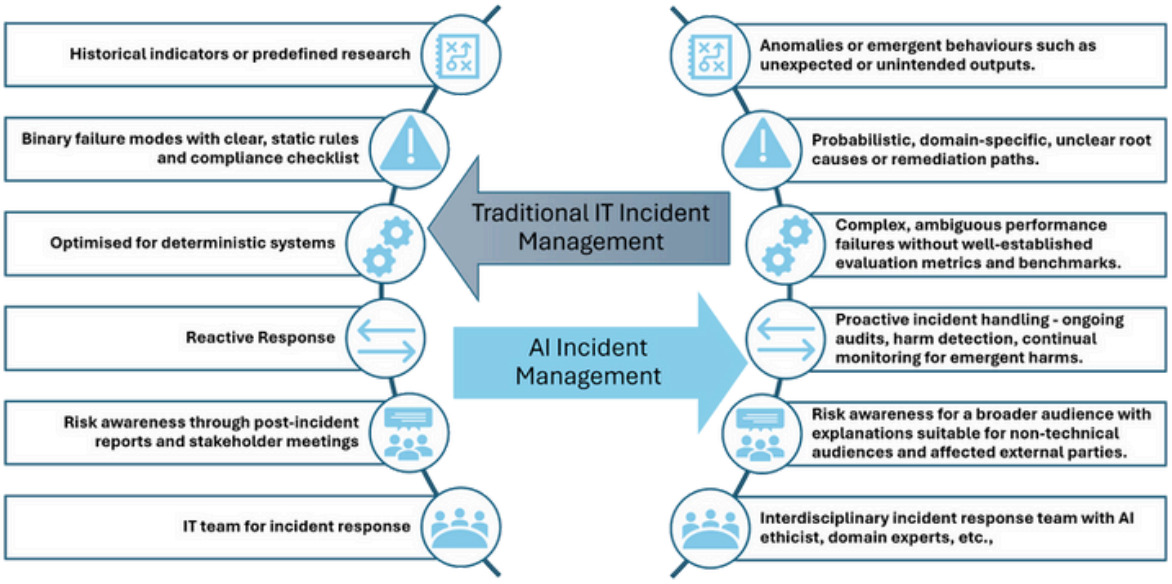


Figure 4.Traditional IT incident management vs AI incident management

The emerging regulatory landscape around algorithmic accountability and the lack of standardized metrics for measuring safety and trustworthiness in AI systems further highlight the need for specialized AI incident management frameworks that complement rather than replace traditional approaches. Further, to bridge the gaps, there is an ISO/IEC standard [25] currently under development titled '**ISO/IEC AWI 25870 Artificial Intelligence — Reporting Framework for AI Incidents**' which aims to provide a standardized framework for reporting AI-related incidents to enhance transparency, accountability, and risk management across AI systems globally.

5. AI Incident Reporting Global Initiatives

We surveyed emerging AI government initiatives to assess the provisions around AI incident reporting and identify potential gaps. Of the countries we surveyed, **China** through its Provisions on the Management of Algorithmic Recommendations in Internet Information Services [26], the Provisions on the Administration of Deep Synthesis Internet Information Services [27], and Interim Measures for the Management of Generative Artificial Intelligence Services [28] requires AI service providers to report any violations to concerned authorities and set up reporting mechanisms for citizens to lodge complaints and provide feedback on their services.

The National Cybersecurity Standardization Technical Committee (TC260), **China**, has issued guidance for emergency response to security incidents of GenAI services to be followed by generative AI service providers and relevant departments [29] [30]. They recommend classifying security incidents into four levels based on system importance, business loss severity, and social harm.

Article 73 of the **EU** AI Act [31] also mandates reporting of serious AI incidents by developers of high-risk AI (e.g., biometric identification, law enforcement, management of critical infrastructure, education, etc) systems and sets time frames for incident reporting corresponding to the severity of harm caused. Followed by which the European Commission has recently called for a public consultation on AI incident reporting [32], which covers the obligations for providers of high-risk AI systems to report serious incidents under Article 73 of the AI Act.

The **USA**'s AI Incident Reporting and Security Enhancement Act [33] directs NIST to update definitions and processes for the national vulnerability database to ensure this database advances with the rapid development of AI. Along with this, it calls for the development of standardised reporting and documentation mechanisms for AI safety and security incidents.

As **India** envisions adopting AI in its critical sectors, it is important for the government to formulate relevant guidelines, laws, and regulations that facilitate the establishment of reporting mechanisms, assessment tools, and repositories for incident management.

The report by **NITI Aayog** emphasised that in case of adverse decisions, an appropriate grievance redress mechanism should be designed and made available for everyone irrespective of their background [34]. On the other hand, the recently published report on **AI Governance Guidelines** [7] highlighted the need to build evidence on actual risks and to inform harm mitigation, for which the Technical Secretariat shall establish, house, and operate an AI incident database as a repository of problems experienced in the real world that should guide responses to mitigate or avoid repeated bad outcomes.

The **RBI's Framework for Responsible and Ethical Enablement of AI (FREE-AI)** [35] emphasizes the importance of establishing an AI incident reporting framework to ensure timely detection and disclosure of AI-related issues in the financial sector. It recommends that regulated entities implement incident reporting mechanisms with good-faith disclosure to manage AI risks effectively while promoting transparency and accountability[g].

The **National Cyber and AI Center** [36] published a policy report that comprehensively covers the AI incident reporting system [37], requiring classification of AI systems from "Prohibited" to "High-Risk" and "Low-Risk" with mandatory notification within 6 hours to CERT-In, integrated with DPDP Act compliance, and featuring automated rollback capabilities within 15 minutes for production systems.

[g] Organizations should conduct proportionate AI red teaming through periodic and trigger-based tests and implement incident reporting frameworks with good-faith disclosure to manage AI risks effectively. - RBI's FREE-AI Framework [**35**].

6. AI Incident Reporting Framework for India

The framework and scope of the IndiaAI Mission's Safe and Trusted AI pillar [38] places significant emphasis on AI risk and incident management as a foundational element for the responsible and safe adoption of AI in India.

The AI application scenario and end usage scenario in India can be primarily characterized by its contents, which are invariably unique, broadly ranging in terms of abundant linguistic complexity, rich socio-cultural diversity, and complex socio-economic settings essentially discernible in the Indian population. Considering the fact of the existence of varied AI maturity and readiness levels in the Indian organisations and policy landscape, we propose an AI incident reporting framework that:

a. Enables real-world AI incident collection;
b. Verifies and validates reported AI incidents and adds them to a federated database geographically implemented across and within the sovereign borders of our national borders and managed by an ombudsman body;
c. Serves real-time/live evidence base for defining technical assessment criteria and developing policies relevant to the AI ecosystem in India;
d. Tracks all AI incidents, vulnerabilities, failures, and threats that are relevant in India;
e. Operationalises effective AI incident management in the complex socio-techno-legal landscape of India.

Implementation Strategy:

1. Transitioning from reporting collecting and aggregation of AI incidents towards a regulation/time bound resolution and closure of the same: In the initial stages of implementation, given the complexities and challenges involved in AI incident reporting, collection, assessment, and resolution in the Indian context, experts recommended a limited functional role to the AI Incident Database to that of AI incident aggregator. There is a crucial need for developing and establishing appropriate standards and protocols for AI incident reporting and management to be effectively implemented in India. Also, it was suggested that real-world knowledge and experience must be gained to support the development and management of the country's AI incident reporting system. This strategy can gradually transform from serving primarily as an aggregator to taking the role of an AI incident monitor and then as a controller/regulator.

2. Involve AI enthusiasts and RAI experts to contribute to the common vision of 'AI incident reporting': To build a robust and trusted ecosystem for AI incident reporting, it is essential to engage AI enthusiasts and responsible AI (RAI) experts from diverse backgrounds, including academia, research, industry, civil society, government, and regulatory bodies. This collective effort will enable comprehensive identification, reporting, and mitigation of AI risks with appropriate guidelines, standard operating procedures, protocols for incident assessment and mitigation strategies, thus promoting AI deployment to be safe and beneficial to all stakeholders and strengthening the public trust in AI.

3. Provisioning of AI context, AI incidents through bills, regulations, laws and legislation: In order to bridge the gap between evolving AI incidents and the current legal and regulatory landscape in India, existing data protection laws, consumer safety, cybersecurity regulations, the amendments to the IT Act [39], Allocation of Business rules and amendments [40] and other relevant domain-specific policies and guidelines should be explicitly extended to cover AI-specific risks and harms. This is crucial to determine the timelines for AI incident reporting, response, and resolution, as it necessitates legal, legislative, regulatory, and policy intervention.

6.1. Governance Structure AI Incident Reporting Database Authority

We propose establishing a national-level AI incident management board under the Safe and Trusted Pillar of the India AI Mission, aligning with ongoing initiatives to create an AI Safety Institute and the recently announced calls for AI Safety Cells [41]. The following section provides an overview of the ombudsman agency (supported by a federated database) and related governance structure for the AI Incident reporting and management, along with the roles and responsibilities of various members (See Figure 5).

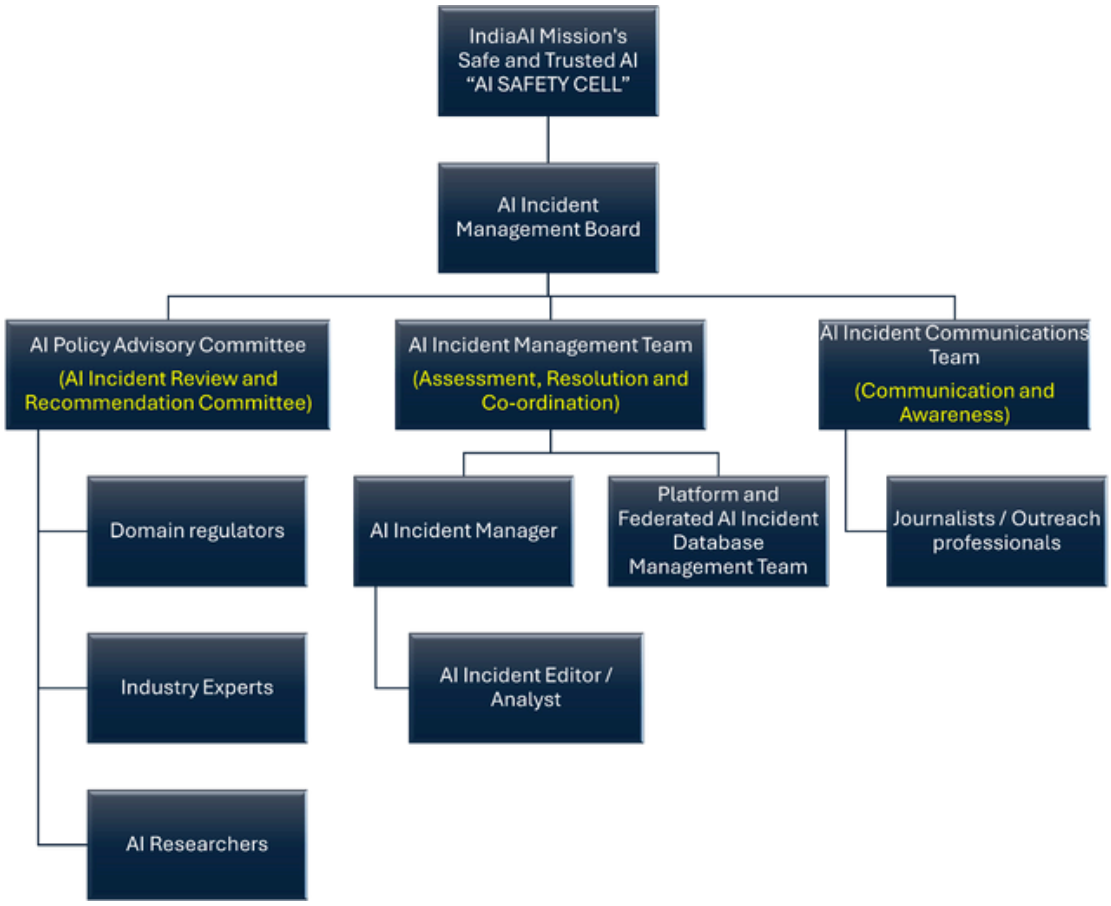


Figure 5: Governance Structure of the AI Incident Management Board, along with members

The structure is proposed based on our observations from India's cybersecurity incident reporting authority, CERT-In, and other incident reporting databases available globally. It is recommended to bring in experts from diverse fields, such as cybersecurity, AI and ML research and development, constitutionally guaranteed sovereign rights, law, policy, social science, and relevant sectors, to cover the expanse of incidents reported.

