

FRAMEWORK-DRIVEN GUIDELINE GENERATION FOR AI ADOPTION: A RISK-BASED PERSPECTIVE

DAVID LAU KEAT JIN¹, GANTHAN NARAYANA SAMY², FIZA ABDUL RAHIM³,
MAHISWARAN SELVANANTHAN⁴, NURAZEAN MAAROP⁵, MUGILRAJ RADHA
KRISHNAN⁶ & SUNDRESAN PERUMAL⁷

^{1,6}Researcher, Faculty of Artificial Intelligence, Universiti Teknologi Malaysia, Malaysia

^{2,3,5}Lecturer, Faculty of Artificial Intelligence, Universiti Teknologi Malaysia, Malaysia

⁴Lecturer, Faculty of Social Sciences and Humanities, Universiti Teknologi Malaysia

⁷Faculty of Science and Technology, Universiti Sains Islam Malaysia, Malaysia

¹davidkeat@graduate.utm.my, ²ganthan.kl@utm.my, ³fiza.abdulrahim@utm.my, ⁴mahiswaran@utm.my,

⁵nurazean.kl@utm.my, ⁶mugilraj2@graduate.utm.my, ⁷sundresan.p@usim.edu.my

ABSTRACT

The adoption of artificial intelligence (AI) presents unique risks that existing frameworks inadequately address, including issues of accountability, accuracy, fairness, safety, and privacy. According to AI Incident Database, there is an increase of 156% of published AI incidents from the year 2020 to 2024. This study bridges the gap between reported AI incidents and actionable countermeasures by analyzing an AI incident repository and contextualizing risks with mitigative strategies drawn from the literature. A knowledge graph was developed to integrate contextual data, risks, and countermeasures, enabling the generation of customizable, risk-based guidelines tailored to specific applications and stakeholders. Key findings include the identification of countermeasures for diverse AI risks, emphasizing the need for systematic risk assessment throughout the AI life cycle. The developed prototype serves as both a risk assessment tool and risk reference database in an enhanced enterprise risk management framework which facilitates responsible AI adoption, guiding developers, risk managers, and policymakers in advancing ethical and sustainable AI practices. This work lays the groundwork for automated tools that enhance scalability and usability in addressing AI risks in various organizational contexts.

Keywords: *Responsible AI; Risk; Countermeasure; Framework; Guideline*

1. INTRODUCTION

The growing reliance on artificial intelligence (AI) technologies introduces unique risks that are not fully addressed by existing risk management frameworks [1] [2]. Unlike traditional systems, the nondeterministic nature of AI outputs creates challenges in evaluating their reliability, fairness, and safety [3]. Current research highlights various approaches to responsible AI, including the European Union's AI Act, which categorizes risks into unacceptable, high, limited, and minimal levels [4] [5], alongside other guidelines provided by organizations such as the World Economic Forum [6] and ISO [7]. However, these efforts primarily focus on high-level principles and compliance requirements, leaving a significant gap in providing actionable, context-specific countermeasures for managing AI risks effectively. This is reiterated in a study that mapped the provisions in NIST NSF 2.0, COBIT 2019, ISO 27001:2022 and ISO 42001:2023 to risks of Large Language Model (LLM) which

indicated significant gaps in risk management [8]. More recently, LLM is leveraged in autonomous decision-making in various forms of agentic workflow [9]. In this regard, AI agents are equipped to learn, reason, and update their knowledge bases dynamically. In fact, AI agents will manage production lines, optimize supply-chain operations with minimum human supervision and handle customer support and fraud detection by the year 2028, where 33% of enterprise software applications are expected to include agentic AI [10]. However, existing frameworks struggle to model the unpredictable actions of autonomous agents arising from their independent decision-making [11, 12].

Although incident repositories such as the AI Incident Database (AIID) [13] and the OECD AI Incidents Monitor [14] provide valuable documentation of real-world AI harms, they fail to offer targeted recommendations or frameworks for mitigating these issues in practice. Moreover, the published AI incidents has increased from 109 in 2020 to 279 in 2024 which represents an increased

of 156% in the period. This prompted prior studies on AI governance to introduce conceptual frameworks, such as the AI TRiSM [15] and SOTEC models [16]. However, these frameworks lack systematic methods for integrating real-world data to inform lifecycle-based risk assessments. Similarly, ethical risk management frameworks, such as the enhanced Enterprise Risk Management (ERM) model, address broad organizational risks but do not comprehensively incorporate AI-specific technological and analytical risks [17]. In fact, the study asserted that when ethical risks were identified, no solutions were at hand which led to the formation of an enhanced ERM framework in accordance with Figure 1. To bridge these gaps, this study introduces a framework-driven approach to generating customized, risk-based guidelines for AI adoption that supports the formation of a risk assessment tool and risk reference database as envisaged in Figure 1.

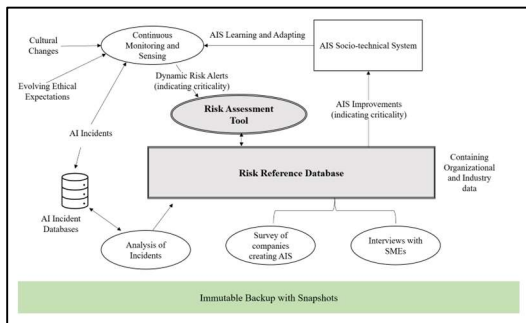


Figure 1: Enhanced ERM Framework [17]

Based on published standards [3, 7, 18] and guidelines [6, 19, 20], a proposed risk-based guideline should consider the dimensions of risk management processes, phases in the AI life cycle, and stakeholders that are responsible for the required activities. The three-dimensional approach is illustrated in Figure 2.

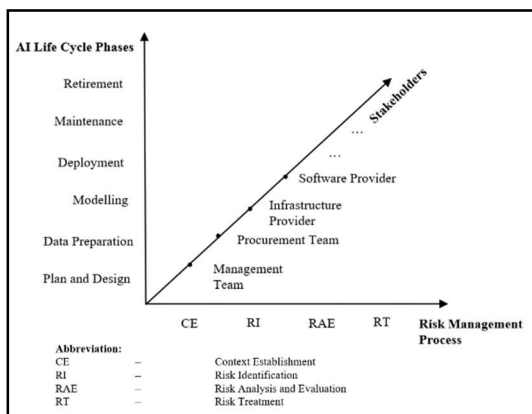


Figure 2: Three-Dimensional Consideration for a Risk-Based Guideline

[17] analyzed 277 responses from 229 businesses to identify gaps in enterprise risk management (ERM) practices and proposed an enhanced ERM framework focused on ethical risks. On the other hand, [16] categorized risks in autonomous and intelligent systems (AIS) in the automotive and healthcare domains using the SOTEC framework (structure, organizational, technological, epidemiological and cultural), but did not address the need to evaluate risks throughout the AI life cycle.

