15TH ITU ACADEMIC CONFERENCE

ITU KALEIDOSCOPE
NEW DELHI 2024

Innovation and digital transformation for a sustainable world

21-23 October 2024
New Delhi, India

Advancing Trustworthy AI for Sustainable Development: Recommendations for Standardising AI Incident Reporting

23 October 2024

Avinash Agarwal
Telecommunication Engineering Centre
Ministry of Communications
New Delhi, India

Manisha J. Nene
Defence Institute of Advanced Technology
Ministry of Defence
Pune, India

Session #S9.2

Agenda

- Introduction
- Original contributions and novelty
- Literature review
- Methodology
- Results
- Gap Analysis and Recommendations
- Conclusion and future directions
- References

Introduction

- Increasing use of AI technologies has led to rising AI incidents (e.g., biases, deepfakes, system failures), posing risks and harm to individuals, organizations, and society.
- AI can accelerate progress toward the UN Sustainable Development Goals (SDGs), but responsible and ethical deployment is critical to avoid negative impacts on goals like employment and equality.
- Responsible AI principles emphasize fairness, transparency, and accountability; standards and benchmarks are needed to ensure AI aligns with these principles.
- Learning from AI incidents can prevent repeats, similar to the aviation and cybersecurity industries with established reporting systems.
- Transparent AI incident reporting and systematic data analysis are crucial for developing mitigation strategies and ensuring trustworthy AI.
- This paper addresses the lack of standardized AI incident reporting protocols and explores the intersection of AI, sustainable digital transformation, and SDGs.

Original contributions and novelty

1. The study identifies nine gaps in existing AI incident reporting practices, offering insights into areas for improvement.

2. It proposes nine actionable recommendations to enhance standardization efforts in AI incident reporting, addressing the identified gaps.

3. It facilitates the development of strategies and mechanisms to prevent similar incidents from occurring in the future, thereby promoting trustworthy AI and aligning with the UN SDGs.

Our research promotes the development of standards to prevent future AI incidents and promote trustworthy AI, thus facilitating achieving the UN SDGs.

Literature review

AI incident definitions

- Reviewed multiple definitions of “AI incident"
  - OECD [15]
  - AI Incident Database (AIID) [16]
  - AI, Algorithmic, and Automation Incidents and Controversies (AIAAIC) [17]
- The review reveals the gap related to a lack of standard terms, definitions, and taxonomies.

