ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

An Incident-Based Ethical Risk Screening Matrix for AI-Based Information Systems Using Public AI Incident Records

Romelyn J. Banaybanay[1], Reagan B. Ricafort[2]

[1]Initao College, Initao, Misamis Oriental, Philippines

[2]AMA University, Makati City, Philippines

DOI: https://doi.org/10.51583/IJLTEMAS.2026.150600080

Received: 21 June 2026; Accepted: 26 June 2026; Published: 07 July 2026

ABSTRACT

This study developed an incident-based ethical risk screening matrix for AI-based information systems using 100 publicly reported AI incident records from 2020 to 2026. It used descriptive content analysis and criterion-based purposive sampling. Each incident was coded by year, system domain, ethical risk type, affected stakeholder, ethical principle violated, severity level, and evidence strength. The study aimed to identify common ethical risk patterns and translate them into a practical screening tool for IT managers and organizations. Results showed that the most common ethical risks were misinformation or deception, followed by misuse or malicious use, safety failure, privacy violation, and accountability failure. Generative AI and chatbots had the highest number of incidents. Notable risk exposure was also found in law enforcement and surveillance, finance, government and public service, education, and social media platforms. The most affected stakeholder group was the general public. Most cases were assessed as high severity, while some were assessed as critical severity. Another supporting pattern was the prevalence of deepfakes and synthetic media, mainly linked to impersonation, fraud, misinformation, privacy harm, and reputational damage. The study concludes that AI-based information systems need ethical screening early in adoption, deployment, or expansion. The proposed matrix helps users identify warning signs, ask targeted screening questions, and apply management actions. It also gives organizations a structured way to connect documented AI failures to practical review steps. The matrix is not a substitute for a full technical, legal, or cybersecurity review. It uses documented AI failures as practical evidence to support early risk identification and responsible IT management.

Keywords: AI ethics, AI incidents, ethical risk, information systems, screening matrix

INTRODUCTION

Artificial intelligence is now part of many information systems used in education, health care, finance, employment, government service, transportation, social media, business, and public communication. These systems process data, create content, support decisions, classify users, recommend actions, and automate services. As use increases, ethical risks become easier to observe. These risks include misinformation, privacy violation, bias, discrimination, safety failure, weak transparency, weak accountability, misuse, and harm to vulnerable groups.

AI risk management has become a practical concern for organizations. The NIST AI Risk Management Framework identifies safety, security, resilience, accountability, transparency, explainability, privacy, and fairness as important characteristics of trustworthy AI [20]. The OECD also supports common reporting criteria for AI incidents [22]. These frameworks show that AI risk is technical, ethical, and organizational.

Public AI incident records show how ethical risks appear in actual use. The AI Incident Database records real-world harms and near-harms caused by deployed AI systems [1]. Prior studies use incident records to classify risks, identify repeated failures, and improve AI incident reporting [7, 9, 16, 23]. Other studies use AI incident cases to examine ethical issues, sector context, responsibility, and consequences [5, 11, 15, 30].

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

The AI ethics literature supports structured risk assessment. Researchers point to accountability, fairness, transparency, human oversight, and harm reduction as basic requirements for responsible AI use [10, 24, 25]. Studies on ethics-based auditing show the need to turn broad principles into tools used in actual organizational practice [18, 19].

Generative AI adds another layer of risk [3]. These systems create text, images, audio, video, and recommendations. They also raise concerns about hallucination, misuse, user reliance, downstream use, and synthetic media [4, 13, 29]. Deepfakes are included only as a supporting pattern because several encoded incidents involved synthetic media. Studies on deepfakes show risks involving deception, identity misuse, privacy harm, public trust, reputational damage, and disinformation [6, 12, 14, 17, 27, 28].

AI incident records and AI risk frameworks are available, but organizations still need simple tools that turn real failures into early screening questions. Many frameworks explain risk at a broad level. Incident databases give actual cases of harm. This study connects these two sources by using public AI incident records to develop an ethical risk screening matrix for AI-based information systems.

This study analyzes 100 publicly reported AI incident records from 2020 to 2026. It uses descriptive content analysis and criterion-based purposive sampling. Each incident is classified by system domain, ethical risk type, affected stakeholder, ethical principle violated, severity level, and evidence strength. The results are then used to develop the Incident-Based Ethical Risk Screening Matrix for AI-Based Information Systems.