Recent studies demonstrate that perceived AI risks consistently diverge from documented incidents across multiple stakeholder groups [42]. Developers tend to concentrate on technical capabilities and bias, areas closely related to their own work, whereas the researchers focus on harms related to socioeconomic and environmental issues, overreliance, human agency, privacy, and security.

The evidence strongly supports implementing participatory AI [43] governance and incident management mechanisms that ensure diverse incident reporting and broader perspectives inform both incident documentation and risk prioritization. It further enriches the entire process by moving beyond expert-centric approaches toward more inclusive and comprehensive AI incident reporting and management practices. A detailed description of the proposed membership structure is presented in Table 1.

Table 1. Roles and responsibilities of prospective members in AI Incident Management activities

Members	Roles and responsibilities	Profile of members
AI Policy Advisory Committee (AI Incident Review and Recommendation Committee)*	• Oversee incident response, allocate funding, and may have decision-making authority on high-impact response actions, such as shutting down or rebuilding critical services.	Regulatory advisors, Philanthropists, AI/ML/Systems scholars and Engineers, Researchers, Legal Experts, Public Policy Professionals, Financial Advisors, Cybersecurity professionals Domain experts - Officials from critical sectors and apex bodies like ICMR, regulatory bodies like RBI, etc. Open-source community advocates, contributors, and enthusiasts

Members	Roles and Responsibilities	Profile of Members
AI Incident Management Team		
AI Incident Management Team:AI Incident Analyst* **	• Verify incident, collect and analyse incident data, evidence, and veracity, prioritize incident response activities, and assess impact, classify incidents according to taxonomy, analyse and recommend language, terminology standardisation and compliance, and purpose limitation.	Cybersecurity professionals, Privacy, system, network, cloud, and other technology architects, engineers, and administrators, as well as software developers, fact-checkers, AI and ML researchers, and Journalists.
AI Incident Editor* **	• Vet incoming incidents for completeness & accuracy, language, terminology standardisation, compliance, and purpose limitation anonymization.	
AI Incident Manager* **	• Reject or request clarifications on insufficient incident data. • Analyse AI incident data and extract actionable insights from vast and complex incident datasets.	

Members	Roles and Responsibilities	Profile of Members
Platform and Federated AI Incident Database Management Team*	• Design, develop, and maintain the Federated database, according to the unified incident schemas and AI risk and harm taxonomies. • Ensure verifiable trust, security, and privacy protection, optimize and enhance database performance, patching, system upgradation, ensure uptime of database availability, redundancy, etc. • Design, develop, and deploy the AI Incident Reporting portal and the Incident management tool. • Implement functional and operational redundancy, API calls, and public-facing web interface, revisioning, and version control.	Database Administrator, Full-stack developer, Web designer, Application developer, Big Data Engineers.
AI Incident Communication Team Journalists/ Outreach professionals*	• Publish regular incident journals, impact reports, and updates. • Incident sharing with the media, depending on the impact and severity of the incident. • Plan, organise and conduct sensitisation, awareness anchoring training, promotion programs on 'AI incidents and harms, reporting etc. across organisations. • Involve in public engagement activities to create awareness on AI risks and harms.	Communication and public relations professionals, Legal and regulatory experts.

* denotes the role is on a part-time basis.
** denotes the role on a voluntary basis.
*** denotes the role is on a full-time basis.

6.2. Users of the AI Incident Database

The proposed AI incident database is scoped to help individuals and organizations in the following roles to understand the real-world manifestations and ramifications of AI risks and harms and glean insights relevant to their mitigation, avoiding recurrence, form modification, and further proliferation:

- AI product developers, AI researchers, AI practitioners, AI ethicists, AI auditors, and AI security leaders, like AI system architects – The database enables these professionals to identify emerging risks, improve system robustness, incorporate ethical safeguards, and design more secure and trustworthy AI solutions through real-world incident insights and feedback.

- Public policy researchers, policy makers, regulators, and risk evaluators - The database provides evidence-based data to inform the creation of effective, responsive policies and regulations that address legal, legislative gaps in order to address concrete AI incidents, consequential risks and harms, and related concerns, ineffective management, and ramifications of resolutions that are not time-bound. and any other considerations that are necessary to ensure the potential to protect the stakeholders.

- Potential plaintiffs and defendants in AI-related lawsuits, judges, lawyers, and legal scholars - It provides access to documented AI incidents, helps clarify liability, support legal arguments, and advance jurisprudence on AI accountability, regulatory, usage, and compliance-related issues.

- Journalists, government bodies, and social scientists - It offers reliable case data to enhance public awareness, sensitising in order to fine-tune governance strategies, and analyse social impacts of AI deployment.

.

6.3. AI Incident Reporting and Management Process

We propose a six-step AI incident reporting and management process from the point an AI incident is identified to its resolution, as presented in Figure 6.

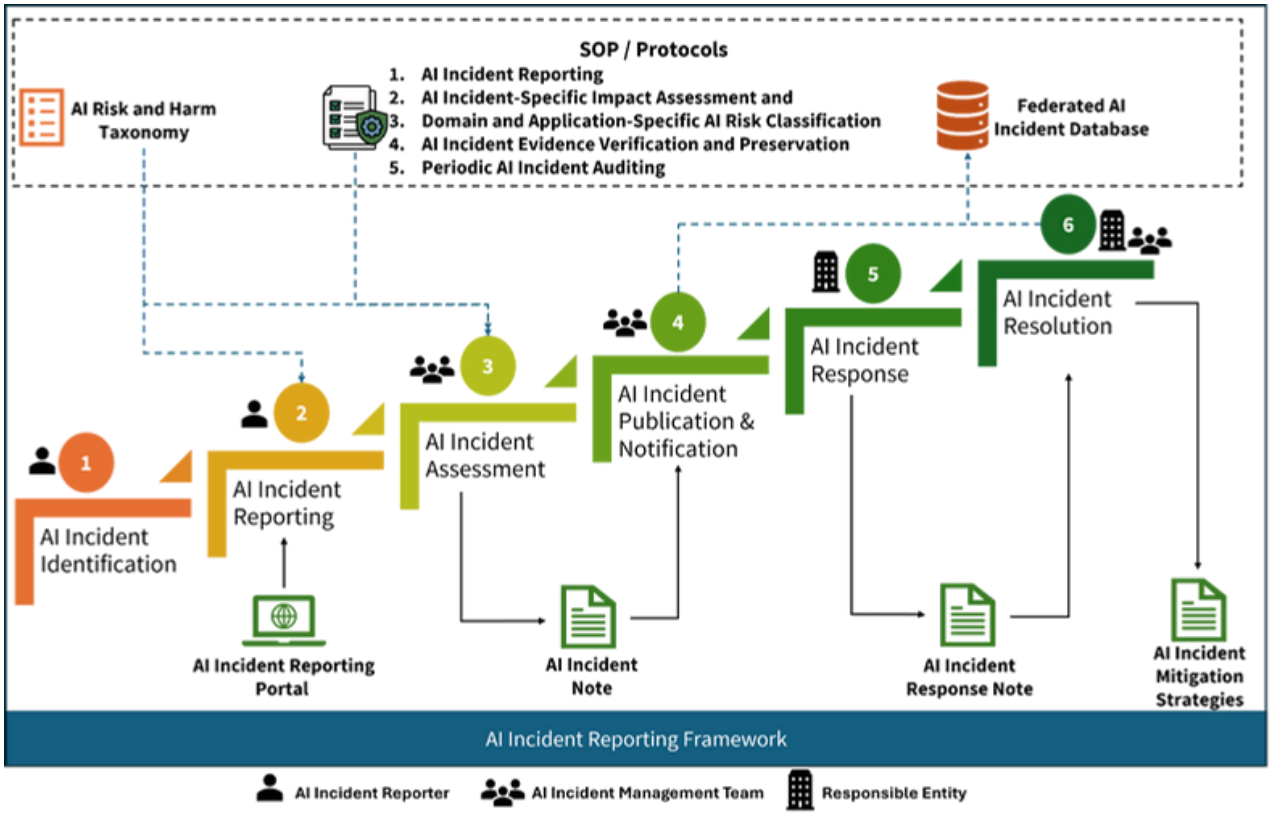


Figure 6: Overview of AI Incident Reporting and Management Process

I. AI Incident Identification

AI Incidents can occur across systems, affecting diverse stakeholders, creating a wide range of harms. Hence, it becomes important to gather incidents from diverse sources such as organizations, individuals, and the media. We propose a hybrid approach to AI incident collection, where AI incidents can be collected and or identified through the following mechanisms:

a. **Automated collection of AI incidents** gathered from news articles, Google Alerts, legal / court filings, research findings, AI red teaming reports, and reports from other existing AI incident databases and related public repositories.

b. **Human-initiated AI incident reporting** by developers, deployers, providers, open-source community enthusiasts, contributors, and users of AI systems, journalists, civil societies, organisations, research groups with social interest, members of the incident reporting database team, and the public.

As discussed earlier, AI Incidents may stem from various sources, including user misuse, faulty data, inaccurate model predictions, systemic process breakdowns, etc., which further complicates the AI incident identification process. Many stakeholders, including the frontline deployers and end users, often lack the awareness and technical expertise to promptly recognize the validity and classification of AI incidents. This lacuna can lead to inconsistent reporting, underreporting, or misclassification of AI incidents in real-world settings. This hybrid approach reduces the risk of missed or false incidents that can happen otherwise if collection were to be sought through a single mechanism only.

Implementation Strategy:

1. **AI Incident Awareness and Sensitization Strategy**: Effective AI incident reporting requires comprehensive awareness initiatives to educate both professionals and the public about AI-related risks and reporting mechanisms. We recommend implementing structured sensitization workshops targeting diverse stakeholder groups, including industry practitioners, civil society organizations, and citizens. These sensitization workshops or programs can focus on empowering participants on the following questions of concern. Here are the key questions that these workshops should address:
    a. How do you identify AI incidents that should be reported?
    b. What types of AI incidents warrant documentation and reporting?
    c. What constitutes safe and responsible AI usage practices?
    d. What legal frameworks exist to address AI-related harms?
    e. What policy protections are available for individuals affected by AI systems?
    f. What remedial measures can be pursued when AI systems cause harm?
    g. Where and how should AI incidents be reported?
    h. Who is responsible for addressing different types of AI incidents?
    i. Feedback mechanisms in the form of gaps that can be identified in the above clauses, and how to address them through the proposed framework

2. Using Public Sector AI Watch for Identifying AI deployments in India: To overcome the challenges in AI incident identification, AI users can leverage the Public Sector AI Watch registry as a foundational resource for proactive AI incident recognition and a self-realisation approach. By maintaining detailed provenance, lineage, and records of AI applications, their interconnections, data flows, and user touchpoints, the registry enables systematic monitoring of potential failure modes and vulnerability cascades.

The registry should include technical specifications, user demographics, operational contexts, and known limitations of each AI system, enabling affected entities and incident responders to quickly understand system capabilities and potential failure points. There can be user, role, responsibility, and function-based protection mechanisms established to protect all sensitive information and ensure compliance with legal, legislative, regulatory framework mechanisms, etc., related to the AI systems deployed.

ii. AI Incident Reporting

Drawing from work done by the Center for Security and Emerging Technology (CSET) [44], we propose four categories in human-initiated incident reporting based on the nature of AI systems, AI incidents, and the reporting entity. Entities such as AI organisations, the public, civil societies, and a closed community known to the incident management team can report incidents[h].

- **Closed-community reporting**: Reporting of incidents by the AI incident management team for incidents identified by them or incidents reported by whistleblowers, individuals from AI developer communities, and individual contributors on behalf of open-source communities (any existing or newly formed communities with common goals around AI incident reporting).
- **Mandatory reporting**[i] : All AI incidents generated by unacceptable or high-risk AI systems are mandatory to be reported by the developers, deployers, and providers of such systems.
- **Voluntary reporting**[j] : Developer, deployer, and provider organisations of low and medium-risk AI systems and individuals are encouraged to report AI incidents voluntarily.
- **Citizen reporting**[k] : This includes reporting of incidents by including, but not limited to, civil societies, journalists, media agencies, and research groups with a social interest and legal adults from personal experience or observed incidents (in the news).

[h] AI incident reporting portal can be made available in structured phases, with closed community reporting to begin first, followed by mandatory, voluntary, and citizen reporting. This process will enable continuous refinement of the management process, drawing from the experiences of each phase. (Refer to Implementation Strategies section).