[15] proposed the Artificial Intelligence Trust, Risk, and Security Management (AI TRiSM) framework, highlighting challenges and potential improvements, including adversarial attacks and related threats. Although it emphasized adaptability and scalability for evolving AI technologies, it lacked specific references to stakeholder roles. [21] identified metrics to measure sustainability, accuracy, fairness, and explainability—key factors opposing AI risks, and provided tests for the AI life cycle. However, it did not identify stakeholders as risk owners or assign responsibility for countermeasures. [22] presented computational methods for risk analysis with examples for Automated Driving Systems (ADS).

Table 1 summarizes the composition of these related studies in terms of these dimensions. Each study contributed to risk management processes within an AI context, with the "in context" criterion assessing whether its approaches were validated using real incidents, field settings, or specific AI models. With the exception of [17] which focused on non-technical risks, none of the listed studies in Table 1 fully account for the three-dimensional aspects of the risk management process, AI life cycle, and stakeholders.

Table 1:

Comparative summary of related studies

Ref.	Dimensional Consideration			
	Risk Management Process	AI Life Cycle	Stakeholder (Risk Owner)	Applied the proposed solution in context
[15]	√	√	X	√
[16]	√	X	√	√
[17]	√	√	√	√
[21]	√	√	X	√
[22]	√	X	X	√

This research aims to develop a structured framework that enables organizational

procurement teams to systematically identify risks and determine effective mitigation strategies during the implementation of AI systems. This study aligns risk management activities with the AI life cycle phases, ensuring continuous monitoring and evaluation. Additionally, it seeks to establish clear role delineation and accountability among internal and external stakeholders—such as data, tools, model, and infrastructure providers—to enhance the reliability and governance of AI deployment. Hence, this study aims to address the following research questions:

1. How can actionable countermeasures for AI-related risks be identified through the analysis of real-world incidents?
2. How can the generation of risk-based guidelines be facilitated in a manner that is dependent upon the selected context, identified risks, and corresponding countermeasures?
3. How can the integration of the AI life cycle and stakeholder responsibilities be ensured within the development of these guidelines?

2. MATERIALS AND METHODS

In this study, the AIAAIC repository was used due its ease of downloading as well as the availability of crucial information related to the sector and technology. In total, there were twelve steps involved in this study which can be divided into three categories, as illustrated in Figure 3. As of 24 June 2024, the downloaded repository contained 1,534 rows with 16 fields or columns.

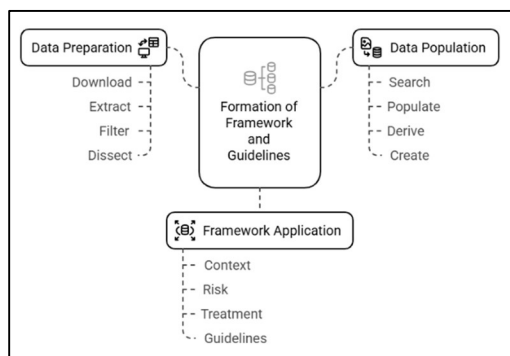


Figure 3: Research procedure

A. Data Preparation

The AI incident repository, sourced from the AIAAIC website (<https://www.aiaaic.org>), contained 16 fields detailing application contexts, associated risks, and other metadata. The

technology field served as a basis to examine the use cases and the associated issues reported. The dataset was prepared by extracting unique technology entries, splitting composite entries into discrete rows, and renaming the 'technology' field to 'application' for clarity. Multiple issues listed in a single column were similarly separated into individual entries, with the 'issue(s)' field renamed 'risk' to emphasize their relevance in risk management. Table 2 presents a sample record from the AIAAIC repository. The record illustrates the structure and content of an AI-related incident in the online database prior to any data processing implemented in this study.

Table 2: A Sample Record from AIAAIC Repository

AIAAIC ID	AIAAIC1539
Headline	Dream Machine AI video generator makes porn
Type	Issue
Released	2024
Occured	2024
Country(ies)	USA
Sector(s)	Media / entertainment / sports / art
Deployer(s)	-
Developer(s)	Luma Labs
System name(s)	Dream Machine
Technology(ies)	Text-to-video
Purpose(s)	Generate video
Media trigger(s)	User comments/complaints
Issue(s)	Privacy; Safety
Transparency	Governance
Description/links	https://www.aiaaic.org/aiaaic-repository/ai-algorithmic-and-automation-incidents/dream-machine-ai-video-generator-makes-porn

Table 3 illustrates the format of the table in the end of the data preparation exercise. Multiple records may be produced after this exercise for each original record as the distinct application and risk are separated.

Table 3: A sample record after data preparation

application	Text-to-video
purpose	Generate video
sector	Media / entertainment / sports / art
risk name	Privacy
risk phase	{human annotated}
ctms name	{curated from literature}
ctms phase	{human annotated}
stakeholder	{human annotated}

B. Data Population

As discussed in Section I, this study focused on two specific groups of risks. Thus, issues related to competition, collusion, malicious use of AI, surveillance for national interests, employment, human rights, and environmental hazards were excluded. Terms in the "issue" field, such as "accuracy," "reliability," and "appropriateness," were converted to risk-related terms like "inaccuracy," "unreliability," and "requirements gap." Entries lacking descriptive URLs as detailed references were also excluded.

A 'countermeasure' is defined as a targeted method to address specific risks associated with AI applications. For each risk-application pair, countermeasures were identified using systematic literature searches on platforms like Google Scholar, with a focus on studies cited in at least five peer-reviewed sources to ensure validity. Keywords were adapted as needed to capture related concepts. The resulting countermeasures were mapped to phases in the AI lifecycle and categorized by responsible stakeholders, ensuring a structured and actionable dataset. For example, for the application "Automatic License Plate Recognition" (ALPR) and the risk "inaccuracy," searches included terms like "ALPR and inaccuracy," and if no results were found, "ALPR and accuracy" or "ALPR and accurate" were used. The research tool *Publish or Perish* facilitated this search [23].