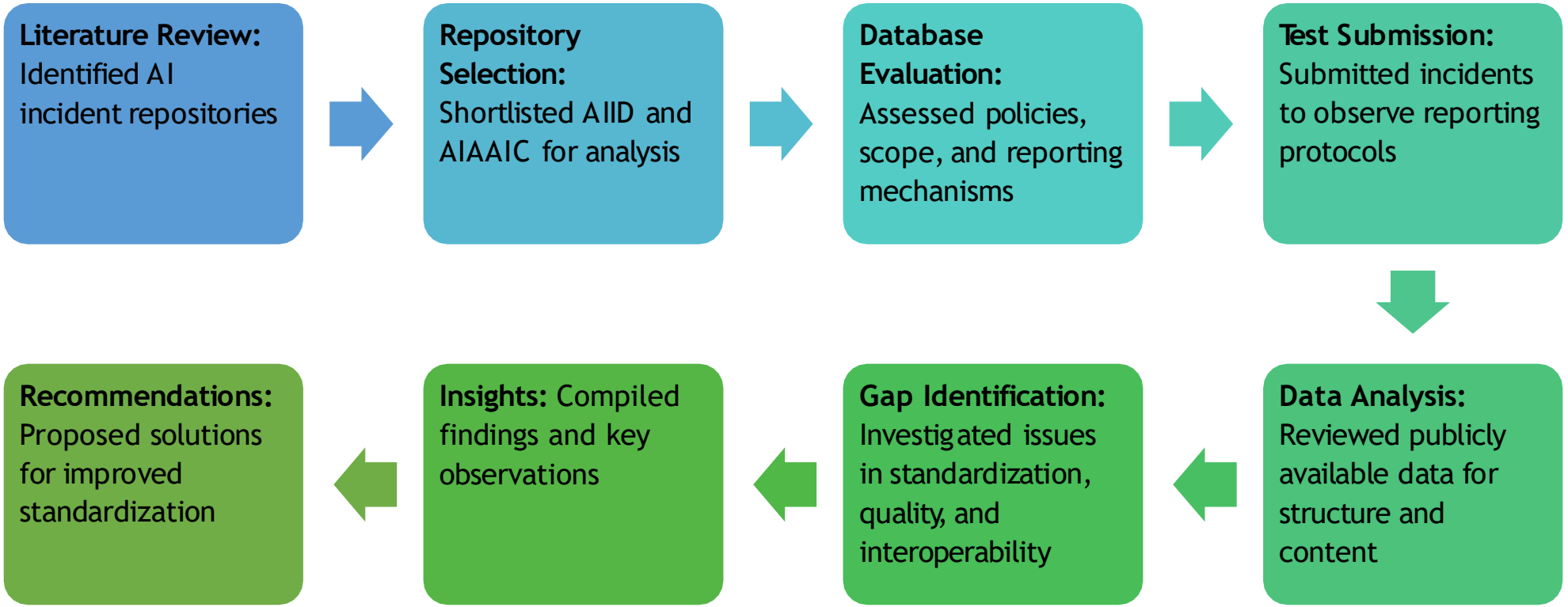
Need for AI incident reporting

- Published work recognizes that incident sharing is vital to prevent vulnerabilities, biases, and privacy concerns in AI systems, to ensure their trustworthiness and enhancing user experience.

AI incident repositories

- Reviewed the available AI incident repositories.
- Noted their shortcomings highlighted by researchers.

Methodology



Results (1/10)

This section presents the observations and results of the study

Incident reporting	AI-Incident snapshot	Interoperability and data sharing	Contributors to the Databases
Sources of the reports submitted to the databases	Sector Coverage	Geographical coverage	Data sharing

Results (2/10)

Table 1 – Incident reporting in AIAAIC and AIID

	AIAAIC	AIID
What can be reported	Incidents and controversies driven by and relating to AI.	Real-world harms or near harms caused by AI systems.
Incidents reported (as on 05-05-2024)	905	657
Who can report incidents	Anyone	Anyone
Submissions reviewed before publishing?	Yes	Yes
Nature of reporting	Voluntary	Voluntary
Incentive for reporting	None	None

Obs: Incident reporting in both databases is voluntary and lacks incentives.

Inf: Reporting relies on the submitter's discretion and motivation, potentially resulting in underreporting.

Results (3/10)

Table 2 – Snapshot of Incidents reported in AIAAIC

AIAAIC ID#	Headline	Ref.
AIAAIC1449	Adobe trained Firefly AI model on competitor images	[28]
AIAAIC1439	OpenAI scrapes YouTube to train GPT-4	[29]
AIAAIC1414	Leonardo AI generates celebrity non-consensual porn images	[30]
AIAAIC1395	Scientific journals publish papers with AI-generated introductions	[31]
AIAAIC1368	Microsoft Copilot generates fake Putin comments on Navalny death	[32]
AIAAIC1356	ChatGPT 'goes crazy', speaks gibberish	[33]

Obs: [28], [29], and [31] may not qualify as AI incidents depending on the definition considered.

Inf: Lack of consistency in qualifying the reported events as incidents. Biases of the submitters or the reviewers. Also, depending on their exposure, capabilities, and understanding

Results (4/10)

Table 3 – Comparison of data fields available in AIID and AIAAIC

Fields available in both AIID and AIAAIC	Fields available only in AIID	Fields available only in AIAAIC
Incident ID; Title/ Headline; Description; Occurrence date; System deployer; System developer;	Alleged harmed or nearly or nearly harmed parties	Type; Released (year); Country(ies); Sector(s); System name(s); Technology(ies); Purpose(s); Media trigger(s); Issue(s); Transparency; External harms; Internal harms

Obs: 1. Only six fields are compatible between the two datasets. 2. Important fields missing in both.

Inf: 1. Aggregating data from multiple databases difficult. 2. Captured data insufficient for assessing the severity and proper categorization of the incidents.

Results (5/10)

Table 4 – Top seven submitters of the incidents in AIID

Submitters	Incidents	%age
Daniel Atherton	149	23%
Anonymous	96	15%
Khoa Lam	93	14%
Ingrid Dickinson CSET	49	7%
Roman Yampolskiy	29	4%
AIAAIC	25	4%
Kate Perkins	21	3%

Obs: Seven submitters reported more than 70% of all the incidents in AIID.

Inf: Process reforms required to widen the base of incident reporting.

Results (6/10)

Table 5 – Top seven source-domains of the reports in AIID

Source domain	Reports
theguardian.com	143
theverge.com	95
nytimes.com	94
washingtonpost.com	71
wired.com	69
vice.com	54
reuters.com	53
bbc.com	53

Obs: Most reports submitted to AIID are from American or European newspapers.