The study is useful for IT managers, system developers, administrators, and organizations that plan to adopt or use AI-based information systems. The matrix guides them in identifying warning signs, asking ethical screening questions, and planning preventive actions before AI systems create harm.

Objectives of the study

The general objective of this study is to develop an Incident-Based Ethical Risk Screening Matrix for AI-Based Information Systems using publicly reported AI incident records.

Specifically, this study aims to:

1. Identify the common ethical risk types found in publicly reported AI incident records.

2. Classify AI incidents by system domain, affected stakeholder, ethical principle violated, and severity level.

3. Analyze patterns among ethical risk types, system domains, affected stakeholders, and severity levels.

4. Develop an ethical risk screening matrix based on the classified AI incident records.

5. Recommend how IT managers and organizations should use the matrix for early ethical risk assessment.

Research questions

This study seeks to answer the following questions:

1. What ethical risk types appear in publicly reported AI incident records?

2. What AI-based information system domains are most commonly involved in reported AI incidents?

3. Who are the affected stakeholders in the reported AI incidents?

4. How severe are the ethical risks found in the incident records?

5. What ethical risk screening matrix can be developed from the classified AI incident records?

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

Scope and limitations

This study is limited to 100 publicly reported AI incident records from 2020 to 2026. It uses public records as documentary data. It does not include system development, technical testing of AI models, surveys, or interviews.

The study focuses on ethical risk classification and screening matrix development. It does not represent all AI incidents worldwide. Because the sample was selected through criterion-based purposive sampling, the findings are intended for descriptive analysis and practical risk screening.

The study depends on the completeness and accuracy of available public reports. Some records contain limited technical detail. For this reason, the coding and severity ratings are based only on information found in the public records.

METHODOLOGY

Research design

This study used descriptive content analysis. It examined publicly reported AI incident records to identify recurring ethical risks in AI-based information systems. This design was suitable because the study organized, interpreted, and classified documented AI incidents by ethical risk pattern, stakeholder, system domain, and severity level.

Descriptive analysis summarized the frequencies and patterns in the dataset. These results became the basis for the Incident-Based Ethical Risk Screening Matrix.

Data source

The study used publicly reported AI incident records as its primary data. The main source was the AI Incident Database, which records real-world harms and near-harms caused by deployed AI systems [1]. The OECD AI incident reporting framework, MIT AI Risk Repository, NIST AI Risk Management Framework, and AI incident reporting studies supported the coding structure [7, 20, 22, 26].

The dataset consisted of 100 AI incident records from 2020 to 2026. The records came from AIID incident IDs 1254 to 1353. These IDs were part of a documented AIID batch of 108 new incidents reported in the November 2025 to January 2026 roundup [2].

Sampling technique

The study used criterion-based purposive sampling because the analysis required incident records with enough detail for ethical risk classification. The goal was to select relevant and usable cases, not to create a statistical sample of all AI incidents.

The sample was limited to 100 AI incident records. This size was enough for frequency analysis, cross-tabulation, severity assessment, and matrix development, while keeping the work manageable.

Inclusion and exclusion criteria

An incident was included if it involved an AI-based information system, was publicly reported, occurred or was reported from 2020 to 2026, had enough detail for coding, and showed at least one ethical risk. These risks included bias, privacy violation, misinformation, safety failure, accountability failure, lack of transparency, misuse, security risk, overreliance, and social harm.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

An incident was excluded if it did not clearly involve AI, lacked enough detail for classification, duplicated another included incident, fell outside the selected coverage period, or described general AI concerns without a specific reported case.

Unit of analysis and coding framework

The unit of analysis was one publicly reported AI incident. Each incident formed one row in the coding sheet.

Each incident was coded by incident number, year, incident title, AI system or platform, short description, source link, system domain, ethical risk type, affected stakeholder, ethical principle violated, severity level, reason for severity rating, and evidence strength.

The system domain categories were education, health care, finance, employment and hiring, law enforcement and surveillance, government and public service, social media and online platforms, transportation, business and customer service, generative AI and chatbots, and others. Ethical risk categories were bias and discrimination, privacy violation, misinformation or deception, safety failure, accountability failure, lack of transparency or explainability, misuse or malicious use, security risk, overreliance or harmful human-AI interaction, and social, economic, or environmental harm.

The ethical principles used in the study were fairness, privacy, transparency, accountability, safety, human oversight, security, and social responsibility. These categories were guided by NIST trustworthy AI characteristics, AIID taxonomies, OECD incident reporting criteria, and MIT AI risk domains [1, 20, 22, 26].