To ensure practicality and potential adoption, the selected countermeasure had to be cited in at least five other studies. Once a suitable countermeasure was identified, the search was stopped, and an additional reference citing it was included for verification. Table 3 serves as a practical blueprint for constructing a knowledge graph that maps complex relationships between AI risks and countermeasures. It also bridges the gap between theoretical frameworks and their application in real-world scenarios. Referring to Table 4, the attributes were translated into label of nodes while the fields were used to form properties in the associated nodes in the knowledge graph.

Table 4: The information required for the knowledge graph

Attribute	Field
Context	Application, Purpose, and Sector
Risk	Name, Phase
Countermeasure	Name, Phase
Stakeholder	Name

To account for the many-to-many relationships between the entities required in this

study, a graph-based method was chosen due to its effectiveness [19]. Hence, a knowledge graph was created using the Neo4j Desktop and Python programming language using VSCode as code editor. The source code is available as a reference for interested readers at the website: <https://www.github.com/renaissance2005/fd-guidelines>.

C. Framework Application

The framework integrated a knowledge graph with a local Large Language Model (LLM) to generate dynamic, context-specific guidelines. Users interact with the system through a series of selections: (1) defining the application context, (2) identifying relevant risks, and (3) choosing countermeasures for each risk. The knowledge graph dynamically retrieves data at each step, guiding user inputs and informing the LLM. The system outputs guidelines tailored to the user's specified parameters, enhancing the framework's practicality for diverse organizational needs. The LLM used was the Llama 3.1:8b variant, running via the Ollama platform [24]. Llama 3.1 was chosen for its open-source nature and ability to run locally, which ensures data privacy and avoids the latency and cost of external APIs. In fact, a study showed that it performed better than GPT-3.5, a proprietary model [25]. While GPT-4 or Claude 2 offer advanced capabilities, they require cloud access and entail usage fees, making them less practical for this implementation. At the time of this study, Llama 3.1 was the latest open-source LLM released by Meta AI and available for download from Hugging Face website (<http://www.huggingface.co>). Ollama provided a user-friendly interface to run the local LLM with minimal configuration.

The choice to use a local LLM and database ensured confidentiality and avoided rate-limit issues common with proprietary LLMs. Figure 4 illustrates the system setup to generate risk-based guidelines. The system allowed users to make selections from information retrieved from the knowledge graph, which was then passed to the local LLM to generate coherent guideline sentences.

To use the system, the user selected the application context, followed by the risks to consider. They then chose countermeasures for each selected risk. Based on these inputs, the system generated guidelines by combining context, risks, and countermeasures. Information from the knowledge graph was retrieved at each

step to support the user's selections and guide the guideline-generation process.

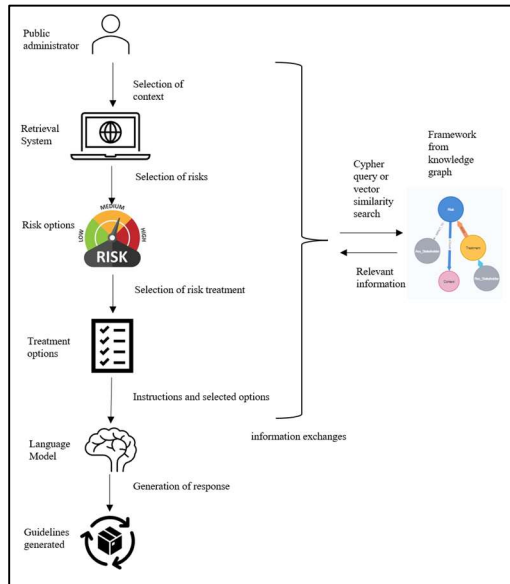
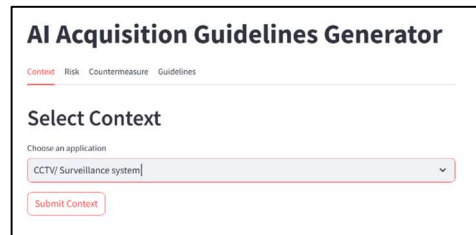


Figure 4: Generation of framework-driven guideline

3. RESULTS AND DISCUSSION

From 1,534 list of issues recorded in the repository, there were 88 distinct lists of technology entries extracted, of which 38 unique entries were curated and renamed as application. Additionally, the entries with countermeasure obtained from extant literature are listed in Appendix 1. This table is a critical resource for practitioners, offering actionable solutions for common AI risks across diverse applications, such as chatbots, autonomous systems, and surveillance technologies. Due to conceptual fuzziness in AI, there were different combinations of keyword applied on the full text of the article to search for the relevant countermeasure [26].

Consequently, the data were entered into an excel file which was used to generate the knowledge graph. The excel file and Python code used to build the system are given in the same preceding URL. In a nutshell, the system developed was divided into 4 tabs, with the first three tabs involve selections by the user while the final tab generated the guidelines based on prior selections. Figure 5 depicts the sample view from the 4 tabs as executed.



AI Acquisition Guidelines Generator

Context Risk Countermeasure Guidelines

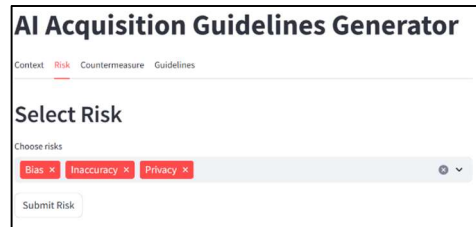
Select Context

Choose an application

CCTV/ Surveillance system

Submit Context

(a)



AI Acquisition Guidelines Generator

Context Risk Countermeasure Guidelines

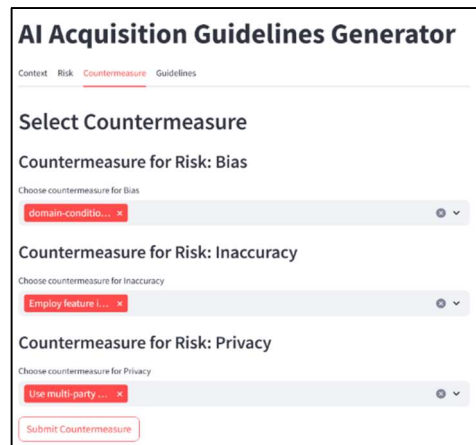
Select Risk

Choose risks

Bias X Inaccuracy X Privacy X

Submit Risk

(b)



AI Acquisition Guidelines Generator

Context Risk Countermeasure Guidelines

Select Countermeasure

Countermeasure for Risk: Bias

Choose countermeasure for Bias

domain-conditions...