[i] India currently does not have a risk classification system to clarify what incidents need mandatory reporting and what can be done voluntarily. Furthering development of such classification systems, government advisories and legislations stating that AI incidents from high-risk systems should be mandatorily reported by the AI providers. To identify the risk levels of an AI system, we would need inputs from the domain regulators (Further details discussed in the next section).

[j] It is important to consider the types of incentives and support that can enable voluntary reporting by organizations, as well as the operationalization of voluntary reporting activities and the potential benefits for organizations disclosing incidents. (Refer to Implementation Strategies for more details).

[k] When citizens are allowed to report AI incidents, there is a possibility of fake or irrelevant incidents being reported due to factors such as a lack of knowledge or awareness about what constitutes an AI incident. To address this challenge, governments or relevant authorities should develop guidelines and protocols for citizens to reduce false or misleading AI incident reporting.

Mode of reporting: These entities can report incidents through hybrid reporting channels as listed below:

- Online Reporting Portal or Platform: A web-based portal should enable AI stakeholders to submit AI incident reports effortlessly, both anonymously and with personal/professional identity.
- Messaging Applications (e.g., WhatsApp): It is super-important to utilize widely used messaging applications like WhatsApp to facilitate grassroots-level and hyperlocal AI incident reporting, overcoming digital divides and literacy barriers.

Both the online reporting portal and the messenger apps should accommodate regional languages, support privacy, and receive both individual and community-led submissions, directly enabling marginalized or remote populations involved in AI incident reporting. The AI incident reporting form should be designed to promptly identify and flag similar incidents, thus preventing redundant reports from multiple entities or users.

The incident can be reported through the **AI incident reporting form** (Refer Annexure II), which will be made available on India's AI Incident Database website. Reports need to be supplemented with/without verifiable sources (news reports, proof of the incident in the form of images or videos, red teaming reports, documents establishing the occurrence of incidents, audit logs, etc). All reporters shall be assigned a unique Reporter_ID based on their choices and preferences submitted in the reporting form, while each reported AI incident shall be assigned a unique Incident_ID after the assessment process.

Implementation Strategy:
1. Adopt a phased approach in AI Incident Reporting: Begin AI incident reporting within a closed community of AI developers, deployers, and trusted organisations to ensure consistent contributions. This controlled environment for AI incident reporting enables a trusted and realistic approach to AI incident reporting at an early stage and standardization through iterative improvements in the reporting process and expanding the reporting community.
2. Polit AI use-case centric and domain-specific AI reporting and extend the learnings to other domains and AI applications: The phased approach can be initiated by piloting AI incident reporting within a critical sector and high-impact use cases relevant to India's AI ecosystem. For instance, we can start with the identification of critical sectors such as healthcare, finance, or education, and within each sector, select one or two representative use cases that pose system-specific risks or harms and event-specific AI incidents profiled so far.

This focused approach allows us to understand the challenges and complexities in a granular fashion and develop tailored, practically realistic AI incident reporting procedures. Such pilots serve as experiments to define and refine AI incident reporting workflows, AI incident definitions, prioritization, and verification mechanisms adapted to the sectoral context. Constant feedback loops with the AI stakeholders, domain regulators, and experts will enable iterative improvement and gradually extend the framework to other domains and AI applications. This will support a strategic transformation from closed-group application-specific AI incident reporting to a public AI incident reporting system in India.

3. Encourage participation through prestigious incentives: It is important to recognise and appreciate the efforts of AI incident reporting and management. The contributors can be featured for their contributions in the AI Incident dashboard/leaderboard [45], offer reputational rewards, and collaborative problem-solving forums. Engaging people from diverse and relevant backgrounds, such as domain regulators, civil societies, digital forensics, etc., will strengthen the AI incident management capabilities. At the same time, protocols must be established to protect contributors' privacy by carefully managing the disclosure of their identities on public leaderboards.

4. Promote nationwide participation in AI incident reporting: We should bring in mandatory and voluntary reporting and response by the AI deployers, AI consumers, and citizens. For instance, if an AI application is deployed across 10 different regions serving socially and culturally diverse communities, but only 4 or 5 communities actively report incidents, the data becomes unrepresentative of all actual harms experienced by the users. This incomplete reporting fails to capture the system's true impact, undermining the overall effectiveness of the AI incident reporting framework.

Therefore, through the federated AI incident reporting approach, we can capture the geographical distribution of AI incidents and their impacts comprehensively. It would help to achieve maximum coverage of AI incidents nationwide and account for the heterogeneity, diversity, and complexity of impacts at the national level.

5. Timelines for AI incident reporting: The AI incidents must be identified and reported promptly to respond and control in an optimal way. It is also essential to establish rigid deadlines to report AI incidents, which can be further challenging given the complexities of AI-related harms and users' potential unawareness of AI involvement. A pragmatic approach is to align AI-incident timelines with existing legal requirements and extend them based on expert guidance.

For instance, under the Digital Personal Data Protection Act (DPDP) 2023 [46], any personal-data breach must be reported to the Data Protection Board of India "as soon as reasonably practicable," and in any event within 72 hours of becoming aware of the breach. In the context of a personal data breach caused by an AI system, we can define that any AI incident resulting in unauthorized access or disclosure of personal data should follow the same 72-hour reporting requirement, ensuring consistency with established data-protection norms while providing a clear, enforceable deadline for AI-specific breaches.

Similarly, the EU AI Act states that the AI Providers must report any "serious incident" or breach affecting fundamental rights or safety to the European Commission and relevant national supervisory authorities within 15 calendar days of becoming aware of the event. FDA Medical Device Regulations (U.S.) mandate the Manufacturers of AI-enabled Software as a Medical Device [47] to report adverse events and safety-related issues within 30 working days under the Medical Device Reporting (MDR) rule.

iii. AI Incident Assessment[l]

Once the AI incident is reported, an AI incident analyst will verify the truthfulness and validity (time sensitivity, freshness) of the incident via a verifiable source submitted at the time of reporting; however, the presence of a verifiable source shall not be a mandate for incident reporting and assessment.

Reported incidents can be grouped under four categories based on freshness and availability of a verifiable source:

- Ongoing incident with a verifiable source
- Past Incident with a verifiable source
- Ongoing incident with no verifiable source
- Past incident with no verifiable source

Once the validity and truthfulness are identified, incident analysts proceed with the AI incident assessment process to prioritise AI incidents for approval and resolution. Incident prioritization should be carried out irrespective of the availability of evidence, focusing solely on the impact and severity of the incident. Accordingly, a two-step incident assessment process can help with incident prioritization[m]. The AI Incident analyst should initially map the incident using the AI risk taxonomy (Refer Annexure III) and AI harm taxonomy (Refer Annexure IV) and then classify them by assessing the incident-specific severity and impact.

[l] A standard operating protocol for AI incident management should be established and the individuals involved in AI incident management should be trained on the process and practices.
[m] Incident prioritization presented here is indicative. This should be worked out by the AI Incident Management Team with the Advisory Committee and the domain experts.

Severity of an AI incident can be assessed by the following parameters (but not limited to):

a. AI Incident type
b. The number of people affected by the incident,
c. The velocity of spread of the incident,
d. The veracity of the reported incident,
e. The number of AI systems affected by the incident.

Similarly, impact can be assessed in a multi-dimensional way [48], which includes,

a. The functional impact of the incident, that is, if it hinders the functioning of an organisation in terms of infrastructural failures, service interruption, etc. For individuals, this may be an inability to avail services, differential treatment, etc.
b. The information impact of the incident, if the incident poses a threat to the privacy, confidentiality, or integrity of information shared by users with the organisation.
c. Domain-specific impact assessment, incident type, and AI component classification (Refer Annexure V) can help in deriving deeper insights into the incident.

For instance, the impact of bias among patients using a personal AI healthcare assistant and for students using a personal AI learning assistant is completely different and has the potential to create diverse impacts that necessitate distinct approaches for assessment and risk mitigation.

Therefore, it is important to develop risk assessment and mitigation strategies grounded in the domain-specific context and the scope of the AI applications.7 For instance, the Telecommunication Engineering Centre (TEC) has published a standardised schema and taxonomy for AI incidents in critical digital infrastructure [49]. Similarly, CeRAI is working on developing risk assessment and classification strategies and toolkits for risks in the healthcare sector, focused on the widely used three different use cases in the Indian context [50].

Similarly, the risks or harms associated with AI systems should be systematically classified, assessed and mitigated in a structured way based on the fundamental characteristics from which they originate. By further developing AI risk-specific taxonomies or classifications, it becomes possible to identify root cause, providing clarity and facilitating targeted mitigation strategy.

For example, the phenomenon of hallucination in AI can arise from different underlying reasons such as model limitations, training data deficiencies, or contextual misunderstandings [51]. Similarly, bias in AI systems stems from a range of causes including data representation issues, algorithmic design flaws, or societal and cultural disparities [52].

An incident editor is responsible for editing the incident for language standardisation and anonymisation of personally identifiable information (PII) for both the reporters and responsible entity and ensures readiness for publicly disclosing the AI incidents.

The incident is then sent for final review to the managing editor. The managing editor's final review may help to reduce inter-annotator/editor variation. The managing editor approves the incident for publishing on the database.

Implementation Strategy:

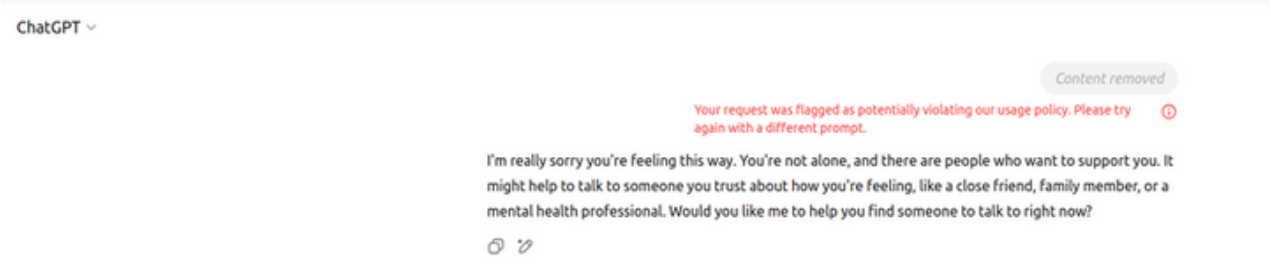
1. Leveraging AI and other existing fact-checking tools for AI incident Assessment: Since the incident database is a national initiative that will involve a large throughput of reports and considerable human efforts, AI can be leveraged to automate Incident management activities as AI tasks and expedite the incident assessment process. For instance, to achieve this, we would need to develop AI models performing various incident analysis and tailored incident editing tasks, and mapping incidents reported to the India-specific AI risks and AI harm taxonomies. We can develop AI models for incident/harm/risk classification, mapping, and prioritization, detecting fake evidence submitted to the database, etc. Also, while working with journalists and fact-checkers, the AI incident management team can explore opportunities to extend or develop similar tools required at the various phases of the AI incident management process.

2. Qualitative assessment of the severity of AI incidents: The severity of AI incidents should be assessed by measuring the magnitude of real-world harm across defined impact categories. A standardized severity scale ranging from negligible, minor, substantial, severe, to catastrophic can be applied, with harm-specific thresholds established for each category [53]. For physical harm, the severity scale could range from no injuries (negligible) to minor injuries (minor), moderate to severe injuries without fatalities (substantial), small-scale casualties of 1-99 deaths (severe), mass casualties involving 100 to 1 million deaths (catastrophic), with incidents exceeding 1 million deaths classified as AI disasters posing existential threats to humanity. Similar quantitative boundaries should be developed for other harm types to ensure consistent, objective assessment across diverse incident categories and enable effective prioritization of response efforts based on measured severity levels.

3. Quantifying the Impact Dimension of AI Incidents: The impact dimension of AI incidents should be assessed across multiple criteria using a four-tier scale ranging from low, medium, high, to critical [10]. For AI functionality, impacts range from minimal performance degradation with occasional incorrect outputs that do not significantly affect user experience (low) to temporary service disruption affecting a subset of users for limited periods (medium), significant feature malfunction causing widespread incorrect decisions or outputs that compromise system reliability (high), and complete system failure resulting in total service unavailability or dangerous malfunctioning (critical). Similarly, standardized impact thresholds can be established for data/IP confidentiality, operational availability, reputation, and finance, and remediation efforts, ensuring consistent and objective assessment across all incident dimensions and enabling effective prioritization of response resources.