Severity classification

The study used a four-level severity scale. Level 1, Low, referred to minor inconvenience, limited harm, or small scope. Level 2, Moderate, referred to clear but limited or reversible harm. Level 3, High, referred to serious harm, legal risk, financial loss, privacy exposure, discrimination, safety concern, or broad effect. Level 4, Critical, referred to severe harm, major rights violation, widespread public impact, serious safety risk, harm to vulnerable groups, or major loss of public trust.

Data collection and analysis procedure

The researcher identified eligible AI incident records from public AI incident repositories and related public reports. Each selected incident was entered into the coding sheet. Duplicate records, unclear cases, and records with insufficient detail were removed.

The encoded dataset was analyzed using frequency counts and percentages. Cross-tabulation was used to examine patterns between ethical risk type and severity level, system domain and ethical risk type, affected stakeholder and severity level, and ethical principle and risk type.

The frequency tables and cross-tabulations were used to develop the Ethical Risk Screening Matrix. The matrix turned observed patterns into warning indicators, screening questions, risk levels, and suggested management actions.

Ethical considerations

This study used publicly available records only. No private personal data were collected. The analysis focused on incident descriptions, ethical risk categories, system domains, stakeholder groups, and severity levels based on public documentation.

Source links were recorded for traceability. The study avoided unsupported claims and did not assign blame beyond the public record. Severity ratings were based only on available evidence.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

RESULTS

This section presents the results from the 100 encoded AI incident records. The analysis focused on year, system domain, ethical risk type, affected stakeholder, ethical principle, severity level, and selected cross-tabulation patterns. Table 1 presents the summary frequency distribution.

Table 1. Summary Frequency Distribution of the 100 AI Incident Records

Variable	Category	Frequency	Percent
Year	2025	80	80.0%
Year	2026	9	9.0%
Year	2024	6	6.0%
Year	2023	4	4.0%
Year	2020	1	1.0%
System Domain	GEN - Generative AI and chatbots	28	28.0%
System Domain	LAW - Law enforcement and surveillance	16	16.0%
System Domain	FIN - Finance	13	13.0%
System Domain	GOV - Government and public service	13	13.0%
System Domain	EDU - Education	10	10.0%
System Domain	SOC - Social media and online platforms	8	8.0%
System Domain	TRN - Transportation	5	5.0%
System Domain	BUS - Business and customer service	4	4.0%
System Domain	HLT - Health care	3	3.0%
Ethical Risk Type	MISINFO - Misinformation or deception	33	33.0%
Ethical Risk Type	MISUSE - Misuse or malicious use	26	26.0%
Ethical Risk Type	SAFETY - Safety failure	12	12.0%
Ethical Risk Type	PRIV - Privacy violation	12	12.0%
Ethical Risk Type	ACC - Accountability failure	11	11.0%
Ethical Risk Type	BIAS - Bias and discrimination	4	4.0%

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

Ethical Risk Type	SEC - Security risk	2	2.0%
Affected Stakeholder	PUBLIC - General public	60	60.0%
Affected Stakeholder	CUSTOMER - Customer or client	14	14.0%
Affected Stakeholder	CHILD - Children or minors	10	10.0%
Affected Stakeholder	ORG - Organization or company	4	4.0%
Affected Stakeholder	PATIENT - Patient	4	4.0%
Affected Stakeholder	VUL - Vulnerable group	3	3.0%
Affected Stakeholder	STUDENT - Student or learner	2	2.0%
Affected Stakeholder	USER - Individual user	2	2.0%
Affected Stakeholder	EMPLOYEE - Employee or job applicant	1	1.0%
Ethical Principle Violated	SECURE - Security	28	28.0%
Ethical Principle Violated	TRANS - Transparency	22	22.0%
Ethical Principle Violated	SAFE - Safety	12	12.0%
Ethical Principle Violated	PRIVACY - Privacy	12	12.0%
Ethical Principle Violated	ACC - Accountability	11	11.0%
Ethical Principle Violated	RESP - Social responsibility	11	11.0%
Ethical Principle Violated	FAIR - Fairness	4	4.0%
Severity Level	3 - High	75	75.0%
Severity Level	4 - Critical	21	21.0%
Severity Level	2 - Moderate	4	4.0%
Evidence Strength	MODERATE	89	89.0%
Evidence Strength	STRONG	11	11.0%

Year distribution

The dataset covered incidents from 2020 to 2026. Most records were reported in 2025, with 80 cases or 80%. This was followed by 2026 with 9 cases, 2024 with 6 cases, 2023 with 4 cases, and 2020 with 1 case. No encoded

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

record in the final dataset was listed for 2021 or 2022. Since 2026 was still a partial year, its count should be treated as preliminary.