Countermeasure for Risk: Inaccuracy

Choose countermeasure for Inaccuracy

Employ feature I...

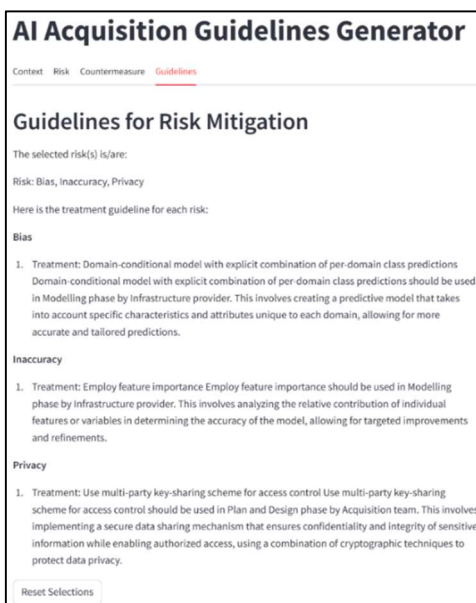
Countermeasure for Risk: Privacy

Choose countermeasure for Privacy

Use multi-party...

Submit Countermeasure

(c)



AI Acquisition Guidelines Generator

Context Risk Countermeasure Guidelines

Guidelines for Risk Mitigation

The selected risk(s) is/are:

Risk: Bias, Inaccuracy, Privacy

Here is the treatment guideline for each risk:

Bias

1. Treatment: Domain-conditional model with explicit combination of per-domain class predictions

Domain-conditional model with explicit combination of per-domain class predictions should be used in Modelling phase by Infrastructure provider. This involves creating a predictive model that takes into account specific characteristics and attributes unique to each domain, allowing for more accurate and tailored predictions.

Inaccuracy

1. Treatment: Employ feature importance

Employ feature importance should be used in Modelling phase by Infrastructure provider. This involves analyzing the relative contribution of individual features or variables in determining the accuracy of the model, allowing for targeted improvements and refinements.

Privacy

1. Treatment: Use multi-party key-sharing scheme for access control

Use multi-party key-sharing scheme for access control should be used in Plan and Design phase by Acquisition team. This involves implementing a secure data sharing mechanism that ensures confidentiality and integrity of sensitive information while enabling authorized access, using a combination of cryptographic techniques to protect data privacy.

Reset Selections

(d)

Figure 5: The AI Acquisition Guidelines Generator with (a) Context (b) Risk (c) Countermeasure (d) Guidelines

As highlighted in Table 1, none of the previous studies considered all three dimensions of the risk management process, the AI life cycle, and stakeholder participation. In addition, the risks analyzed in this study were derived from real-world incidents reported in an online repository. Furthermore, the applicable risks and countermeasures are context-dependent, as the options are displayed to users based on the selected context, thereby minimizing human error and guiding attention toward the relevant risks and controls.

Comparatively, this study advanced the work of previous studies in the following manner:

- i. **Advancements over Ethical Risk Frameworks [17, 20]:** This study not only highlighted ethical risks but also provided actionable countermeasures and stakeholder-specific guidance, addressing the operational gap in [17] and [20].
- ii. **Life-cycle Integration [15, 16]:** While [15] and [16] provided theoretical risk management frameworks, this study incorporated life-cycle considerations which facilitates monitoring and assessment of required activities.
- iii. **Stakeholder-specific Roles [15, 21]:** This study surpasses [15] and [21] by clearly defining stakeholder responsibilities, a critical component for operationalizing AI governance that is often overlooked in existing models.
- iv. **Use of Real-world Data [15, 22]:** Unlike [15] and [22], which primarily rely on conceptual models, this study integrates real-world data from the AIAAIC repository, underscoring its practical benefits.

4. LIMITATIONS AND WAY FORWARD

Admittedly, an empirical comparison with other frameworks was not available at the time of this study, as it would require the practical application of all stated frameworks in a real-world environment. Additionally, this study only considered the various academic databases available for the author's institution which include ArXiv, Scopus, Web of Science, IEEE Explore and ScienceDirect. In addition, the search for risk

controls and countermeasures were not exhaustive as it was not the intention of this study to produce a catalog of mitigation strategies for all available risks.

Future enhancement may explore automated scrapping of published incidents and risks as well as extractions of abstracts from academic papers regarding the risk mitigation strategies and controls proposed. Additionally, LLM can be used to derive the associated stages in the life cycle as well as stakeholders without the requirement for human annotation.

5. CONCLUSION

In a nutshell, this study made the following contributions to the field of responsible AI:

1. **Context-specific Countermeasures:** By linking AI incidents to specific risks and identifying actionable countermeasures, the research addresses the lack of operational solutions in existing guidelines.
2. **Lifecycle-based Risk Assessment:** The proposed framework accounts for risks and mitigation strategies throughout the entire AI life cycle, ensuring a holistic approach to responsible AI governance.
3. **Stakeholder Integration:** Unlike previous frameworks, this study explicitly identifies stakeholders responsible for implementing countermeasures, ensuring accountability and operational clarity.
4. **Practical Applicability:** The system integrates a local LLM for generating customized guidelines, providing a scalable and adaptable tool for organizations to manage AI risks effectively.

By using this framework, procurement team, risk managers, project managers, internal developers and maintenance team will be able to set the required expectations regarding the risk pertaining to the use of AI in the organizations. By addressing both ethical and technological risks, the proposed framework extends beyond the scope of existing frameworks and guidelines, contributing to the development of adaptive and actionable solutions for responsible AI adoption.

ACKNOWLEDGMENT

We would like to thank the Public Services Department of Malaysia for sponsoring this study.