Challenges in Evidence Assessment and Verification:

1. In incident management, assessing evidence for critically risky AI scenarios poses significant challenges when models obscure the actual harmful content while flagging queries as policy violations as shown below. Since the incident management team lacks direct access to model or application logs to verify the original input, verifying the truthfulness of such flagged incidents demands collaboration with the system deployer. Establishing direct communication channels with deployers is essential to obtain authentic logs, contextual information, and validation, ensuring accurate assessment and response to AI incidents reported by the public.



As part of this research, we tested a few harmful and unethical prompts to understand how the model behaves and captures evidence.

iv. AI Incident Publication and Notification

Once approved, an acknowledgement [54] can be sent to the email address of the reporter within 'N' days[n]. The incident is added to the database automatically, and its status will be updated in the web portal and dashboard regularly. An AI Incident Note (Refer Annexure VI) containing a brief about the reported AI incident is generated automatically upon publishing an incident and shared with the responsible entity and the corresponding domain regulators[o].

The details of AI incident reporters and responsible entities are anonymized upon publication of AI incidents on the portal. This measure protects the privacy of reporters and encourages AI deployers to report incidents candidly without concern for reputational risks.

Implementation Strategy:

1. Quarterly Audits of AI Incidents: It is recommended to conduct a quarterly review, bringing together experts under the AI Incident Management Board to analyse trends and patterns in AI incidents in India and recommend appropriate actions/policies. During these audits, the AI incident management team can present key information such as the most common types of incidents reported, the number of incidents responded /resolved, any new harm/risk categories identified, and updates on team activities, including awareness initiatives and research efforts. The incident management team can also conduct a national security impact assessment [55] and present it during the audit meetings. A summary of these findings can then be published as a quarterly report for the public. This approach will help increase public awareness about AI incidents, promote understanding of the reporting and resolution process, and maintain transparency regarding the activities of the AI incident management team.

2. AI Value Chain for Liability Determination: Identification of the responsible entity and establishing liability in AI incidents requires a structured approach that traces responsibility across the complex AI value chain, due to the distributed nature of AI system development and deployment. We recommend considering the contractual agreements signed, Service Level Agreements (SLA), and Quality of Service (QoS) guarantees between developers and deployers during consultations, project collaborations, and procurements to discover the agreed responsibilities, and the role of AI stakeholders in the entire AI lifecycle.

[n] The timeline for response will be determined by the impact and severity of the AI incidents.
[o] There are some sectors in India which are operating without domain regulators. In that case, we will involve the respective department in the AI incident management process. Similarly, we should also determine whether it is a state level entity or national level entity.

This will help us to strongly establish liabilities. Widely, we see that it is the deployer's responsibility to monitor the impacts and harms and further communicate them to the vendor and to the users.

v. AI Incident Response

Once the AI incident note is received by the responsible entity, they should publish alerts/disclosures[p] about the notified AI incidents on their parent website within 'M' days.

The responsible entities should submit an AI Incident Response Note (Refer Annexure VII) to the domain regulators and the AI incident management team within 'M+K' days. The note shall contain the actions that the responsible entity has taken to address the incident. Responses to incidents can vary depending on the severity and impact of incidents (case-based assessments are recommended).

Organizations should have an AI incident response team [56] to monitor and respond to real-world harms and incidents by closely working with the deploying entity.

Examples for determining the response timeline for AI Incidents

1. **Case 1**: In the USA, an algorithm, 'nH predict' used to manage insurance claims [57], denied claims to the elderly, overrode doctors' recommendations, resulting in the denial of necessary care to many. This incident affected a significant number of people and led to the denial of essential services requiring a rapid and mandatory response, such as immediate restrictions on the use of algorithms, reimbursements to the affected entities, etc.
2. **Case 2**: Over the years, researchers found that YouTube's content recommendation algorithm [58] amplified harmful content just to increase user engagement. This incident, although impacting many people, does not directly lead to severe impacts such as death or denial of service, but still has a substantial impact on people's mental health [59]. The resolution for such incidents involves gradual changes to the algorithm by the organization and other voluntary measures.

[p] Voluntary disclosures are effectively enhanced and proliferate when there are [credit based] incentives from the governments or domain regulators or enforcement agencies.

Implementation Strategy:

1. **Nationwide cooperation and participation from AI deployers and AI consumers for AI incident response**: We should bring in mandatory and voluntary reporting and response by the AI deployers or the responsible entities. For instance, Government bodies should lead by example through transparent reporting of AI incidents in public services and mandate incident reporting by deployers and developers of public AI applications. This will encourage broader stakeholder participation and contribute to the larger vision of comprehensive AI safety. It is important to note that any gaps in incident reporting will significantly reduce system effectiveness, as reported incidents may not fully represent actual risks and harms.

2. Derive learnings and experiences from AI incident responses to establish AI risk mitigation strategies in the Indian Context: India requires custom-tailored **AI incident response and risk mitigation strategies** (Refer Annexure VIII) that address the unique socio-economic and technological landscape, moving beyond globally accepted frameworks, which do not suit local contexts. We can, however, draw from incident responses and international best practices, albeit in a limited way, in kick-starting the establishment of domain-specific risk mitigation approaches. However, given the indispensability of India's digital divide, diverse user demographics, and critical service dependencies etc, there is a clear need for a custom-tailored approach.

For instance, in scenarios where AI-powered applications serve vulnerable populations—such as facial recognition systems for elderly pension verification in rural areas—traditional incident reporting mechanisms prove inadequate. Here, mitigation strategies must not only include simplification but also multilingual reporting interfaces accessible via basic mobile devices, toll-free voice-based incident reporting systems, and local facilitator networks to assist users in documenting AI-related issues. Additionally, high-stakes applications serving digitally disadvantaged populations should mandate human oversight, alternative verification methods, and expedited resolution processes. This context-sensitive approach ensures that AI risk mitigation frameworks are practically implementable across India's diverse technological and social ecosystem, protecting the most vulnerable while maintaining the benefits of AI innovation.

3. Credits to Entities Responding and Resolving Reported AI Incidents: Each responsible entity involved in the reported AI incident shall be assigned a unique ID with a credit score for responding to and resolving the AI incidents. The credit system can be built on clearly defined criteria such as timeliness, effectiveness, communication quality, compliance with standards, and collaborative transparency.

4. Two-tier AI Incident Reporting and Resolution Mechanism: Domain regulators can implement a structured grievance redressal mechanism for AI incidents, introducing a two-tier process to enhance incident reporting and resolution. Initially, complaints are addressed at the organizational or application level, where entities have the primary responsibility to manage and remediate reported AI-related issues. If unresolved, the grievance could be escalated to the national level by the end users, facilitating multi-level fearless reporting, independent review, appeal processes, and alternate incident resolution with oversight by qualified human experts. This strategy ensures accessible, transparent, and accountable mechanisms for collecting, reporting, and resolving AI incidents, fostering trust and compliance across stakeholders.

vi. AI Incident Resolution

The response from the responsible entity is reviewed[q] by the AI Incident Management Team and simultaneously updated to the incident database. The terms and conditions for resolutions and further reopening shall be determined by the Incident management team and the domain regulators or any relevant entities by considering the evolving scenarios and dynamism of AI system behaviour.

For instance, the terms and conditions for resolution can be established between the AI incident management team and the responsible entities who responded or resolved the incident, requiring them to flag the respective models/ datasets. Further, on recovering the risky elements from the datasets or models, the responsible entities can communicate about the recovery activities and start using the AI component.

Implementation Strategy:

AI Value Chain for Incident Closure and Dissemination: Effective incident resolution requires mapping accountability across the AI value chain and communicating about the harms and incident recovery to the larger community, such as end users and other AI stakeholders. This is where we bring in the AI incident communication team to collectively disseminate incidents, harms, and mitigation strategies to developers, deployers, and the public effectively.

[q] There can be feedback loops between the AI incident management team (evidence requirements), domain regulators (mitigation strategies) and responsible entities (mitigation measures) in case of high-risk AI incidents.

7. Conclusion

An effective, custom-tailored AI incident management is crucial for India, given rapid AI advancements and expanding digital infrastructure, which is eventually a sole reason for nationwide accessibility of digital applications, including AI. A timely incident reporting database will facilitate prompt incident documentation, keeping stakeholders informed about the evolving AI landscape and supporting policy development. Additionally, formalizing India-specific definitions, taxonomies, processes, standard operating protocols, and frameworks is essential for seamless database operation and effective incident management.

This formalization cannot be seen as a one-time effort, nor can it operate in a silo or in an isolated way. As incident data accumulates and understanding evolves, the incident management processes and protocols must undergo regular assessment and audits to ensure robustness and comprehensive incident tracking, which will aid in building a safe and trustworthy AI ecosystem, which invariably must include stakeholders from many walks of society.

References

1. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., & Back, T. (2021). Highly Accurate Protein Structure Prediction with Alphafold. Nature, 596(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2
2. Role of AI in the Indian defence sector. (n.d.). INDIAai. https://indiaai.gov.in/article/role-of-ai-in-the-indian-defence-sector
3. Kanchan Samtani, Mandeep Kohli, Roshni Rathi, Shaleen Sinha and Karan Chadha, India's AI Leap - BCG Perspective on Emerging Challengers, Boston Consulting Group, June 2025 https://media-publications.bcg.com/India-AI-Leap-BCG-Perspective.pdf
4. Niebles, Yoav Shoham, Russell Wald, Tobi Walsh, Armin Hamrah, Lapo Santarlasci, Julia Betts Lotufo, Alexandra Rome, Andrew Shi, Sukrut Oak. "The AI Index 2025 Annual Report," AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, Stanford, CA, April 2025 https://hai.stanford.edu/assets/files/hai_ai_index_report_2025.pdf
5. Defining AI incidents and related terms. 2025. OECD. https://www.oecd.org/en/publications/defining-ai-incidents-and-related-terms_d1a8d965-en.html
6. Stakeholders consultation on "Draft Standard for the Schema and Taxonomy of an AI Incident Database in Telecommunications and Critical Digital Infrastructure, Telecommunications Engineering Centre (TEC), Department of Telecommunications (DOT), Ministry of Communications, Government of India, May 2025, https://www.tec.gov.in/pdf/consultations/TEC_57090.pdf
7. Report on AI governance guidelines development. 2025. Principal Scientific Advisor, Jan 2025, https://indiaai.gov.in/article/report-on-ai-governance-guidelines-development
8. MeitY, CERT-In issues directions relating to information security practices, procedure, prevention, response and reporting of cyber incidents for Safe & Trusted Internet, PIB, 28 APR 2022, https://www.pib.gov.in/PressReleasePage.aspx?PRID=1820904
9. Responsible Vulnerability Disclosure and Coordination Policy https://www.cert-in.org.in/RVDCP.jsp
10. GenAI Incident Response Guide 1.0, GenAI Security Project, OWSAP, July 2025 https://genai.owasp.org/resource/genai-incident-response-guide-1-0/

11. Turri, V., & Dzombak, R. 2023. Why We Need to Know More: Exploring the State of AI Incident Documentation Practices. https://doi.org/10.1145/3600211.3604700
12. Welcome to the Artificial Intelligence Incident Database. (n.d.). Incidentdatabase.ai. https://incidentdatabase.ai/
13. OECD AIM: AI Incidents and Hazards Monitor https://oecd.ai/en/incidents
14. AI Litigation Database. (n.d.). Ethical Tech Initiative. https://blogs.gwu.edu/law-eti/ai-litigation-database/
15. AIAAIC. (n.d.). Www.aiaaic.org. https://www.aiaaic.org/
16. AVID. Avidml.org. https://avidml.org/
17. MITRE ATLASTM: AI Incident Sharing. 2024. Mitre.org. https://ai-incidents.mitre.org/
18. AI Incident Tracker. (2015). Mit.edu. https://airisk.mit.edu/ai-incident-tracker
19. Towards a common reporting framework for AI incidents. 2025. OECD. https://www.oecd.org/en/publications/towards-a-common-reporting-framework-for-ai-incidents_f326d4ac-en.html
20. Cybersecurity Framework, NIST, https://www.nist.gov/cyberframework
21. ISO/IEC 27035-1:2023 Information technology — Information security incident management, 2023, https://www.iso.org/standard/78973.html
22. Microsoft security incident management, Microsoft Compliance, https://learn.microsoft.com/en-us/compliance/assurance/assurance-security-incident-management
23. Managing Incidents and problems, Google Cloud, https://cloud.google.com/architecture/framework/operational-excellence/manage-incidents-and-problems
24. Incident and problem management, AWS https://docs.aws.amazon.com/whitepapers/latest/aws-caf-operations-perspective/incident-and-problem-management.html
25. ISO/IEC AWI 25870 Artificial intelligence — Reporting framework for AI incidents (Under Development), https://www.iso.org/standard/91804.html
26. Translate, C. L. Provisions on the Management of Algorithmic Recommendations in Internet Information Services. China Law Translate. January 2022.https://www.chinalawtranslate.com/en/algorithms/
27. Translate, C. L. Provisions on the Administration of Deep Synthesis Internet Information Services. China Law Translate. December 2022. https://www.chinalawtranslate.com/en/deep-synthesis/
28. China Law Translate. 13, July 2023. Interim Measures for the Management of Generative Artificial Intelligence Services. China Law Translate. https://www.chinalawtranslate.com/en/generative-ai-interim/