System domain frequency

Generative AI and chatbots recorded the highest number of cases with 28 incidents. Law enforcement and surveillance followed with 16 cases.

Finance and government/public service had 13 cases each. Education had 10 cases, social media and online platforms had 8 cases, transportation had 5 cases, business and customer service had 4 cases, and health care had 3 cases.

These results show higher risk exposure in public-facing systems, surveillance systems, financial systems, and public service systems. These domains involve public trust, personal data, safety, rights, and organizational accountability.

Ethical risk type frequency

Misinformation or deception was the most frequent risk type, with 33 cases. Misuse or malicious use followed with 26 cases. Safety failure and privacy violation had 12 cases each. Accountability failure had 11 cases. Bias and discrimination had 4 cases, while security risk had 2 cases.

The high count for misinformation and deception is linked to generative AI, synthetic media, deepfakes, and automated content generation. The high count for misuse shows that AI risk comes from harmful use by people as well as system error.

Affected stakeholder frequency

The general public was the most affected stakeholder group, with 60 cases. Customers or clients followed with 14 cases. Children or minors had 10 cases. Organizations and patients had 4 cases each. Vulnerable groups had 3 cases. Students or learners and individual users had 2 cases each, while employees or job applicants had 1 case.

This pattern shows that AI ethical risks often extend beyond direct users. Public-facing systems, synthetic media, surveillance tools, and online platforms affect communities and public trust. The presence of minors in the dataset also shows the need for stronger safeguards.

Ethical principle and severity frequency

Security was the most frequently violated ethical principle, with 28 cases. Transparency followed with 22 cases. Safety and privacy had 12 cases each. Accountability and social responsibility had 11 cases each. Fairness had 4 cases.

Most incidents were rated High severity, with 75 cases. Critical incidents accounted for 21 cases, while Moderate incidents accounted for 4 cases. No incident was rated Low severity. This pattern suggests that public AI incident records often capture visible, serious, or widely reported harm.

Table 2 shows the cross-tabulation between ethical risk type and severity level.

Table 2. Ethical Risk Type by Severity Level

Ethical Risk Type	Moderate	High	Critical	Total
MISINFO - Misinformation or deception	4	29	0	33

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

MISUSE - Misuse or malicious use	0	24	2	26
SAFETY - Safety failure	0	6	6	12
PRIV - Privacy violation	0	3	9	12
ACC - Accountability failure	0	10	1	11
BIAS - Bias and discrimination	0	2	2	4
SEC - Security risk	0	1	1	2

Cross-tabulation patterns

Privacy violation had a high share of Critical severity cases. Safety failure also showed serious severity because half of its cases were rated Critical. Bias and discrimination had fewer cases, but half were Critical. These patterns show that privacy, safety, and fairness risks need close review even when they appear less often.

Generative AI and chatbots were often linked to misinformation or deception, accountability failure, and privacy violation. Finance was strongly linked to misuse or malicious use.

Transportation was linked to safety failure. Education was linked mainly to privacy violation and safety failure. Table 3 summarizes the dominant risk pattern by system domain.

Table 3. System Domain and Dominant Ethical Risk Pattern

System Domain	Frequency	Percent	Dominant Ethical Risk Pattern
GEN - Generative AI and chatbots	28	28.0	MISINFO - Misinformation or deception (11); ACC - Accountability failure (6); PRIV - Privacy violation (4)
LAW - Law enforcement and surveillance	16	16.0	MISINFO - Misinformation or deception (5); MISUSE - Misuse or malicious use (4); BIAS - Bias and discrimination (4)
FIN - Finance	13	13.0	MISUSE - Misuse or malicious use (12); MISINFO - Misinformation or deception (1)
GOV - Government and public service	13	13.0	MISINFO - Misinformation or deception (7); MISUSE - Misuse or malicious use (5); SAFETY - Safety failure (1)
EDU - Education	10	10.0	PRIV - Privacy violation (6); SAFETY - Safety failure (3); MISINFO - Misinformation or deception (1)
SOC - Social media and online platforms	8	8.0	MISINFO - Misinformation or deception (5); PRIV - Privacy violation (2); MISUSE - Misuse or malicious use (1)
TRN - Transportation	5	5.0	SAFETY - Safety failure (5)