REFERENCES

- [1] OWASP. "OWASP Top 10 for Large Language Model Applications." <https://owasp.org/www-project-top-10-for-large-language-model-applications/> accessed December 30, 2023.
- [2] MITRE. "ATLAS Matrix." <https://atlas.mitre.org/matrices/ATLAS> accessed 4 April, 2024.
- [3] NIST, "Artificial Intelligence Risk Management Framework (AI RMF 1.0)," 2023.
- [4] J. Laux, S. Wachter, and B. Mittelstadt, "Trustworthy artificial intelligence and the European Union AI act: On the conflation of trustworthiness and acceptability of risk," *Regulation and Governance*, Article vol. 18, no. 1, pp. 3-32, 2024, doi: 10.1111/rego.12512.
- [5] E. Parliament. "EU AI Act: first regulation on artificial intelligence." <https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence> accessed March 2, 2024.
- [6] WEF. "Guidelines for AI Procurement." https://www3.weforum.org/docs/WEF_Guidelines_for_AI_Procurement.pdf accessed 31 May 2023.
- [7] ISO/IEC 22989-Information technology — Artificial intelligence — Vocabulary, ISO/IEC, 2022.
- [8] T. R. McIntosh *et al.*, "From cobit to iso 42001: Evaluating cybersecurity frameworks for opportunities, risks, and regulatory compliance in commercializing large language models," *Computers & Security*, vol. 144, p. 103964, 2024.
- [9] A. Kumar, "Building Autonomous AI Agents based AI Infrastructure."
- [10] Gartner. "Intelligent Agents In AI Really Can Work Alone. Here's How. ." <https://www.gartner.com/en/articles/intelligent-agent-in-ai> accessed April 21, 2025.
- [11] I. D. S. Portugal, P. Alencar, and D. Cowan, "An Agentic AI-based Multi-Agent Framework for Recommender Systems," in *2024 IEEE International Conference on Big Data (BigData)*, 2024: IEEE, pp. 5375-5382.
- [12] A. Chan *et al.*, "Infrastructure for AI Agents," *arXiv preprint arXiv:2501.10114*, 2025.
- [13] M. Feffer, N. Martelaro, and H. Heidari, "The AI Incident Database as an Educational Tool to Raise Awareness of AI Harms: A Classroom Exploration of Efficacy, Limitations, & Future Improvements," in *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, 2023, pp. 1-11.
- [14] R. Rowena, R. Anais, and S. Nicole, "When Artificial Intelligence Fails: The Emerging Role of Incident Databases," *Public Governance, Administration and Finances Law Review*, vol. 8, no. 2, pp. 17-28, 2023.
- [15] A. Habbal, M. K. Ali, and M. A. Abuzaraida, "Artificial Intelligence Trust, risk and security management (AI trism): Frameworks, applications, challenges and future research directions," *Expert Sys Appl*, vol. 240, p. 122442, 2024.
- [16] C. Macrae, "Managing risk and resilience in autonomous and intelligent systems: Exploring safety in the development, deployment, and use of artificial intelligence in healthcare," *Risk Anal.*, 2024 Jan 2024, doi: 10.1111/risa.14273.
- [17] Q. McGrath, A. R. Hevner, and G.-J. de Vreede, "Managing Ethical Risks of Artificial Intelligence in Business Applications," *Authorea Preprints*, 2024.
- [18] *ISO 31000:2018 Risk management — Guidelines*, ISO, 2018.
- [19] R. S. Ross, "Guide for conducting risk assessments," 2012.
- [20] M. Yurrita, D. Murray-Rust, A. Balayn, and A. Bozzon, "Towards a multi-stakeholder value-based assessment framework for algorithmic systems," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022, pp. 535-563.
- [21] P. Giudici, M. Centurelli, and S. Turchetta, "Artificial Intelligence risk measurement," *Expert Sys Appl*, vol. 235, p. 121220, 2024.
- [22] J. M. Camacho, A. Couce-Vieira, D. Arroyo, and D. R. Insua, "A Cybersecurity Risk Analysis Framework for Systems with Artificial Intelligence Components," *arXiv preprint arXiv:2401.01630*, 2024.
- [23] A.-W. Harzing, *The publish or perish book*. Tarma Software Research Pty Limited Melbourne, Australia, 2010.
- [24] Meta. "Introducing Llama 3.1: Our most capable models to date." <https://ai.meta.com/blog/meta-llama-3-1/> accessed July 26, 2024.
- [25] A. Purwar, "Evaluating the efficacy of open-source llms in enterprise-specific rag systems: A comparative study of

- performance and scalability," *arXiv preprint arXiv:2406.11424*, 2024.
- [26] B. C. Stahl *et al.*, "A systematic review of artificial intelligence impact assessments," *Artificial Intelligence Review*, pp. 1-33, 2023.
- [27] S. Al-Maadeed, R. Boubezari, S. Kunhoth, and A. Bouridane, "Robust feature point detectors for car make recognition," *Computers in Industry*, vol. 100, pp. 129-136, 2018.
- [28] J. Tang, L. Wan, J. Schooling, P. Zhao, J. Chen, and S. Wei, "Automatic number plate recognition (ANPR) in smart cities: A systematic review on technological advancements and application cases," *Cities*, vol. 129, p. 103833, 2022.
- [29] T. Mustafa and M. Karabatak, "Deep Learning Model for Automatic Number/License Plate Detection and Recognition System in Campus Gates," in *2023 11th International Symposium on Digital Forensics and Security (ISDFS)*, 2023: IEEE, pp. 1-5.
- [30] B. Custers, M. A. Hasani, and B. Aishah, "Privacy Issues of Traffic Monitoring," in *Abstracts & Proceedings of the 3rd Conference on Law & Technology (CLT3, 2008)*, Kuala Lumpur, Malaysia, 2008: Unknown Publisher, pp. 85-94.
- [31] B. C. Newell, "Local law enforcement jumps on the big data bandwagon: Automated license plate recognition systems, information privacy, and access to government information," *Me. L. Rev.*, vol. 66, p. 397, 2013.
- [32] P. Varsha, "How can we manage biases in artificial intelligence systems—A systematic literature review," *International Journal of Information Management Data Insights*, vol. 3, no. 1, p. 100165, 2023.
- [33] R. Gupta, K. Nair, M. Mishra, B. Ibrahim, and S. Bhardwaj, "Adoption and impacts of generative artificial intelligence: Theoretical underpinnings and research agenda," *International Journal of Information Management Data Insights*, vol. 4, no. 1, p. 100232, 2024.
- [34] B. Jiao, "Anti-motion interference wearable device for monitoring blood oxygen saturation based on sliding window algorithm," *IEEE Access*, vol. 8, pp. 124675-124687, 2020.
- [35] M. F. Ahmed, M. O. Ali, M. H. Rahman, and Y. M. Jang, "Real-time health monitoring system design based on optical camera communication," in *2021 International Conference on Information Networking (ICOIN)*, 2021: IEEE, pp. 870-873.
- [36] C. H. Li, Y. T. Yu, S. L. Song, and Y. C. Liu, "A Digital Filtering Approach in Measurement and Control System," *Advanced Materials Research*, vol. 712, pp. 2615-2618, 2013.
- [37] K. Ahmad, M. Abdelrazek, C. Arora, M. Bano, and J. Grundy, "Requirements engineering for artificial intelligence systems: A systematic mapping study," *Information and Software Technology*, vol. 158, p. 107176, 2023.
- [38] K. Ahmad, M. Abdelrazek, C. Arora, M. Bano, and J. Grundy, "Requirements practices and gaps when engineering human-centered Artificial Intelligence systems," *Applied Soft Computing*, vol. 143, p. 110421, 2023.
- [39] T. Cui *et al.*, "Risk taxonomy, mitigation, and assessment benchmarks of large language model systems," *arXiv preprint arXiv:2401.05778*, 2024.
- [40] S. Quach, P. Thaichon, K. D. Martin, S. Weaven, and R. W. Palmatier, "Digital technologies: tensions in privacy and data," *Journal of the Academy of Marketing Science*, vol. 50, no. 6, pp. 1299-1323, 2022.
- [41] H. U. Keval and M. A. Sasse, "Can we ID from CCTV? Image quality in digital CCTV and face identification performance," in *Mobile Multimedia/Image Processing, Security, and Applications 2008*, 2008, vol. 6982: SPIE, pp. 189-203.
- [42] L. Moccagatta and H. Chen, "MPEG-4 visual texture coding: more than just compression," in *1999 Digest of Technical Papers. International Conference on Consumer Electronics (Cat. No. 99CH36277)*, 1999: IEEE, pp. 302-303.
- [43] Z. Wang *et al.*, "Towards fairness in visual recognition: Effective strategies for bias mitigation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 8919-8928.
- [44] T. Garg, S. Masud, T. Suresh, and T. Chakraborty, "Handling bias in toxic speech detection: A survey," *ACM Computing Surveys*, vol. 55, no. 13s, pp. 1-32, 2023.
- [45] A. Costin, "Security of CCTV and Video Surveillance Systems: Threats, Vulnerabilities, Attacks, and Mitigations," presented at the Proceedings of the 6th International Workshop on Trustworthy