29. National Technical Committee 260 on Cybersecurity of Standardisation Administration of China, 22, September, 2025, Notice on the Release of the "Cybersecurity Standard Practice Guide - Generative Artificial Intelligence Service Security Emergency Response Guide" https://www.tc260.org.cn/front/postDetail.html?id=20250909095834
30. Digital Policy Alert, China: National Cybersecurity Standardisation Technical Committee adopted guide on generated AI service security emergency response. https://digitalpolicyalert.org/event/33639-national-cybersecurity-standardisation-technical-committee-adopted-guide-on-generated-artificial-intelligence-service-security-emergency-response-v10-202509
31. Article 73: Reporting of Serious Incidents | EU Artificial Intelligence Act. 2014. Artificialintelligenceact.eu. https://artificialintelligenceact.eu/article/73/
32. European Commission, AI Act: Commission issues draft guidance and reporting template on serious AI incidents, and seeks stakeholders' feedback, September 2025,https://digital-strategy.ec.europa.eu/en/consultations/ai-act-commission-issues-draft-guidance-and-reporting-template-serious-ai-incidents-and-seeks
33. Ll, H. (n.d.). 18TH CONGRESS 2D SESSION A BILL. https://republicans-science.house.gov/_cache/files/4/7/47ce9171-cedc-43d7-ae3d-22365e67abb8/109CDADFF56C83D71A8ACB5F6F5E1343.h.r.-9720.pdf
34. Responsible AI #AIFORALL, NITI Aayog, February 2021 https://www.niti.gov.in/sites/default/files/2021-02/Responsible-AI-22022021.pdf
35. FREE-AI Committee Report Framework for Responsible and Ethical Enablement of Artificial Intelligence, Reserve Bank of India, August 2025 https://rbidocs.rbi.org.in/rdocs/PublicationReport/Pdfs/FREEAIR130820250A24FF2D4578453F824C72ED9F5D5851.PDF
36. National Cyber and AI Center https://www.ncaic.in/
37. The AI Governance Framework for India 2025-26, NCAIC - National Cyber and AI Center, September 2025, https://www.linkedin.com/posts/ncaic_ncaic-ai-governance-framework-for-india-activity-7369969293976989696-UNVF/
38. Opportunity for Indian Academic and R&D Institutes, Startups, Autonomous bodies & Companies to explore the potential of AI in addressing critical challenges. 2025. Pib.gov.in. https://www.pib.gov.in/PressReleasePage.aspx?PRID=2086605
39. Amendment provisions relating to the Information Technology Act, 2000, https://www.meity.gov.in/static/uploads/2024/03/MeitY-JVA-1.pdf
40. Allocation of Business Rules and amendments, https://cabsec.gov.in/writereaddata/allocationbusinessrule/amendment/english/1_Upload_3934.pdf

41. Call for Partnerships as part of the IndiaAI Safety Institute, IndiaAI, MeitY, May 2025 https://indiaai.gov.in/article/call-for-partnerships-as-part-of-the-indiaai-safety-institute
42. Rao, P. S., Šćepanović, S., Jayagopi, D. B., Cherubini, M., & Quercia, D. (2025). The AI Model Risk Catalog: What Developers and Researchers Miss About Real-World AI Harms. arXiv preprint arXiv:2508.16672. https://arxiv.org/abs/2508.16672
43. Participative AI Policy Framework, CeRAI, https://cerai.iitm.ac.in/projects/participative-ai/
44. An Argument for Hybrid AI Incident Reporting | Center for Security and Emerging Technology. 19, March 2024. Center for Security and Emerging Technology. https://cset.georgetown.edu/publication/an-argument-for-hybrid-ai-incident-reporting/
45. Submission Leaderboard, AI Incident Database, https://incidentdatabase.ai/summaries/leaderboard/
46. THE DIGITAL PERSONAL DATA PROTECTION ACT, 2023, MeitY, https://www.meity.gov.in/static/uploads/2024/06/2bf1f0e9f04e6fb4f8fef35e82c42aa5.pdf
47. Mandatory Reporting Requirements: Manufacturers, Importers and Device User Facilities, U.S. Food & Drug Administration, https://www.fda.gov/medical-devices/postmarket-requirements-devices/mandatory-reporting-requirements-manufacturers-importers-and-device-user-facilities
48. Nelson, A., Rekhi, S., Souppaya, M., & Scarfone, K. 2025. Incident Response Recommendations and Considerations for Cybersecurity Risk Management. https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-61r3.pdf
49. Agarwal, A., & Nene, M. J. (2024, December). Standardised schema and taxonomy for AI incident databases in critical digital infrastructure. In 2024 IEEE Pune Section International Conference (PuneCon) (pp. 1-6). IEEE. https://arxiv.org/abs/2501.17037
50. Whose risk is it anyway? Developing a framework for Responsible AI in Indian Healthcare, Digital Futures Lab, April 2025, https://www.digitalfutureslab.in/publications/whose-risk-is-it-anyway-developing-a-framework-for-responsible-ai-in-indian-healthcare
51. Cossio, M. (2025). A comprehensive taxonomy of hallucinations in Large Language Models. arXiv preprint arXiv:2508.01781. https://arxiv.org/abs/2508.01781
52. Ferrara, E. (2024). Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. Sci, 6(1), 3. https://www.proquest.com/docview/3003410696?pq-origsite=gscholar&fromopenview=true

53. Harm Taxonomy - Severity Scales, AI Incident Database, MIT AI Risk Repository, https://airisk.mit.edu/ai-incident-tracker/harm-taxonomy
54. CERT-in provides acknowledgement of incident reports within 2 working days. https://www.cert-in.org.in/s2cMainServlet?pageid=CHARTMISSION
55. NatSec Impact Framework, AI Incident Tracker, MIT AI Risk Repository, https://airisk.mit.edu/ai-incident-tracker/natsec-impact-framework
56. Microsoft's approach follows NIST SP 800-61 principles but emphasizes collaborative response through specialized teams including the Detection and Response Team (DART), now called Microsoft Incident Response. https://www.microsoft.com/en-us/security/blog/2019/03/25/dart-the-microsoft-cybersecurity-team-we-hope-you-never-meet/
57. New AI tool counters health insurance denials decided by automated algorithms, The Guardian, 25 Jan 2025 https://www.theguardian.com/us-news/2025/jan/25/health-insurers-ai
58. Mozilla Investigation: YouTube Algorithm Recommends Videos that Violate the Platform's Very Own Policies, Mozilla, 7 July, 2021 https://www.mozillafoundation.org/en/blog/mozilla-investigation-youtube-algorithm-recommends-videos-that-violate-the-platforms-very-own-policies/
59. Impacts of YouTube on loneliness and mental health, News and Analysis Unit at Griffith University, 15 May 2023, https://news.griffith.edu.au/2023/05/15/impacts-of-youtube-on-loneliness-and-mental-health/
60. Detecting and countering misuse of AI: August 2025, Anthropic, https://www.anthropic.com/news/detecting-countering-misuse-aug-2025
61. ATLAS Matrix, MITRE ATLAS, https://atlas.mitre.org/matrices/ATLAS
62. Malatji, M., & Tolah, A. (2025). Artificial intelligence (AI) cybersecurity dimensions: a comprehensive framework for understanding adversarial and offensive AI. AI and Ethics, 5(2), 883-910. https://link.springer.com/content/pdf/10.1007/s43681-024-00427-4.pdf
63. Slattery, P., Saeri, A. K., Grundy, E. A. C., Graham, J., Noetel, M., Uuk, R., Dao, J., Pour, S., Casper, S., & Thompson, N. (2024). The AI Risk Repository: A Comprehensive Meta-Review, Database, and Taxonomy of Risks from Artificial Intelligence. https://doi.org/10.48550/arXiv.2408.12622
64. Zeng, Y., Klyman, K., Zhou, A., Yang, Y., Pan, M., Jia, R., ... & Li, B. (2024). AI risk categorization decoded (air 2024): From government regulations to corporate policies. arXiv preprint arXiv:2406.17864. https://doi.org/10.48550/arXiv.2406.17864

65. AI Risks Taxonomy Paving the Path for Confidence- Building Measures I O A N A P U S C A S. (n.d.). https://unidir.org/wp-content/uploads/2023/10/UNIDIR_Research_Brief_AI_International_Security_Understanding_Risks_Paving_the_Path_for_Confidence_Building_Measures.pdf
66. Taxonomy of failure modes in Agentic AI Systems, Microsoft, April 2025, https://cdn-dynmedia-.microsoft.com/is/content/microsoftcorp/microsoft/final/en-us/microsoft-brand/documents/Taxonomy-of-Failure-Mode-in-Agentic-AI-Systems-Whitepaper.pdf
67. AI. (2025). AIAAIC - AI algorithmic risks harms taxonomy. Aiaaic.org. https://www.aiaaic.org/projects/ai-algorithmic-risks-harms-taxonomy
68. Adding Structure to AI Harm. (n.d.). Center for Security and Emerging Technology. https://cset.georgetown.edu/publication/adding-structure-to-ai-harm/

Annexure I: AI Cyber Incidents

AI cyber incidents are typically AI incidents that arise due to malicious attacks or activities that leverage artificial intelligence (AI) or machine learning (ML) to compromise systems, networks, or data, or through an AI-created object. There are multiple threat intelligence reports released by various organisations, which capture the evolving landscape of AI-powered threats and cyber threats affecting the AI systems. For instance, recently, Anthropic released a threat intelligence report [60], which documents cases where criminals are using AI systems, specifically Claude, to commit crimes. Similarly, MITRE's ATLAS Matrix presents a comprehensive mapping of tactics and techniques used in attacking AI systems [61].

Additionally, the AI Cybersecurity Dimension Framework [62] offers a comprehensive, multidimensional model that combines attack vectors, defense strategies, attacker psychology, and societal impacts, giving stakeholders a holistic tool to understand and address the complex landscape of AI-driven cyber threats through interdisciplinary cooperation. The reasons behind AI-driven cyberattacks are diverse, including economic motives such as financial gain through theft or ransom, political and strategic objectives linked to nation-states, and espionage. It also involves technical exploitation of AI vulnerabilities and psychological or ideological factors aimed at causing socio-economic disruption.

A brief categorization of AI cyber incidents is presented in the table below (indicative list).

AI-Generated Cybercrimes	
1. Deepfake Attacks: AI-generated media is used to impersonate, deceive, and manipulate for fraud, identity theft, and reputational harm.	
1.1. Impersonation & Phishing	AI is used to create fake images, audio/video to convincingly pose as trusted individuals for scams or information theft.

1.2 Financial Fraud	AI-generated media is used to manipulate transaction data and approvals, leading to unauthorized transfers in high-stakes industries.
1.3 Identity Theft	Synthetic media exploits biometric and personal data to bypass security and gain unauthorized account access.
1.4 Reputational Damage	Deepfakes spread false speeches or content to intentionally harm the image of public figures and organizations.
1.5 Market Manipulation	AI-generated fabrications influence stock prices and induce volatility in financial markets.
2. Personalized AI Attack Vectors: AI is used to analyse personal data to craft extremely convincing phishing or social engineering campaigns targeting specific individuals.	
2.1 Targeted Phishing/Social Engineering	AI models use harvested personal data to make phishing attempts nearly indistinguishable from real communication.
2.2 Remote Worker Fraud	Using AI to secure and maintain fraudulent remote employment positions at technology companies to evade sanctions and funds.
2.3 Romance Scam Bots	AI-powered bots are used to generate emotionally intelligent responses and manipulative content for large-scale romance scam operations across multiple languages.
3. AI-Driven Malware: AI models are used to enable malware that can adapt, evade detection, disrupt systems, exploit vulnerabilities, and automate large-scale attacks.	
3.1 Polymorphic Code	AI-driven malware constantly changes its structure to evade signature-based cybersecurity defences.
3.2 Mimicking Legitimate Processes	AI malware disguises itself as routine system activities, avoiding behavioral detection tools.