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

BUS - Business and customer service	4	4.0	MISINFO - Misinformation or deception (3); ACC - Accountability failure (1)
HLT - Health care	3	3.0	ACC - Accountability failure (2); MISUSE - Misuse or malicious use (1)

Deepfake-related ethical risk pattern

Deepfake-related incidents were included only as a supporting risk pattern under generative AI and synthetic media systems. In the encoded dataset, 37 out of 100 records mentioned deepfake or synthetic media. These cases were linked to misinformation, impersonation, fraud, privacy violation, reputational harm, and public manipulation.

Deepfakes were not treated as the main focus of the study. They were included because several records showed AI-generated or manipulated media being used to imitate real persons, distort information, support scams, or target individuals. Prior studies on deepfakes report similar concerns involving deception, impersonation, privacy harm, disinformation, and detection challenges [14, 17, 21, 27]. This keeps the study focused on AI-based information systems while recognizing deepfakes as a recurring ethical risk.

DISCUSSION

The findings show that ethical risks in AI-based information systems appear across several domains. The results support incident-based AI ethics research, which treats documented failures as useful evidence for risk analysis [5, 9, 16, 23]. They also support sociotechnical harm frameworks, which explain AI risk through the interaction of data, models, users, institutions, and affected communities [25, 26].

The high frequency of misinformation, deception, and misuse reflects the growth of generative AI and synthetic media. These systems produce text, images, audio, video, and recommendations. Users sometimes treat their outputs as factual. Institutions sometimes rely on automated content. Malicious actors use the same tools to mislead the public, impersonate people, or commit fraud.

The high number of incidents affecting the general public shows that AI ethical risk is not limited to direct users. Public-facing AI systems often affect communities, institutions, public trust, and vulnerable groups. This supports the need for early screening before AI systems reach broad use.

The high and critical severity ratings also show the need for stronger controls. Privacy violations, safety failures, child-related harms, fraud, surveillance, and public misinformation require careful review. These findings align with risk-based governance approaches such as the NIST AI RMF and the EU AI Act [8, 20].

The findings also show that general AI ethics principles are not enough by themselves. Organizations need tools that turn principles into operational questions. The screening matrix developed in this study addresses this need by linking system domains, risk types, stakeholders, warning indicators, screening questions, and management actions.

Generative AI and large language model studies support this result. These studies identify risks related to scale, hallucination, misuse, user dependence, and downstream adaptation [4, 13, 19, 29]. The matrix helps organizations examine these risks before deployment or expanded use.

Incident-Based Ethical Risk Screening Matrix

The Incident-Based Ethical Risk Screening Matrix was developed from the analysis of 100 publicly reported AI incident records. It turns observed risk patterns into screening questions and management actions. The matrix supports early ethical risk identification before AI systems are adopted, deployed, or expanded.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

Table 4. Incident-Based Ethical Risk Screening Matrix for AI-Based Information Systems, Part 1: Risk Identification

AI System Domain	Common Ethical Risks	Affected Stakeholders	Risk Level
Generative AI and Chatbots	Misinformation, deception, privacy violation, accountability failure, misuse	General public, students, users, organizations	High
Deepfake and Synthetic Media Systems	Identity misuse, deception, fraud, privacy violation, reputational harm	Individuals, public figures, minors, organizations, general public	High to Critical
Law Enforcement and Surveillance	Bias, discrimination, privacy violation, lack of transparency, accountability failure	General public, vulnerable groups, government agencies, individuals	Critical
Finance	Fraud, misuse, privacy violation, security risk, accountability failure	Customers, clients, financial institutions, general public	High
Education	Privacy violation, safety failure, misinformation, overreliance, lack of transparency	Students, teachers, schools, children or minors	High
Health Care	Safety failure, privacy violation, accountability failure, bias, lack of transparency	Patients, health professionals, health institutions	Critical
Government and Public Service	Misinformation, privacy violation, accountability failure, lack of transparency, social harm	Citizens, public agencies, vulnerable groups, general public	High
Social Media and Online Platforms	Misinformation, deception, misuse, social harm, privacy violation	General public, minors, platform users, vulnerable groups	High
Employment and Hiring	Bias, discrimination, lack of transparency, privacy violation, accountability failure	Job applicants, employees, employers, vulnerable groups	High
Transportation	Safety failure, accountability failure, security risk, lack of transparency	Passengers, pedestrians, drivers, transport operators, general public	Critical
Business and Customer Service	Privacy violation, misinformation, accountability failure, overreliance, misuse	Customers, clients, organizations, employees	Moderate to High