Inf: Technological interventions required for incident reporting

Results (7/10)

Table 6 – Top seven sectors of the incidents in AIAAIC

Sectors	Incidents	%age
Media/entertainment/sports/arts	193	21.3%
Automotive	86	9.5%
Politics	75	8.3%
Technology	60	6.6%
Education	58	6.4%
Banking/financial services	40	4.4%
Business/professional services	35	3.9%

Obs: Skewed represent-ations of application sectors.

Inf: They predominantly report consumer-oriented sectors; underrepresent critical infrastructure sectors.

Results (8/10)

Table 7 – Top seven deployers of the AI systems in AIID

Deployer of AI system	incidents	%age
tesla	39	6%
facebook	36	6%
google	28	4%
unknown	23	4%
amazon	21	3%
openai	20	3%
cruise	12	2%

Obs: Maximum incidents reported in AIID relate to self-driving cars and social media.

Inf: Critical infrastructure sectors (telecom, power, etc.) underrepresented.

Results (9/10)

Table 8 – Top seven countries of the incidents in AIAAIC

Countries	Incidents	%age
USA	424	46.9%
UK	59	6.5%
China	53	5.9%
USA; Global	26	2.9%
Global	21	2.3%
India	21	2.3%
Canada	18	2.0%

Obs: Just three countries account for 60% of the incidents reported in AIAAIC.

Inf: Lack representation from developing and underdeveloped countries.

Results (10/10)

Table 9 – Sharing of incident data by AIAAIC and AIID

Data sharing	AIAAIC	AIID
Format	Available as a Google Sheet.	Weekly snapshots of the database in JSON, MongoDB, and CSV format
Information not accessible	Contributor details are not public. Harm data is only accessible to premium members.	-
APIs	None	None

Obs: Different formats for downloading data; both do not provide APIs for accessing data.

Inf: Lack of standardized mechanisms for sharing incident data among stakeholders.

Gap Analysis and Recommendations (1/2)

S. No	Gaps observed	Recommendations
1	Lack of definitions and taxonomies	Standardise AI-incident and AI-harms taxonomies
2	Bias, inconsistencies, and misclassification	Define guidelines for AI-incident database quality audits
3	Insufficient and incompatible data fields	Standardise AI-incident database structures
4	Inadequate motive to report incidents	Develop regulatory and policy frameworks for AI-incident reporting
5	Narrow base of the incidents reported	Develop Standards for automated incident reporting
6	Inadequate data-sharing protocols	Standardise data sharing mechanisms
7	Sectoral underrepresentation	Sector-specific AI-incident databases
8	Demographic underrepresentation	ITU-led inclusive AI incident reporting
9	Lack of awareness	Awareness Programs

Gap Analysis and Recommendations (2/2)

It is further recommended to include incident reporting as an integral part of the AI lifecycle as illustrated in figure-1 so that it gets appropriate focus in the future.