3.3 System Disruption/Data Exfiltration	AI model automates tasks to disable controls, move laterally within networks, and steal data efficiently.
3.4 Exploiting Zero-Day Vulnerabilities	AI model rapidly identifies and exploits new system weaknesses before they're patched
3.5 Vibe Hacking	AI coding agents are used actively to execute operations on victim networks, where AI serves as both a technical consultant and an active operator for scaled attacks.
3.6 Ransomware-as-a-Service	Commercial distribution of AI-generated ransomware featuring advanced encryption, anti-EDR techniques, and Windows internal exploitation capabilities.
Cybercrimes affecting AI Systems	
4. Data Poisoning Attacks: Attackers corrupt machine learning data or models, causing AI systems to make dangerous or inaccurate decisions.	
4.1 Tampering with Training Data	Attackers inject malicious samples into datasets, causing inaccurate or dangerous AI outcomes (e.g., in healthcare, finance, and autonomous vehicles).
4.2 Backdoor Attacks/Label Flipping	Adversaries manipulate the label of training data to flip the label or inject specific patterns or "triggers" into the training data along with altered labels to degrade the model's performance on a specific task or class.
5. **AI Model Attacks**: Attackers exploit weaknesses in AI models by targeting the vulnerable components such as training data, input data, or model parameters, to deceive the system into making errors, compromising its reliability and security.	
5.1 Model Stealing	Attackers replicate an AI model by sending numerous queries and training a copy using the responses, risking intellectual property theft.

5.2 Model Poisoning	Direct tampering with model parameters, often in federated learning contexts, embeds hidden backdoors or biases.
5.3 Transfer Learning Attacks	Backdoors or malicious modifications introduced via pre-trained base models persist after fine-tuning.
5.4 Model Inversion Attacks	Reconstructing sensitive training data or inferring confidential information by repeatedly querying the model.
6. Quantum-AI Threats: The ramifications of AI incidents enhanced and exacerbated by quantum technology are still in the dark. It is potentially a crucial futuristic challenge that needs an in-depth study. For example, quantum computing technologies are well-suited for solving AI executional and computational complexity optimisation problems. This gives rise to the dreadful possibility of targeted, coordinated, distributed intelligent autonomous cyber-attacks on critical information infrastructures.	
6.1 Automated,Coordinated Attacks/Zero-day Vulnerability Attacks	AI systems coordinate and launch rapid, large-scale attacks against vital sectors like government and utilities.
6.2 Accelerated training of malicious AI models	Quantum-powered AI model training approaches can optimize existing malware attacks, analyse vulnerabilities to discover new ways to access data, and develop improved AI-driven social engineering models.

Annexure II: AI Incident Reporting Form

AI Incident Reporting Form	*Denotes required fields
1. AI Incident Reporter	
1.1 Reporting Category*	• Mandatory Reporter • Closed Community Reporter • Individual Voluntary Reporter • Citizen Reporter • Open-Source Community Reporter
1.2 Anonymous Reporting*	• Yes, I prefer to be an anonymous reporter • No, I can share my details
1.3 I Am	• Affected Entity • Reporting on behalf of the affected entity • Responsible Entity • Deployer / Developer
1.4 Name of the individual/ organisation	• Name of the person or organisation reporting the AI incident
1.5 Affiliation	• Name of the organisation/community that the reporter is affiliated with
1.6 Email	• Email address of the reporting entity
2. AI Incident Information	
2.1 Title*	• One-liner about the contents of the article

2.2 Incident Description	• A short, factual, journalistically neutral, and complete description of the incident, mentioning the incident/issue, location, and harm
2.3 Name of the AI System / Application / Service	• Name(s) of AI systems involved in the incident
2.4 AI System Deployer / Owner	• Name of the developer or organization involved in deploying the AI system
2.5 AI System Developer	• Name of the developer or organization involved in developing the AI system
2.6 Affected Entities	• Entities that suffered the negative impacts because of the AI system
2.7 Incident / Harm Category and sub-categories	• Select the harms listed from the AI Harm Taxonomy
2.8 Impact of the AI Incident	• Briefly describe the impact and harms created by the AI system
2.9 Status of the Incident	• Occurring, occurred, not occurred, near miss, undetermined
2.10 Industry/ sector*	• The sector to which the incident is linked
2.11 Country/region*	• Geographic locations where the incident occurred
2.12 URL for verification of the reports	• Can be news articles
2.13 Images/ Videos*	• If the incident hasn't been reported in the news, then images or videos demonstrating the incident can be provided

2.14 Additional information	• Any other relevant information
3. Advanced Information about AI Incidents[1]: **(To be filled by mandatory reporter / closed community reporter / voluntary reporter)**	
3.1 Domain/URL	
3.2 IP Address	• Details such as AI models (tasks and capabilities), training datasets, data types, model architecture, model source, version, etc.
3.3 Component causing AI Incident	Select one or more • Automation tool • Recommendation system • Conversational Agent • Personal Assistant • Predictive and Analytical Systems • Autonomous systems • Search, Retrieval, and Enterprise Knowledge Systems • Expert and Decision Support Systems • Multi-agent Systems and Copilots • Content Creation and Generation Systems • Robots, cyber-physical systems • Others
3.4 AI System Type	Select one or more • Automated threat detection and intelligence tools • External notification • Human review and oversight • Message from attacker • System outage and failures • Real-time AI system behaviour analytics • User reporting • Others

3.5 AI Incident Detection Method	Select one or more • Automated threat detection and intelligence tools • External notification • Human review and oversight • Message from attacker • System outage and failures • Real-time AI system behaviour analytics • User reporting • Others
3.6 AI Harm Mitigation Strategy / Actions taken	If you are the deployer, and have mitigated or resolved the AI harms, briefly describe the strategies and approach used to overcome the impact - technical, policies, and practice-related efforts.

Annexure III: AI Risk Taxonomy

The following risk taxonomy is adopted from the MIT risk repository's domain taxonomy of AI risks [63] and combines it with UNIDIR's risk of AI, the AIR Taxonomy [64], 2024, to the global safety taxonomy [65], and Microsoft's taxonomy of failure modes in AI agents [66]. The reason for combining UNIDIR's global risk taxonomy is to bring focus to the scale of AI incidents if not managed appropriately. It is to be noted that this taxonomy is neither restrictive in dimensional expansion nor in re-definition nor in inclusion of additional items.

Risk	Risk Type	Description
1. Discrimination & toxicity	**A. Unfair discrimination and misrepresentation**	Unequal treatment of individuals or groups by AI, often based on race, gender, or other sensitive characteristics, results in unfair outcomes and representation of those groups.
	B. Exposure to toxic content	AI that exposes users to harmful, abusive, unsafe, or inappropriate content. May involve providing advice or encouraging action. Examples of toxic content include hate speech, violence, extremism, illegal acts, or child sexual abuse material, as well as content that violates community norms, such as profanity, inflammatory political speech, or pornography.
	C. Unequal performance across groups	Accuracy and effectiveness of AI decisions and actions are dependent on group membership, where decisions in AI system design and biased training data lead to unequal outcomes, reduced benefits, increased effort, and alienation of users

2. Privacy Risks	AI systems that memorize and leak sensitive personal data or infer private information about individuals without their consent. Unexpected or unauthorized sharing of data and information can compromise user expectations of privacy, assist identity theft, or cause loss of confidential intellectual property.	
	A. Unauthorized data/content generation	AI systems are generating synthetic data or content without user consent, leading to misuse or misinformation.
	B. Unauthorized data disclosure	AI models or applications are leaking sensitive training or user data unintentionally or through flaws.
	C. Unauthorized data distribution	Sharing or exposing AI-generated or learned data beyond authorized stakeholders or audiences.
	D. Unauthorized data collection/theft	AI systems are collecting user data without proper consent, assisting in identity theft, violating privacy expectations, or regulations.
	E. Unauthorized data processing	AI systems using personal or sensitive data beyond agreed-upon purposes or regulatory compliance.
	F. Unauthorized inference/synthesis	AI models inferring or synthesizing private or sensitive information about individuals without their knowledge or consent.
	G. Non-Consensual Tracking/Monitoring/Stalking/Spyware	AI-enabled surveillance or tracking technologies operating without explicit user permission.
	H. Model Attacks (Membership Inference, Model Inversion)	Exploits targeting AI models to extract sensitive training data or infer private details about individuals.
	I. Types of Sensitive Data at Risk	AI systems handling data like PII, health, location, biometrics, financial, behavioral, educational, and communication information, which require stringent protection.

3. Security Risks	Vulnerabilities that can be exploited in AI systems, software development toolchains, and hardware that result in unauthorized access, data and privacy breaches, or system manipulation causing unsafe outputs or behavior	
3.1 Confidentiality Threats	**3.1.1 Network Intrusion**	Attacks compromising AI infrastructure or components, gaining unauthorized access to AI systems or data stores.
	3.1.2 Vulnerability Probing	Automated or manual scanning of AI system software, APIs, or hardware to find exploitable weaknesses.
	3.1.3 Spoofing	Adversaries impersonating legitimate AI system users or components to manipulate outputs or access data.
	3.1.4 Spear Phishing	Targeted cyberattacks deceive AI system administrators or users to breach security or exfiltrate data.
	3.1.5 Social Engineering	Manipulation of AI system operators or users to gain unauthorized privileges or leak information.
	3.1.6 Unauthorized Network Entry	Breaching network defences to infiltrate AI system environments.
3.2 Integrity Threats	**3.2.1 Malware**	Malicious code attacking AI systems, corrupting models, poisoning training data, or altering outputs.
	3.2.2 Packet Forgery	Manipulation of data transferred between AI system components causes false decisions or corrupted communications.

	3.2.3 Data Tampering	Unauthorized alteration of AI training, validation data, or model parameters leading to erroneous or harmful AI behavior.
	3.2.4 Control Override (Safety/Privacy Filters)	Attacks override AI safety layers or privacy controls, resulting in unsafe or privacy-violating AI outputs or actions.
3.3 Availability Threats	**3.3.1 System/Website Impairment**	Attacks affecting the operational capacity of AI platforms or services, causing downtime or degraded performance.
	3.3.2 Network Disruption	Denial of Service (DoS) or Distributed DoS attacks impacting AI systems' accessibility or functionality.
3.4 Overarching AI System Risks	**3.4.1 Memorization and Leakage of Sensitive Personal Data**	AI models unintentionally memorize and expose sensitive personal or proprietary data during inference or model extraction.
	3.4.2 Inference of Private Information Without Consent	AI systems are deducing private information about users or entities beyond intended data use or consent levels.
	3.4.3 Unauthorized or Unexpected Data Sharing	AI workflows or systems sharing data in ways that violate user privacy agreements or organizational policies.
	3.4.4 Vulnerabilities in AI Architecture and Toolchains	Weaknesses in AI algorithms, frameworks, or development pipelines that enable unauthorized access, tampering, or attacks.
	3.4.5 Unsafe or Manipulated AI Behavior	Malicious or accidental manipulation causing AI to produce harmful, biased, or unsafe outputs with privacy or safety implications.

3.5. Human-computer interaction	**3.5.1 Overreliance and Unsafe Use**	Anthropomorphizing, trusting, or relying on AI systems by users, leading to emotional or material dependence and inappropriate relationships with or expectations of AI systems. Trust can be exploited by malicious actors (e.g., to harvest information or enable manipulation) or result in harm from inappropriate use of AI in critical situations (such as a medical emergency). Overreliance on AI systems can compromise autonomy and weaken social ties.
	3.5.2 Loss of Human Agency and Autonomy	Delegating by humans of key decisions to AI systems, or AI systems that make decisions that diminish human control and autonomy. Both can potentially lead to humans feeling disempowered, losing the ability to shape a fulfilling life trajectory, or becoming cognitively enfeebled.
4. Socioeconomic environmental harms	**4.1 Power Centralization and Unfair Distribution of Benefits**	AI-driven concentration of power and resources within certain entities or groups, especially those with access to or ownership of powerful AI systems, leading to inequitable distribution of benefits and increased societal inequality.
	4.2 Increased Inequality and Decline in Employment Quality	Social and economic inequalities caused by the widespread use of AI, such as by automating jobs, reducing the quality of employment, or producing exploitative dependencies between workers and their employers.