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

Table 5. Incident-Based Ethical Risk Screening Matrix for AI-Based Information Systems, Part 2: Screening and Management Actions

AI System Domain	Warning Indicators	Screening Questions	Suggested Management Action
Generative AI and Chatbots	Generates text, images, audio, or video that users may treat as factual information.	Is AI-generated content clearly labeled? Is there human review for sensitive outputs? Who is accountable for harmful output?	Require human review. Label AI content. Create reporting channels. Limit use in high-risk decisions.
Deepfake and Synthetic Media Systems	Imitates a real person's face, voice, image, or likeness.	Was consent obtained? Is the content labeled? Can the source be verified? Is there a takedown process?	Require consent verification, source verification, labels or watermarks, rapid takedown, and identity protection.
Law Enforcement and Surveillance	Identifies, tracks, profiles, scores, or monitors people.	Was the system tested for bias? Is there human review? Is there an appeal or correction process?	Require bias audit, strict human oversight, transparency reports, and appeal mechanisms.
Finance	Handles identity, transactions, credit decisions, fraud detection, or financial data.	Does the system protect financial data? Is fraud monitoring active? Can users challenge automated decisions?	Strengthen identity verification. Maintain audit logs. Monitor fraud patterns. Require human review.
Education	Collects student data, monitors learners, grades outputs, or generates academic content.	Does the system protect student data? Are students informed of AI use? Is there teacher review?	Apply data privacy safeguards. Require teacher oversight. Avoid fully automated high-stakes grading.
Health Care	Supports diagnosis, triage, treatment recommendation, patient monitoring, or health records.	Does the system influence medical decisions? Was it clinically validated? Who is responsible for wrong recommendations?	Require medical expert review, clinical validation, patient data protection, and audit trails.
Government and Public Service	Affects public services, identity verification, eligibility, benefits, or public information.	Can citizens understand and challenge decisions? Is public data protected? Is there a correction process?	Require transparency review, appeal mechanisms, human validation, and public data safeguards.
Social Media and Online Platforms	Recommends content, moderates posts, generates synthetic media, or influences public opinion.	Does the system amplify harmful content? Are AI-generated materials labeled? Are minors protected?	Improve moderation. Label synthetic content. Add reporting channels. Protect minors and vulnerable users.
Employment and Hiring	Screens applicants, ranks resumes, monitors workers,	Was the system tested for bias? Can applicants appeal? Are	Require bias audit. Keep human decision authority.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

	or scores employee performance.	decision criteria explainable?	Provide appeal options. Limit unnecessary surveillance.
Transportation	Supports autonomous driving, routing, navigation, traffic decisions, or safety-critical movement.	Can system failure cause physical harm? Is there human override? Who is responsible during failure?	Require safety testing, emergency protocols, human override, and clear accountability.
Business and Customer Service	Handles customer data, service requests, product recommendations, or automated responses.	Are customers informed when they interact with AI? Can customers reach a human agent? Is customer data protected?	Disclose AI use. Protect customer data. Review AI responses. Provide human support.

Together, Tables 4 and 5 present the screening matrix in portrait format. Table 4 shows risk identification. Table 5 shows screening questions and management actions.

CONCLUSION AND RECOMMENDATIONS

Conclusion

This study developed an Incident-Based Ethical Risk Screening Matrix for AI-Based Information Systems using 100 publicly reported AI incident records from 2020 to 2026. The incidents were classified by system domain, ethical risk type, affected stakeholder, ethical principle violated, severity level, and evidence strength.

The findings show that ethical risks in AI-based information systems recur across different domains. The most common risks were misinformation or deception, misuse or malicious use, safety failure, privacy violation, and accountability failure. These risks appeared most often in generative AI and chatbots, law enforcement and surveillance, finance, government and public service, education, and social media platforms.

The general public was the most affected stakeholder group. This means that AI risk often reaches beyond direct users and affects communities, institutions, public trust, and vulnerable groups. Incidents involving minors, health care, transportation, surveillance, and public services raised stronger severity concerns because they involve safety, privacy, rights, and social accountability.