- Embedded Devices, Vienna, Austria, 2016. [Online]. Available: <https://doi.org/10.1145/2995289.2995290>.
- [46] J. Kim, D. Lee, and N. Park, "CCTV-RFID enabled multifactor authentication model for secure differential level video access control," *Multimedia Tools and Applications*, vol. 79, pp. 23461-23481, 2020.
- [47] A. Castiglione, M. Cepparulo, A. De Santis, and F. Palmieri, "Towards a Lawfully Secure and Privacy Preserving Video Surveillance System," in *E-Commerce and Web Technologies*, Berlin, Heidelberg, F. Buccafurri and G. Semeraro, Eds., 2010// 2010: Springer Berlin Heidelberg, pp. 73-84.
- [48] P. Lewis *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459-9474, 2020.
- [49] D. Calvaresi, S. Eggenschwiler, Y. Mualla, M. Schumacher, and J.-P. Calbimonte, "Exploring agent-based chatbots: a systematic literature review," *Journal of ambient intelligence and humanized computing*, vol. 14, no. 8, pp. 11207-11226, 2023.
- [50] I. O. Gallegos *et al.*, "Bias and fairness in large language models: A survey," *Computational Linguistics*, pp. 1-79, 2024.
- [51] A. Barrón-Cedeño, P. Gupta, and P. Rosso, "Methods for cross-language plagiarism detection," *Knowledge-Based Syst.*, vol. 50, pp. 211-217, 2013.
- [52] T. Foltýnek, N. Meuschke, and B. Gipp, "Academic plagiarism detection: a systematic literature review," *ACM Computing Surveys (CSUR)*, vol. 52, no. 6, pp. 1-42, 2019.
- [53] H. Zhong *et al.*, "Copyright protection and accountability of generative ai: Attack, watermarking and attribution," in *Companion Proceedings of the ACM Web Conference 2023*, 2023, pp. 94-98.
- [54] S. S. Ghosal, S. Chakraborty, J. Geiping, F. Huang, D. Manocha, and A. S. Bedi, "Towards possibilities & impossibilities of ai-generated text detection: A survey," *arXiv preprint arXiv:2310.15264*, 2023.
- [55] M. Masood, M. Nawaz, K. M. Malik, A. Javed, A. Irtaza, and H. Malik, "Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward," *Applied intelligence*, vol. 53, no. 4, pp. 3974-4026, 2023.
- [56] V. Chamola, V. Hassija, A. R. Sulthana, D. Ghosh, D. Dhingra, and B. Sikdar, "A Review of Trustworthy and Explainable Artificial Intelligence (XAI)," *IEEE Access*, vol. 11, pp. 78994-79015, 2023, doi: 10.1109/ACCESS.2023.3294569.
- [57] J. A. D. Guzman, K. Thilakarathna, and A. Seneviratne, "Security and Privacy Approaches in Mixed Reality: A Literature Survey," *ACM Comput. Surv.*, vol. 52, no. 6, p. Article 110, 2019, doi: 10.1145/3359626.
- [58] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 7464-7475.
- [59] E. O. Appiah and S. Mensah, "Object detection in adverse weather condition for autonomous vehicles," *Multimedia Tools and Applications*, vol. 83, no. 9, pp. 28235-28261, 2024.
- [60] J. Van Brummelen, M. O'brien, D. Gruyer, and H. Najjaran, "Autonomous vehicle perception: The technology of today and tomorrow," *Transportation research part C: emerging technologies*, vol. 89, pp. 384-406, 2018.
- [61] X. Zhao, Y. Fang, H. Min, X. Wu, W. Wang, and R. Teixeira, "Potential sources of sensor data anomalies for autonomous vehicles: An overview from road vehicle safety perspective," *Expert Sys Appl*, p. 121358, 2023.
- [62] H. Wu *et al.*, "Towards explainable in-the-wild video quality assessment: a database and a language-prompted approach," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 1045-1054.
- [63] L. Li *et al.*, "Theme-aware visual attribute reasoning for image aesthetics assessment," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 9, pp. 4798-4811, 2023.
- [64] P. Meer, D. Mintz, A. Rosenfeld, and D. Y. Kim, "Robust regression methods for computer vision: A review," *International journal of computer vision*, vol. 6, pp. 59-70, 1991.
- [65] A. Deviyani, "Assessing dataset bias in computer vision," *arXiv preprint arXiv:2205.01811*, 2022.
- [66] N. Zhuang *et al.*, "Recognition oriented facial image quality assessment via deep