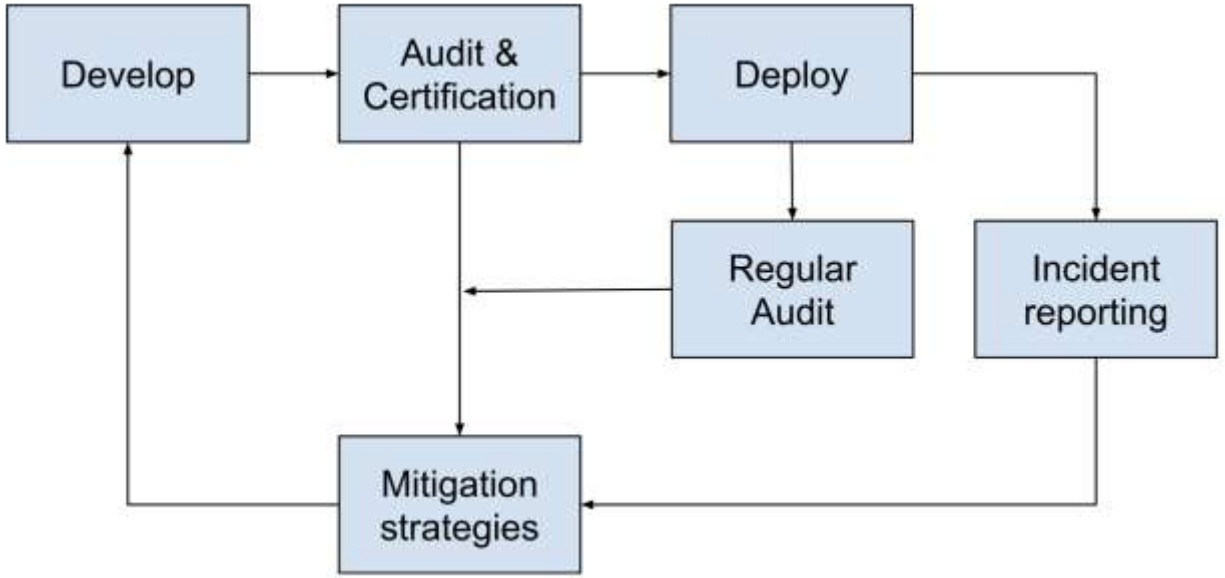


Figure-1: Conceptualised AI lifecycle stages

Conclusion and future directions

Critical Need	Standardized AI incident reporting for data-driven mitigation and research.
Key Findings	Identified 9 standardization gaps in AI incident reporting.
Recommendations	Proposed 9 solutions for improved reporting protocols.
Impact	Standardization enables data-driven strategies, product improvement, and trustworthy AI deployment.
Importance	Enhancing trust, transparency, and accountability in AI.
UN SDG Alignment	Contributes to sustainable development and responsible AI use.
Goal	Align AI frameworks with the UN SDGs for a more inclusive and sustainable future.
Future Work	Prioritize multi-stakeholder engagement and international cooperation in standardization endeavours.

References (1/3)

1. Nina Schwalbe and Brian Wahl. Artificial intelligence and the future of global health. The Lancet, 395(10236):1579-1586, 2020.
2. Tumaini Mwendile Kabudi. Artificial Intelligence for Quality Education: Successes and Challenges for AI in Meeting SDG4. In International Conference on Social Implications of Computers in Developing Countries, pages 347-362. Springer, 2022.
3. Margaret A Goralski and Tay Keong Tan. Artificial intelligence and sustainable development. The International Journal of Management Education, 18(1):100330, 2020.
4. Walter Leal Filho, Tony Wall, Serafino Afonso Rui Mucova, Gustavo J Nagy, Abdul-Lateef Balogun, Johannes M Luetz, Artie W Ng, Marina Kovaleva, Fardous Mohammad Safiul Azam, Fátima Alves, et al. Deploying artificial intelligence for climate change adaptation. Technological Forecasting and Social Change, 180:121662, 2022.
5. Ricardo Vinuesa, Hossein Azizpour, Iolanda Leite, Madeline Balaam, Virginia Dignum, Sami Domisch, Anna Felländer, Simone Daniela Langhans, Max Tegmark, and Francesco Fuso Nerini. The role of artificial intelligence in achieving the Sustainable Development Goals. Nature communications, 11(1):1-10, 2020.
6. Shivam Gupta, Simone D Langhans, Sami Domisch, Francesco Fuso-Nerini, Anna Felländer, Manuela Battaglini, Max Tegmark, and Ricardo Vinuesa. Assessing whether artificial intelligence is an enabler or an inhibitor of sustainability at indicator level. Transportation Engineering, 4:100064, 2021.
7. Estrella Gomez-Herrera and Sabine T Köszegi. A gender perspective on artificial intelligence and jobs: the vicious cycle of digital inequality. Technical report, Bruegel Working Paper, 2022.
8. OECD. OECD AI Principles overview. https://oecd.ai/en/ai-principles.
9. Jon Truby. Governing artificial intelligence to benefit the UN sustainable development goals. Sustainable Development, 28(4):946-959, 2020.
10. Avinash Agarwal and Harsh Agarwal. A seven-layer model with checklists for standardising fairness assessment throughout the AI lifecycle. AI and Ethics, pages 1-16, 2023
11. Yi Gao, Yang Hao, Sen Wang, and Hao Wu. The dynamics between voluntary safety reporting and commercial aviation accidents. Safety science,141:105351, 2021.