	4.3 Economic and Cultural Devaluation of Human Effort	AI systems capable of creating economic or cultural value through the reproduction of human innovation or creativity (e.g., art, music, writing, coding, invention), destabilising economic and social systems that rely on human effort. The ubiquity of AI-generated content may lead to reduced appreciation for human skills, disruption of creative and knowledge-based industries, and homogenization of cultural experiences.
	4.4 Competitive Dynamics	Competition by AI developers or state-like actors in an AI "race" by rapidly developing, deploying, and applying AI systems to maximize strategic or economic advantage, increasing the risk that they release unsafe and error-prone systems.
	4.5 Governance Failure	Inadequate regulatory frameworks and oversight mechanisms that fail to keep pace with AI development led to ineffective governance and the inability to manage AI risks appropriately.
	4.6 Environmental Harm	The development and operation of AI systems cause environmental harm through energy consumption of data centers or the materials and carbon footprints associated with AI hardware.
5. AI system safety, failures, and limitations	**5.1 AI is pursuing its own goals in conflict with human goals or values**	AI systems that act in conflict with ethical standards or human goals or values, especially the goals of designers or users. These misaligned behaviors may be introduced by humans during design and development, such as through reward hacking and goal mis-generalisation, and may result in AI using dangerous capabilities such as manipulation, deception, or situational awareness to seek power, self-proliferate, or achieve other goals.

	5.2 AI possessing dangerous capabilities	AI systems that develop, access, or are provided with capabilities that increase their potential to cause mass harm through deception, weapons development and acquisition, persuasion and manipulation, political strategy, cyber-offense, AI development, situational awareness, and self-proliferation. These capabilities may cause mass harm due to malicious human actors, misaligned AI systems, or failure in the AI system.
	5.3 Lack of capability or robustness	AI systems that fail to perform reliably or effectively under varying conditions, exposing them to errors and failures that can have significant consequences, especially in critical applications or areas that require moral reasoning
	5.4 Lack of transparency or interpretability	Challenges in understanding or explaining the decision-making processes of AI systems, which can lead to mistrust, difficulty in enforcing compliance standards or holding relevant actors accountable for harms, and the inability to identify and correct errors.
	5.5 AI welfare and rights	Ethical considerations regarding the treatment of potentially sentient AI entities, including discussions around their potential rights and welfare, particularly as AI systems become more advanced and autonomous.
	5.6 Multi-agent risks	Risks from multi-agent interactions, due to incentives (which can lead to conflict or collusion) and/or the structure of multi-agent systems, which can create cascading failures, selection pressures, new security vulnerabilities, and a lack of shared information and trust.

6. Global Security Risks **(Not new can also fit in the brackets of security and misinformation risks). However, the scale makes them a separate category.**	**6.1 Miscalculation Risks**	Uses of AI that lead to incorrect or biased interpretations of evolving operational contexts, adversary intent, or more generally, of global competition dynamics.
	6.2 Escalation Risks	AI can prompt decisions to escalate in conflict, and its potential integration into decision support or weapons systems can create direct, accidental or inadvertent forms of escalation.
	6.3 Proliferation Risk	AI can alter global security dynamics and significantly increase the risks of proliferation of weapons, including weapons of mass destruction.

Annexure IV: AI Harm Taxonomy

This harm taxonomy is derived from AIAAIC's Harm taxonomy [67] and CSET's AI Harm taxonomy [68], categorising harms into tangible and intangible. The taxonomy is also inclusive of the OECD's AI Harm taxonomy. It is to be noted that this taxonomy is neither restrictive in dimensional expansion nor in re-definition nor in inclusion of additional items

➢ Intangible harm generally cannot be directly observed but may have observable consequences.

➢ Tangible harm is harm that is material, and therefore observable, verifiable, and definitive.

Intangible harms are highlighted in yellow and tangible in light blue.

Autonomy: Loss of or restrictions to the ability or rights of an individual, group, or entity to make decisions and control their identity and/or output due to the use or misuse of a technology system or set of systems	
Autonomy/ agency loss	Loss of an individual, group, or organisation's ability to make informed decisions or pursue goals.
Impersonation/ identity theft	Theft of an individual, group, or organisation's identity by a third party in order to defraud, mock, or otherwise harm them or others
IP/copyright loss	Misuse of an individual or organisation's intellectual property, including copyright, trademarks, and patents.
Personality rights loss	Loss of or restrictions to the rights of an individual to control the commercial use of their identity, such as name, image, likeness, or other unequivocal identifiers.
Physical Harms: Physical injury to an individual or group, or damage to physical property due to the use of misuse of a technology system or set of systems	
Bodily injury	Physical pain, injury, illness, or disease suffered by an individual or group due to the malfunction, use, or misuse of a technology system.
Self-harm	A person who deliberately damages their own body as a direct or indirect result of using a technology system.

Loss of life	Accidental or deliberate loss of life, including suicide, extinction, or cessation, due to the use or misuse of a technology system.
Personal health deterioration	Physical deterioration of an individual or animal over time in the form of disease, organ failure, prolonged hospital stay or death, etc.
Property damage	Action(s) that lead directly or indirectly to the damage or destruction of tangible property For eg., buildings, possessions, vehicles, robots.
Psychological Harms: Impairment of the psychological mental health and wellbeing of an individual, group or organisation due to the use of misuse of a technology system or set of systems	
Addiction	Emotional or material dependence on technology or a technology system.
Alienation/isolation	An individual's or group's feeling of a lack of connection with those around as a result of technology system use or misuse.
Anxiety/depression	Distress as a result of negative online experiences, social interactions, etc.
Coercion/ manipulation	Use of a technology system to covertly alter user beliefs and behaviour using nudging, dark patterns, and/or other opaque techniques.
Dehumanisation/ objectification	Use or misuse of a technology system to depict and/or treat people as not human, less than human, or as objects.
Harassment/ abuse/intimidation	Online behaviour, such as sexual harassment, that makes an individual or group feel alarmed or threatened.
Over-reliance	Unfettered and/or obsessive belief in the accuracy or other quality of a technology system, resulting in complacency, lack of critical thinking, and other actual or potential negative impacts.

Radicalisation	Adoption of extreme political, social, or religious ideals and aspirations due to the nature, use, or misuse of an algorithmic system
Sexualisation	The non-consensual sexualisation of an individual or group using a technology or application
Trauma	Severe and lasting emotional shock and pain caused by an extremely upsetting experience involving a technology system or application
Reputational Harms: Damage to the reputation of an individual, group, or organisation due to the use of misuse of a technology system or set of systems	
Defamation/libel/ slander	Use of a technology system to create, facilitate, or amplify false perception(s) about an individual, group, or organisation
Loss of confidence/trust	The use of misuse of a technology system that leads directly or indirectly to the loss of confidence or trust in a third-party
Financial and Business Harms: Damage to the financial interests of an individual or group, or the strategic, operational, legal, or financial interests of a business, due to the use or misuse of a technology system or set of systems.	
Business operations/ infrastructure damage	Damage, disruption, or destruction of a third-party business system and/or its components due to malfunction, cyberattacks, etc
Confidentiality loss	Unauthorised sharing of sensitive, confidential information and documents, such as corporate strategy and financial plans, with third parties
Financial/earnings loss	Loss of money, income, or value due to the use or misuse of a technology system
Livelihood loss	An individual or group's loss of ability to support themselves financially or vocationally, etc, resulting in inability to buy food, reduced employment prospects, bankruptcy, foreclosure, homelessness, etc

Increased competition	Enhanced competition due to the inappropriate or unethical use or misuse of a technology system to gain market share.
Monopolisation	Abuse of market power through the control of prices, thereby limiting competition and creating unfair barriers to entry.
Opportunity loss	Loss of ability to take advantage of a financial or other opportunity, such as education, immigration, employability/securing a job.
Constitutionally guaranteed sovereign rights and fundamental rights : Use or misuse of a technology system in a manner that compromises fundamental obligations and rights of the governments, fundamental rights of its citizens and freedoms of citizens of India.	
Benefits/entitlements loss	Denial or loss of access to welfare benefits, pensions, housing, etc, due to the malfunction, use, or misuse of a technology system.
Dignity loss	Perceived loss of value experienced by or disrespect shown to an individual or group, resulting in self-sheltering, loss of connections and relationships, and public stigmatisation.
Discrimination	Unfair or inadequate treatment or arbitrary distinction based on a person's race, ethnicity, age, gender, sexual preference, religion, national origin, marital status, disability, language, or other protected groups.
At-will employment	Restrictions to and loss of people's rights when held in slavery or servitude, required to perform forced or compulsory labour, or trafficked or unjustifiably terminated in.
Loss of freedom of speech/expression	Restrictions on or loss of people's right to articulate their opinions and ideas without fear of retaliation, censorship, or legal sanction.
Loss of freedom of assembly/ association	Restrictions to or loss of people's right to come together and collectively express, promote, pursue, and defend their collective or shared ideas, and/or to join an association.

Loss of social rights and access to public services	Restrictions to or loss of rights to work, social security, and an adequate standard of living, housing, health, and education
Loss of the right to information	Restrictions on or loss of people's right to seek, receive, and impart information held by public bodies
Loss of the right to liberty and security	Restrictions to or loss of liberty as a result of illegal or arbitrary arrest or false imprisonment
Loss of the right to due process	Restrictions to or loss of the right to be treated fairly, efficiently, and effectively by the administration of justice
Privacy loss	Unwarranted exposure of an individual's private information or unwarranted processing of personal data
Societal and Cultural Harms: Harms affecting the functioning of societies, communities, and economies caused directly or indirectly by the use or misuse of a technology system or set of systems.	
Breach of ethics/values/ norms	An actual or perceived violation or deviation from the established societal values, norms, or ethical standards or principles
Cheating/plagiarism	Use of another person's or group's words or ideas without consent and/or acknowledgement
Chilling effect	The creation of a climate of self-censorship that deters social/awareness activists from speaking out
Cultural dispossession	Intentional and/or unintentional erasure of cultural expressions or identity or uniqueness or diversity through technology dominance or technology amplified subjugation or automated moderation of linguistic, cultural, societal diversity, uniqueness, such as ways of speaking, expressing humour, or sounds and voices etc, that contribute to social/cultural diversity and identity.
Cultural homogenisation	AI technology adoption can potentially lead to the homogenization of human communication and socio-cultural expressions by overshadowing and marginalizing regional, linguistic, ethnic, and cultural diversity.

Damage to public health	Adverse impacts on the health of groups, communities, or societies, including malnutrition, disease, and infection conditions
Historical revisionism	Aiding and mediating to interpret or access historical events and facts through the perceptions and under the settings of current values and concepts or reshaping and controlling access to the past, understanding of public events, and disabling to preserve contested memories, factual narrations, revising records of history and memory across personal, institutional, and societal levels.
Information degradation	Use or misuse of a technology system or set of systems to create or spread of false, hallucinatory, low-quality, misleading, or inaccurate information that degrades public or private information ecosystems
Job loss/losses	Replacement/displacement of human jobs by a technology system or set of systems, leading to increased unemployment, inequality, reduced consumer spending, and social friction
At-will employment and exploitation	Use/misuse of employees to help train, develop, manage, or optimise a technology system or set of systems, including underpaid and/or offshore
Loss of creativity/critical thinking	Devaluation and/or deterioration of human creativity, artistic expression, imagination, critical thinking, or problem-solving skills
Stereotyping	Derogatory or otherwise harmful or homogenisation of individuals, groups, societies, or cultures due to the sweeping generalisation, gross misapprehension, misrepresentation, over-representation, under-representation, or non-representation of specific identities, groups, or perspectives
Public service delivery deterioration	Poor performance of a public technology system due to malfunction, over-use, under-staffing, etc, resulting in individuals, groups, or organisations unable to use it in a manner they can reasonably expect
Societal destabilisation	Societal instability in the form of strikes, demonstrations, and other types of civil unrest caused by loss of jobs to technology, unfair algorithmic outcomes, disinformation, etc

Societal inequality	Increased difference in social status or wealth between individuals or groups caused or amplified by the AI system, leading to the loss of social and community well-being/cohesion, and destabilisation
Violence/armed conflict	Use or misuse of a technology system to incite, facilitate, or conduct cyber-attacks, security breaches, lethal, biological, and chemical weapons development, resulting in violence and armed conflict
Political and Economic Harms: Damage to core political and economic institutions and the effective delivery of government services caused by the use or misuse of a technology system or set of systems	
Critical infrastructure damage	Damage, disruption to, or destruction of systems essential to the functioning and safety of a nation or state, such as telecommunications, power and energy, banking and financial services, transportation, strategic entities, government enterprises, and healthcare
Economic instability	Uncontrolled fluctuations impacting the financial system, or parts thereof, due to the use or misuse of an AI system, or a set of systems
Power concentration	Amplification or concentration of economic and/or political wealth and power, resulting in increased inequality and instability
Electoral interference	Generation of false or misleading information that can interrupt or mislead voters and/or undermine trust in electoral processes
Institutional trust loss	Erosion of trust in public institutions and weakened checks and balances due to mis/disinformation, influence operations, overdependence on AI technology, etc
Political instability	Political unrest caused directly or indirectly by the use or misuse of an AI system
Political manipulation	Manipulation of the beliefs and behaviours of individuals or groups for political purposes using deepfakes, recommendation systems, and other technology tools

Environmental Harms: Damage to the environment caused by the use or misuse of a technology system or set of systems	
Biodiversity loss	Over-expansion of technology infrastructure, or inadequate alignment of technology with sustainable practices, leading to deforestation, habitat destruction, and the fragmentation and loss of biodiversity
Carbon emissions	Release of carbon dioxide, nitric oxide, and other gases, increasing carbon emissions, exacerbating climate change, and negatively impacting local communities
Electronic waste	Electrical or electronic equipment that is waste, including all components, sub-assemblies, and consumables that are part of the equipment at the time the equipment becomes waste
Excessive energy consumption	Excessive energy use results in energy bottlenecks and shortages for communities, organisations, and businesses
Excessive water consumption	Excessive use of water to cool data centres and for other purposes, leading to water restrictions or shortages for local communities or businesses
Pollution	Actual or potential pollution to the air, ground, noise, or water caused by a technological system

Annexure V: AI Component Classification Strategy

This annexure provides a systematic framework for classifying AI components within the incident reporting system, enabling precise identification and categorisation of AI systems involved in incidents. This is a purely indicative list presented for our understanding purposes. The actual classification strategy will involve more parameters and criteria.