Most cases were rated High severity. A smaller but important group was rated Critical. This pattern suggests that public AI incident reporting often captures cases with visible harm, public concern, or broad impact.

The study concludes that AI-based information systems need structured ethical screening before adoption, deployment, or expansion. The proposed matrix offers a practical guide for identifying warning indicators, asking screening questions, and selecting management actions.

Recommendations

1. Organizations should conduct ethical risk screening before deploying AI-based information systems. They should check fairness, privacy, transparency, accountability, safety, security, human oversight, and social responsibility.

2. IT managers should use the Ethical Risk Screening Matrix during system planning and procurement. The matrix supports early risk identification before AI systems are acquired, developed, or integrated.

3. High-risk AI domains should undergo stricter review. AI systems used in health care, law enforcement, surveillance, transportation, finance, education, and public service should receive stronger ethical assessment.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

4. Organizations should require human oversight for high-impact AI decisions. AI systems that affect employment, education, health, financial access, public benefits, safety, or legal outcomes should not operate without human review.

5. AI-generated content should be labeled and reviewed in sensitive contexts. Generative AI systems should include disclosure mechanisms, content review, and error reporting channels.

6. Deepfake and synthetic media risks should be controlled through consent verification, source verification, labeling, identity protection, and rapid takedown procedures.

7. Organizations should protect vulnerable groups and minors. Privacy protection, consent, safety checks, and human intervention should receive priority.

8. AI incident reporting should be strengthened. Organizations should document AI-related errors, complaints, harms, and near-harms to support learning and prevention.

9. Future researchers should expand the dataset. Future studies should compare AI incident databases or focus on domains such as education, health care, finance, or generative AI.

10. Future studies should validate the screening matrix through expert review or organizational case studies.

Overall, the study recommends treating AI ethics as a management responsibility. AI systems should be assessed based on their effect on people, organizations, rights, safety, privacy, and public trust.

Data availability

The data used in this study came from publicly available AI incident records. The coded dataset used for the analysis is available in the researchers' working file and will be shared for academic review if required.

Conflict of interest

The authors declare no conflict of interest.

REFERENCES

1. AI Incident Database. (n.d.). Welcome to the Artificial Intelligence Incident Database. https://incidentdatabase.ai/
2. Atherton, D. (2026, February 2). AI Incident Roundup: November and December 2025 and January 2026. AI Incident Database. https://incidentdatabase.ai/blog/incident-report-2025-november-december-2026-january/
3. Autio, C., Schwartz, R., Dunietz, J., Jain, S., Stanley, M., Tabassi, E., Hall, P., & Roberts, K. (2024). Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile (NIST AI 600-1). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.600-1
4. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 610-623. https://doi.org/10.1145/3442188.3445922
5. Burema, D., Debowski-Weimann, N., von Janowski, A., Grabowski, J., Maftei, M., Jacobs, M., van der Smagt, P., & Benbouzid, D. (2023). A sector-based approach to AI ethics: Understanding ethical issues of AI-related incidents within their sectoral context. Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society, 705–714. https://doi.org/10.1145/3600211.3604680
6. Diel, A., Lalgi, T., Mellis, F. S., Teufel, M., & Bäuerle, A. (2025). The harm of deepfakes: A scoping review of deepfakes' negative effects on human mind and behavior. AI & Society. https://doi.org/10.1007/s00146-025-02774-0
7. Dixon, R. B. L., & Frase, H. (2025). AI incidents: Key components for a mandatory reporting regime.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