- convolutional neural network," *Neurocomputing*, vol. 358, pp. 109-118, 2019.
- [67] G. Wang *et al.*, "Two-stage unsupervised facial image quality measurement," *Information Sciences*, vol. 611, pp. 432-445, 2022.
- [68] T. Yang, Y. Zhang, J. Sun, and X. Wang, "Privacy enhanced cloud-based facial recognition," *Neural Processing Letters*, pp. 1-9, 2022.
- [69] B. Meden *et al.*, "Privacy-enhancing face biometrics: A comprehensive survey," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 4147-4183, 2021.
- [70] N. Japkowicz and S. Stephen, "The class imbalance problem: A systematic study," *Intelligent Data Analysis*, vol. 6, pp. 429-449, 2002, doi: 10.3233/IDA-2002-6504.
- [71] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, "Class-balanced loss based on effective number of samples," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 9268-9277.
- [72] J. Benaloh, M. Chase, E. Horvitz, and K. Lauter, "Patient controlled encryption: ensuring privacy of electronic medical records," in *Proceedings of the 2009 ACM workshop on Cloud computing security*, 2009, pp. 103-114.
- [73] I. Keshta and A. Odeh, "Security and privacy of electronic health records: Concerns and challenges," *Egyptian Informatics Journal*, vol. 22, no. 2, pp. 177-183, 2021.
- [74] N. DeMarinis, S. Tellex, V. P. Kemerlis, G. Konidaris, and R. Fonseca, "Scanning the Internet for ROS: A View of Security in Robotics Research," in *2019 International Conference on Robotics and Automation (ICRA)*, 20-24 May 2019 2019, pp. 8514-8521, doi: 10.1109/ICRA.2019.8794451.
- [75] A. Botta, S. Rotbei, S. Zinno, and G. Ventre, "Cyber security of robots: A comprehensive survey," *Intelligent Systems with Applications*, vol. 18, p. 200237, 2023.
- [76] L. Gualtieri, E. Rauch, and R. Vidoni, "Emerging research fields in safety and ergonomics in industrial collaborative robotics: A systematic literature review," *Robotics and Computer-Integrated Manufacturing*, vol. 67, p. 101998, 2021/02/01/ 2021, doi: https://doi.org/10.1016/j.rcim.2020.101998.
- [77] T. A. Puranik, N. Shaik, R. Vankudoth, M. R. Kolhe, N. Yadav, and S. Boopathi, "Study on Harmonizing Human-Robot (Drone) Collaboration: Navigating Seamless Interactions in Collaborative Environments," in *Cybersecurity Issues and Challenges in the Drone Industry*: IGI Global, 2024, pp. 1-26.
- [78] A. Malik, M. Kuribayashi, S. M. Abdullahi, and A. N. Khan, "DeepFake detection for human face images and videos: A survey," *IEEE Access*, vol. 10, pp. 18757-18775, 2022.
- [79] N. T. Chibi, H. E. Ghazi, and W. F. Fihri, "Drone Cyber-Attack: An Intrusion Detection Technique Based on RSSI and Trilateration," in *2021 Third International Conference on Transportation and Smart Technologies (TST)*, 27-28 May 2021 2021, pp. 42-45, doi: 10.1109/TST52996.2021.00014.
- [80] M. A. Khan, H. Menouar, A. Eldeeb, A. Abu-Dayya, and F. D. Salim, "On the detection of unauthorized drones—Techniques and future perspectives: A review," *IEEE Sens. J.*, vol. 22, no. 12, pp. 11439-11455, 2022.
- [81] S. E. Kahou *et al.*, "Emonets: Multimodal deep learning approaches for emotion recognition in video," *Journal on Multimodal User Interfaces*, vol. 10, pp. 99-111, 2016.
- [82] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 2, pp. 423-443, 2018.
- [83] M. Mansoury, H. Abdollahpouri, M. Pechenizkiy, B. Mobasher, and R. Burke, "Feedback loop and bias amplification in recommender systems," in *Proceedings of the 29th ACM international conference on information & knowledge management*, 2020, pp. 2145-2148.
- [84] M. J. Lee, Z. Jin, S. N. Liang, and M. Tistarelli, "Alignment-Robust Cancelable Biometric Scheme for Iris Verification," *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 3449-3464, 2022, doi: 10.1109/TIFS.2022.3208812.
- [85] S. M. Abdullahi, S. Sun, B. Wang, N. Wei, and H. Wang, "Biometric template attacks and recent protection mechanisms: A survey," *Information Fusion*, vol. 103, p. 102144, 2024.
- [86] R. Vilalta, C. V. Apte, J. L. Hellerstein, S. Ma, and S. M. Weiss, "Predictive algorithms in the management of computer systems,"

- IBM Systems Journal*, vol. 41, no. 3, pp. 461-474, 2002.
- [87] F. Salfner, M. Lenk, and M. Malek, "A survey of online failure prediction methods," *ACM Computing Surveys (CSUR)*, vol. 42, no. 3, pp. 1-42, 2010.
- [88] M. Lindholm, R. Richman, A. Tsanakas, and M. V. Wüthrich, "Discrimination-free insurance pricing," *ASTIN Bulletin: The Journal of the IAA*, vol. 52, no. 1, pp. 55-89, 2022.
- [89] X. Xin and F. Huang, "Antidiscrimination insurance pricing: Regulations, fairness criteria, and models," *N. Am. Actuar. J.*, vol. 28, no. 2, pp. 285-319, 2024.
- [90] M. Mozafari, R. Farahbakhsh, and N. Crespi, "Hate speech detection and racial bias mitigation in social media based on BERT model," *PLoS One*, vol. 15, no. 8, p. e0237861, 2020.
- [91] A. K. Sahoo, S. Mallik, C. Pradhan, B. S. P. Mishra, R. K. Barik, and H. Das, "Intelligence-based health recommendation system using big data analytics," in *Big data analytics for intelligent healthcare management*: Elsevier, 2019, pp. 227-246.
- [92] A. K. Sahoo, C. Pradhan, and H. Das, "Performance evaluation of different machine learning methods and deep-learning based convolutional neural network for health decision making," *Nature inspired computing for data science*, pp. 201-212, 2020.
- [93] F. Kamiran and T. Calders, "Classification with no discrimination by preferential sampling," in *Proc. 19th Machine Learning Conf. Belgium and The Netherlands*, 2010, vol. 1, no. 6: Citeseer.
- [94] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM computing surveys (CSUR)*, vol. 54, no. 6, pp. 1-35, 2021.
- [95] B. Zhang, H. Yan, J. Wu, and P. Qu, "Application of Semantic Analysis Technology in Natural Language Processing," *Journal of Computer Technology and Applied Mathematics*, vol. 1, no. 2, pp. 27-34, 2024.
- [96] H. Li, F. Xu, and Z. Lin, "ET-DM: Text to image via diffusion model with efficient Transformer," *Displays*, vol. 80, p. 102568, 2023.
- [97] S. Qamar, Z. Anwar, and M. Afzal, "A systematic threat analysis and defense strategies for the metaverse and extended reality systems," *Computers & Security*, vol. 128, p. 103127, 2023.