References (2/3)

12. Tianxi Dong, Qiwei Yang, Nima Ebadi, Xin RobertLuo, and Paul Rad. Identifying incident causal factors to improve aviation transportation safety: Proposing a deep learning approach. Journal of advanced transportation, 2021:1–15, 2021.
13. Sandra Schmitz-Berndt. Defining the reporting threshold for a cybersecurity incident under the NIS Directive and the NIS 2 Directive. Journal of Cybersecurity, 9(1):tyad009, 2023.
14. Shahar Avin, Haydn Belfield, Miles Brundage, Gretchen Krueger, Jasmine Wang, Adrian Weller, Markus Anderljung, Igor Krawczuk, David Krueger, Jonathan Lebensold, et al. Filling gaps in trustworthy development of AI. Science, 374(6573):1327–1329,2021.
15. OECD. Stocktaking for the development of an AI incident definition. (4), 2023.
16. AIID. AI Incident Database. Incidents(incidentdatabase.ai), 2024. Accessed: 16/5/2024.
17. AIAAIC. AIAAIC Repository. https://www.aiaaic.org/aiaaic-repository, 2024. Accessed:16/5/2024.
18. Giampiero Lupo. Risky artificial intelligence: The role of incidents in the path to AI regulation. Law, Technology and Humans, 5(1):133–152, 2023.
19. Miles Brundage, Shahar Avin, Jasmine Wang, Haydn Belfield, Gretchen Krueger, Gillian Hadfield, Heidy Khlaaf, Jingying Yang, Helen Toner, Ruth Fong, et al. Toward trustworthy AI development: mechanisms for supporting verifiable claims. arXiv preprintarXiv:2004.07213, 2020.
20. Bo Li, Peng Qi, Bo Liu, Shuai Di, Jingen Liu, JiquanPei, Jinfeng Yi, and Bowen Zhou. Trustworthy ai:From principles to practices. ACM Computing Surveys,55(9):1–46, 2023.
21. Violet Turri and Rachel Dzombak. Why We Needto Know More: Exploring the State of AI Incident Documentation Practices. In Proceedings of the 2023AAAI/ACM Conference on AI, Ethics, and Society, pages576–583, 2023.
22. Sean McGregor. Preventing repeated real world AI failures by cataloging incidents: The AI incident database. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, pages 15458–15463,2021.

References (3/3)

23. Syeda Faiza Nasim, Muhammad Rizwan Ali, and Umme Kulsoom. Artificial intelligence incidents & ethics a narrative review. International Journal of Technology, Innovation and Management (IJTIM),2(2):52–64, 2022.
24. AVID. AI Vulnerability Database. https://avidml.org/, 2024. Accessed: 16/5/2024.
25. AILD. AI Litigation Database. https://blogs.gwu.edu/law-eti/ai-litigation-database/, 2024.Accessed: 16/5/2024.
26. OECD. OECD AI Incidents Monitor. https://oecd.ai/en/incidents-methodology, 2024. Accessed:09/3/2024.
27. Avinash Agarwal and Manisha Nene. Addressing AI Risks in Critical Infrastructure: Formalising the AI Incident Reporting Process. In Proceedings of the 10thInternational Conference on Electronics, Computing and Communication Technologies, IEEE CONECCT, July 2024.
28. AIAAIC. AIAAIC incident id AIAAIC1449. https://www.aiaaic.org/aiaaic-repository/ai-algorithmic-and-automation-incidents/adobe-trained-firefly-ai-model-on-competitor-images, 2024. Accessed: 16/5/2024.
29. AIAAIC. AIAAIC incident id AIAAIC1439. https://www.aiaaic.org/aiaaic-repository/ai-algorithmic-and-automation-incidents/openai-scraped-youtube-to-train-gpt-4, 2024.Accessed: 16/5/2024.
30. AIAAIC. AIAAIC incident id AIAAIC1414. https://www.aiaaic.org/aiaaic-repository/ai-algorithmic-and-automation-incidents/leonardo-ai-generates-celebrity-non-consensual-porn-images, 2024. Accessed: 16/5/2024.
31. AIAAIC. AIAAIC incident id AIAAIC1395. https://www.aiaaic.org/aiaaic-repository/ai-algorithmic-and-automation-incidents/scientific-journals-publish-papers-with-ai-generated-introductions, 2024. Accessed: 16/5/2024.
32. AIAAIC. AIAAIC incident id AIAAIC1368. https://www.aiaaic.org/aiaaic-repository/ai-algorithmic-and-automation-incidents/microsoft-copilot-generates-fake-putin-comments-on-navalny-death, 2024. Accessed: 16/5/2024.
33. AIAAIC. AIAAIC incident id AIAAIC1356. https://www.aiaaic.org/aiaaic-repository/ai-algorithmic-and-automation-incidents/chatgpt-goes-crazy-speaks-gibberish, 2024. Accessed:16/5/2024.

Thank you!