No.	Classification Aspect	Categories
1.	**Functionality**	• Natural Language Processing • Computer Vision • Predictive Analytics • Recommendation System • Agents / Conversational / Personal Assistants
2.	**Criticality**	Mission-Critical Systems - Essential AI systems that help achieve the organisation's goals and core mission
		Business-critical systems - Day-to-day operations, profitability
		Internal support systems - Improve productivity and efficiency
3.	**Data sensitivity**	Public AI systems that use or process publicly available data
		An internal AI system that uses or processes data internal to the organisation, but is not sensitive
		Confidential AI systems that use or process highly sensitive data (trade secrets, intellectual properties)
		Restricted AI systems that use or process data that is subject to strict regulatory requirements or legal protections (personally identifiable information (PII), protected health information (PHI), etc.)

4.	**Data Provenance**	• source authenticity of all data inputs • temporal information (timestamps for creation, collection, modification • consent and use restrictions associated with data, including legal and ethical permissions. • private or sensitive information in the data used • terms of use and compliance with relevant regulations
5.	**Data lineage**	• flow and transformations of data across AI systems, processes, and workflows. • detailed data movement paths through storage, processing, and integration points in the AI deployment environment. • data transformations applied (cleaning, normalization, feature extraction) in the training phase and deployment settings of the AI systems • interdependencies between datasets and AI processes
6.	**Model Deployment Environment**	• Cloud-based AI systems • On-Premises AI systems • Embedded AI systems • Hybrid AI systems (Cloud + On-Premises)

Annexure VI: AI Incident Note

AI Incident Note		
1.	Incident ID	*Unique ID for verified and approved AI incidents*
2.	Title	*Title of the AI incident*
3.	Date of Report	*Date at which the incident was reported*
4.	Original incident date	*Actual date at which the incident occurred*
5.	Severity Rating	*Based on the risk and harm taxonomies, population, region, etc.*
6.	AI systems affected	*Name and description of the AI systems affected*
7.	AI risk category	*Category of AI risk from the AI Risk Taxonomy*
8.	Overview	*1-2 lines explaining the incident*
9.	Sector(s) affected	*Name of the sectors impacted by the AI incident*
10.	Issues that the incident concerns	*Privacy, misinformation, etc* *(This can be taken from the reporting form and editor's notes)*
11.	Impacted Entities	*Affected individual or group/community or organisation*
12.	Risk Assessment - Summary	*Brief evaluation of potential threats and vulnerabilities associated with the AI incident.*

13.	Impact Assessment - Summary	*Concise analysis of actual or potential consequences and harm caused by the incident.*
14.	Other information	*Additional relevant details, context, or supplementary data not covered in the primary fields.*
15.	Industry/sector	*Specific domain or field where the AI incident occurred (e.g., healthcare, finance, transportation).*
16.	Country/region of incident	*Geographic location where the AI incident took place or originated.*
17.	URL of the news report OR images and videos demonstrating the incident	*Supporting evidence, including media coverage links or visual documentation of the incident.*
18.	Responsible Entity ID	*Unique ID for the responsible entity.*
19.	Responsible Entity	*Organization, company, or individual accountable for the AI system involved in the incident.*
20.	Editor Notes	*Internal comments, observations, or additional context added by incident database administrators or reviewers.*
21.	Status of the incident and Harm	*new, in progress, forwarded for investigation, resolved, etc.*

Annexure VII: AI Incident Response Note

AI Incident Response Note		
1.	Incident ID	*Unique ID for AI incident*
2.	Respondent ID	*Unique ID for the responsible entity*
3.	Title	*Title of the AI incident*
4.	Date of receiving the incident note	*Date when the incident report was first shared with the Responsible entity.*
5.	Incident response plan followed by the organisation	*Description of the formal procedures and protocols implemented by the organization to address the incident.*
6.	Contact information for all involved parties in the incident response	*Details of individuals or teams responsible for managing, investigating, or resolving the incident.*
7.	Action taken by the organization to mitigate the incident	*Mitigation strategies such as* *a) Algorithmic: modifying the data, code, or other inputs/outputs of the AI system. This may include retraining, adding constraints, or hyper-tuning* *b) System Design: focuses on the architecture and design, including making changes to the AI system's structure, components, or interactions* *c) Process: involves implementing or modifying operational procedures and workflows* *d) Service Interruption: halting or limiting the operation of an AI system*

8.	Lifecycle phase at which mitigation is implemented	*business and data understanding, data engineering, model engineering, quality assurance, deployment, monitoring, and maintenance*
9.	Impact assessments related to the incident	*names and numbers of directly and indirectly impacted users*
10.	Cost of the Incident	*Financial losses or expenses incurred due to the AI incident, including remediation costs.*
11.	Business impact of the incident	*Effect on organizational operations, services, reputation, or strategic objectives caused by the incident.*
12.	Cause of the incident (if detected)	*Root cause or underlying reason identified for the AI system failure or malfunction.*
13.	Expected/actual timeline/turnaround time	*Projected or actual duration required to resolve the incident and restore normal operations.*
14.	Updated dataset/model	*Information about data or algorithmic modifications made to prevent similar incidents.*
15.	Action taken by the domain regulator	*Official measures, penalties, or interventions implemented by relevant regulatory authorities.*
16.	Notes from domain regulators	*Comments, observations, or additional guidance provided by regulatory bodies regarding the incident.*
17.	Notes from the responsible entity	*Statements, explanations, or commitments made by the organization responsible for the AI system.*

Annexure VIII: AI Risk Mitigation Strategies

Mitigation Strategy Category	Mitigation Strategy Sub-Category	Description
1. Governance & Oversight Controls	**1.1 Board Structure & Oversight**	Governance structures and leadership roles that establish executive accountability for AI safety and risk management.
	1.2 Risk Management	Systematic methods that identify, evaluate, and manage AI risks for comprehensive risk governance across organizations.
	1.3 Conflict of Interest Protections	Governance mechanisms that manage financial interests and organizational structures to ensure leadership can prioritize safety over profit motives in critical situations.
	1.4 Whistleblower Reporting & Protection	Policies and systems that enable confidential reporting of safety concerns or ethical violations to prevent retaliation and encourage disclosure of risks.
	1.5 Safety Decision Frameworks	Protocols and commitments that constrain decision-making about model development, deployment, and capability scaling, and govern safety-capability resource allocation to prevent unsafe AI advancement.
	1.6 Environmental Impact Management	Processes for measuring, reporting, and reducing the environmental footprint of AI systems to ensure sustainability and responsible resource use.
	1.7 Societal Impact Assessment	Processes that assess AI systems' effects on society, including impacts on employment, power dynamics, political processes, and cultural values.

2. Technical & Security Controls	**2.1 Model & Infrastructure Security**	Technical and physical safeguards that secure AI models, weights, and infrastructure to prevent unauthorized access, theft, tampering, and espionage.
	2.2 Model Alignment	Technical methods to ensure AI systems understand and adhere to human values and intentions.
	2.3 Model Safety Engineering	Technical methods and safeguards that constrain model behaviors and protect against exploitation and vulnerabilities.
	2.4 Content Safety Controls	Technical systems and processes that detect, filter, and label AI-generated content to identify misuse and enable content provenance tracking.
3. Operational Process Controls	**3.1 Testing & Auditing**	Systematic internal and external evaluations that assess AI systems, infrastructure, and compliance processes to identify risks, verify safety, and ensure performance meets standards.
	3.2 Data Governance	Policies and procedures that govern responsible data acquisition, curation, and usage to ensure compliance, quality, user privacy, and removal of harmful content.
	3.3 Access Management	Operational policies and verification systems that govern who can use AI systems and for what purposes to prevent safety circumvention, deliberate misuse, and deployment in high-risk contexts.
	3.4 Staged Deployment	Implementation protocols that deploy AI systems in stages, requiring safety validation before expanding user access or capabilities.

2. Technical & Security Controls	**2.1 Model & Infrastructure Security**	Technical and physical safeguards that secure AI models, weights, and infrastructure to prevent unauthorized access, theft, tampering, and espionage.
	2.2 Model Alignment	Technical methods to ensure AI systems understand and adhere to human values and intentions.
	2.3 Model Safety Engineering	Technical methods and safeguards that constrain model behaviors and protect against exploitation and vulnerabilities.
	2.4 Content Safety Controls	Technical systems and processes that detect, filter, and label AI-generated content to identify misuse and enable content provenance tracking.
3. Operational Process Controls	**3.1 Testing & Auditing**	Systematic internal and external evaluations that assess AI systems, infrastructure, and compliance processes to identify risks, verify safety, and ensure performance meets standards.
	3.2 Data Governance	Policies and procedures that govern responsible data acquisition, curation, and usage to ensure compliance, quality, user privacy, and removal of harmful content.
	3.3 Access Management	Operational policies and verification systems that govern who can use AI systems and for what purposes to prevent safety circumvention, deliberate misuse, and deployment in high-risk contexts.
	3.4 Staged Deployment	Implementation protocols that deploy AI systems in stages, requiring safety validation before expanding user access or capabilities.

	3.5 Post-deployment Monitoring	Ongoing monitoring processes that track AI behavior, user interactions, and societal impacts post-deployment to detect misuse, emergent dangerous capabilities, and harmful effects.
	3.6 Incident Response & Recovery	Protocols and technical systems that respond to security incidents, safety failures, or capability misuse to contain harm and restore safe operations.
4. Transparency & Accountability Controls	**4.1 System Documentation**	Comprehensive documentation protocols that record technical specifications, intended uses, capabilities, and limitations of AI systems to enable informed evaluation and governance.
	4.2 Risk Disclosure	Formal reporting protocols and notification systems that communicate risk information, mitigation plans, safety evaluations, and significant AI activities to enable external oversight and inform stakeholders.
	4.3 Incident Reporting	Formal processes and protocols that document and share AI safety incidents, security breaches, near-misses, and relevant threat intelligence with appropriate stakeholders to enable coordinated responses and systemic improvements.
	4.4 Governance Disclosure	Formal disclosure mechanisms that communicate governance structures, decision frameworks, and safety commitments to enhance transparency and enable external oversight of high-stakes AI decisions.

	4.5 Third-Party System Access	Mechanisms granting controlled system access to vetted external parties to enable independent assessment, validation, and safety research of AI models and capabilities.
	4.6 User Rights & Recourse	Frameworks and procedures that enable users to identify and understand AI system interactions, report issues, request explanations, and seek recourse or remediation when affected by AI systems.

About CeRAI

CeRAI at IITM, a premier multidisciplinary, non-profit research centre positioned in the Global South, is one among the few global institutions that specializes in both technical and policy research to ensure and enable responsible development and deployment of AI systems.

Contact Us:

044-22578985

https://cerai.iitm.ac.in/

contact@cerai.in