Appendix 1: The countermeasure for the selected application and risk

Application	Risk	Countermeasure	Keyword (full text)	Reference
ALPR/ANPR	Inaccuracy	Robust local feature points	ANPR & accuracy	[27, 28]
	Unreliability	Integrating data from an alternate source	ANPR	[28, 29]
	Privacy	Store data in shortest amount of time	traffic & duration of storage	[30, 31]
Behavioural analysis/ monitoring	Bias	Conduct regular fairness audit	bias & monitoring	[32, 33]
Blood measurement algorithm	Inaccuracy	Anti-Motion Interference Wearable Device	blood oxygen & accuracy	[34, 35]
	Unreliability	Apply digital filtering method	blood oxygen & reliability	[34, 36]
	Appropriateness	Conduct requirements engineering	requirements & artificial intelligence	[37, 38]
Bot/ Agent	Safety	Implement guardrails	chatbot & safety	[39, 40]
CCTV/ Surveillance System	Inaccuracy	Employ feature importance	Surveillance & Accuracy	[27, 28]
	Unreliability	Use MPEG4 over wavelet	CCTV & reliability	[41, 42]
	Bias	Domain-conditional model with explicit combination of per-domain class predictions	visual & bias	[43, 44]
	Security	Implement access control	CCTV & security	[45, 46]
	Privacy	Use multiparty key-sharing scheme for access control	surveillance system & privacy	[46, 47]
Chatbot	Inaccuracy	Correction with external evidence	chatbot & accuracy	[39, 48]
	Unreliability	Multi-agent interaction	chatbot & reliability	[39, 49]
	Requirement gap	Conduct requirements engineering	requirements & artificial intelligence	[37, 38]
	Bias	replacing biased texts in the data set with neutral texts	chatbot & bias	[39, 50]
	Plagiarism	Use the cross-language plagiarism detection mechanism	plagiarism & detection	[51, 52]
	Copyright infringement	Watermarking	copyright & generative AI	[53, 54]
	Disinformation/ misinformation	Reinforcement Learning from Human Feedback	chatbot & misinformation	[39, 55]
	Privacy	Use differential privacy	privacy & artificial intelligence	[56, 57]
	Safety	Implement guardrails	chatbot & safety	[39, 40]
Self-driving system/ Driver assistance system	Inaccuracy	Use YOLOv7 architecture for realtime object detection	autonomous driving & performance	[58, 59]
	Unreliability	Detection mechanism for sensor fault	autonomous driving	[60, 61]
	Safety	Use the YOLOv7 architecture for real-time object detection	autonomous driving	[58, 59]
Computer Vision	Inaccuracy	Combine technical and aesthetic scores	computer vision & accuracy	[62, 63]

Application	Risk	Countermeasure	Keyword (full text)	Reference
	Unreliability	Semantically construct aesthetic features for Image Aesthetics Assessment (IAA)	computer vision & reliability	[64, 65]
	Bias	Semantically construct aesthetic features for image aesthetics assessment (IAA)	computer vision & bias	[64, 65]
	Requirement gap	Perform requirements engineering	requirements & artificial intelligence	[37, 38]
Facial recognition	Inaccuracy	Use facial quality assessment	facial recognition & accuracy	[66, 67]
	Privacy	Use homomorphic encryption	facial recognition & privacy	[68, 69]
Gaze redirection system	Requirement gap	Perform requirements engineering	requirements & artificial intelligence	[37, 38]
Dataset	Bias	Oversample minority class	class imbalance	[70, 71]
	Privacy	Use encryption for medical records	privacy & medical records	[72, 73]
Robotics	Privacy	Detecting exposure by conducting network scans	robotics & privacy	[74, 75]
	Security	Activate security features in the transport and application layers	robotics & security	[74, 75]
	Safety	Incorporate human-in-the-loop	robotics & safety	[76, 77]
Deepfake-audio	Copyright infringement	Watermarking	copyright & generative AI	[53, 54]
Deepfake-video	Disinformation/misinformation	Use a hybrid of signature- and anomaly-based detection mechanism	deepfakes & detection	[55, 78]
Drone	Unauthorized attack	Use the Received Signal Strength Indication (RSSI) technique and the trilateration method	drone & unauthorized	[79, 80]
Emotion recognition	Inaccuracy	Use multimodal deep learning approach	video & identification	[81, 82]
	Bias	Use Reducing Bias Amplification (RBA)	visual & bias	[43, 83]
Iris scanning	Privacy	Apply iris template protection	iris & privacy	[84, 85]
Prediction Algorithm/ Predictive Statistical Analysis	Inaccuracy	Use high-resolution vital signs time series and electronic medical record to train the model	prediction & model	[86, 87]
Pricing Algorithm	Bias	Use discrimination-free pricing formula	pricing & discrimination	[88, 89]
Recommendation algorithm	Bias	Use a reweighting mechanism as bias-mitigating module	bias & detection system	[44, 90]
	Safety	Implement measurement criteria	recommendation & safety	[91, 92]
Text analysis	Privacy	Use encryption for medical records	privacy & medical record	[72, 73]
	Bias	Use preferential sampling	discrimination & models	[93, 94]

Application	Risk	Countermeasure	Keyword (full text)	Reference
Text-to-image/ Image generator	Copyright infringement	Watermarking	copyright & generative AI	[53, 54]
Text-to-image	Bias	Combine the diffusion model and the high-efficiency transformer model for text-to- image synthesis	visual & bias	[95, 96]
Text-to- speech/ Voice generator	Copyright infringement	Watermarking	copyright & generative AI	[53, 54]
Virtual reality	Safety	Use authentication protocol for device-to-device sharing	virtual & safety	[57, 97]
Voice recognition	Copyright infringement	Watermarking	copyright & generative AI	[53, 54]