Center for Security and Emerging Technology. https://doi.org/10.51593/20240023
8. European Parliament & Council of the European Union. (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence. Official Journal of the European Union. https://eur-lex.europa.eu/eli/reg/2024/1689/oj
9. Feffer, M., Martelaro, N., & Heidari, H. (2023). The AI Incident Database as an educational tool to raise awareness of AI harms: A classroom exploration of efficacy, limitations, and future improvements. Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization, Article 3. https://doi.org/10.1145/3617694.3623223
10. Gabriel, I. (2020). Artificial intelligence, values, and alignment. Minds and Machines, 30, 411-437. https://doi.org/10.1007/s11023-020-09539-2
11. Hadan, H., Mogavi, R. H., Zhang-Kennedy, L., & Nacke, L. E. (2025). Who is responsible when AI fails? Mapping causes, entities, and consequences of AI privacy and ethical incidents. International Journal of Human-Computer Interaction. https://doi.org/10.1080/10447318.2025.2549073
12. Helmus, T. C. (2022). Artificial intelligence, deepfakes, and disinformation: A primer. RAND Corporation. https://doi.org/10.7249/PEA1043-1
13. Ji, Z., Lee, N., Frieske, R. M., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), Article 248. https://doi.org/10.1145/3571730
14. Kietzmann, J., Lee, L. W., McCarthy, I. P., & Kietzmann, T. C. (2020). Deepfakes: Trick or treat? Business Horizons, 63(2), 135-146. https://doi.org/10.1016/j.bushor.2019.11.006
15. Knight, S., McGrath, C., Viberg, O., & Cerratto Pargman, T. (2025). Learning about AI ethics from cases: A scoping review of AI incident repositories and cases. AI and Ethics, 5, 2037-2053. https://doi.org/10.1007/s43681-024-00639-8
16. McGregor, S. (2021). Preventing repeated real-world AI failures by cataloging incidents: The AI Incident Database. Proceedings of the AAAI Conference on Artificial Intelligence, 35(17), 15458-15463. https://doi.org/10.1609/aaai.v35i17.17817
17. Mirsky, Y., & Lee, W. (2021). The creation and detection of deepfakes: A survey. ACM Computing Surveys, 54(1), Article 7. https://doi.org/10.1145/3425780
18. Mökander, J., & Floridi, L. (2021). Ethics-based auditing to develop trustworthy AI. Minds and Machines, 31(2), 323–327. https://doi.org/10.1007/s11023-021-09557-8
19. Mökander, J., Schuett, J., Kirk, H. R., & Floridi, L. (2024). Auditing large language models: A three-layered approach. AI and Ethics, 4, 1085-1115. https://doi.org/10.1007/s43681-023-00289-2
20. National Institute of Standards and Technology. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0) (NIST AI 100-1). U.S. Department of Commerce. https://doi.org/10.6028/NIST.AI.100-1
21. Nguyen, T. T., Nguyen, Q. V. H., Nguyen, D. T., Nguyen, D. T., Huynh-The, T., Nahavandi, S., Nguyen, T. T., Pham, Q. V., & Nguyen, C. M. (2022). Deep learning for deepfakes creation and detection: A survey. Computer Vision and Image Understanding, 223, Article 103525. https://doi.org/10.1016/j.cviu.2022.103525
22. Organisation for Economic Co-operation and Development. (2025). Towards a common reporting framework for AI incidents (OECD Artificial Intelligence Papers, No. 34). OECD Publishing. https://doi.org/10.1787/f326d4ac-en
23. Paeth, K., Atherton, D., Pittaras, N., Frase, H., & McGregor, S. (2025). Lessons for editors of AI incidents from the AI Incident Database. Proceedings of the AAAI Conference on Artificial Intelligence, 39(28), 28946-28953. https://doi.org/10.1609/aaai.v39i28.35163
24. Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20) (pp. 33–44). Association for Computing Machinery. https://doi.org/10.1145/3351095.3372873
25. Shelby, R., Rismani, S., Henne, K., Moon, A., Rostamzadeh, N., Nicholas, P., Yilla-Akbari, N., Gallegos, J., Smart, A., Garcia, E., & Virk, G. (2023). Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction. Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

Society, 723-741. https://doi.org/10.1145/3600211.3604673

26. 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. arXiv. https://doi.org/10.48550/arXiv.2408.12622
27. Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., & Ortega-Garcia, J. (2020). Deepfakes and beyond: A survey of face manipulation and fake detection. Information Fusion, 64, 131-148. https://doi.org/10.1016/j.inffus.2020.06.014
28. Vaccari, C., & Chadwick, A. (2020). Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. Social Media + Society, 6(1). https://doi.org/10.1177/2056305120903408
29. Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P.-S., Mellor, J., Glaese, A., Cheng, M., Balle, B., Kasirzadeh, A., Biles, C., Brown, S., Kenton, Z., Hawkins, W., Stepleton, T., Birhane, A., Hendricks, L. A., Rimell, L., Isaac, W., ... Gabriel, I. (2022). Taxonomy of risks posed by language models. Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, 214-229. https://doi.org/10.1145/3531146.3533088
30. Wei, M., & Zhou, Z. (2022). AI ethics issues in real world: Evidence from AI Incident Database. arXiv. https://doi.org/10.48550/arXiv.2206.07635