IJECBE (2025), **3**, **2**, 377–410
Received (12 March 2025) / Revised (13 March 2025)
Accepted (8 June 2025) / Published (30 June 2025)
https://doi.org/10.62146/ijecbe.v3i1.109
https://ijecbe.ui.ac.id
ISSN 3026-5258

International Journal of Electrical, Computer and Biomedical Engineering

RESEARCH ARTICLE

Artificial Intelligence Risk Identification: Challenges, Impacts, and Mitigation Strategies

Ulfia Syukrina and I Gde Dharma Nugraha*

Department of Electrical Engineering, Faculty of Engineering, Universitas Indonesia, Depok, Indonesia
*Corresponding author. Email: i.gde@ui.ac.id

Abstract

Artificial Intelligence (AI) has rapidly transformed various industries, providing significant benefits in automation, decision-making, and efficiency. However, AI also presents numerous risks, including bias, lack of transparency, security vulnerabilities, and regulatory challenges. This study employs a Systematic Literature Review (SLR) approach to identify and categorize key risks associated with AI implementation. The findings indicate that AI risks can be classified into technological, social, and regulatory aspects, each posing unique challenges. Algorithmic bias, privacy concerns, and the lack of global AI governance frameworks highlight the need for more robust risk mitigation strategies. To address these challenges, this study recommends enhancing fairness-aware AI models, strengthening AI governance, and increasing public AI literacy. Future research should focus on improving AI accountability, security measures, and ethical guidelines to ensure responsible AI adoption.

Keywords: Artificial Intelligence, AI Risks, Bias in AI, AI Governance, AI Security, Ethical AI, Transparency, Accountability in AI, Systematic Literature Review

1. Introduction

Artificial Intelligence Systems were officially defined for the first time in 1956 as "the science and engineering of making intelligent machines" by John McCarthy [1]. In a broader sense, the IEEE Guider defines AI as a system that exhibits intelligent behavior by analyzing its environment and taking actions with varying degrees of autonomy to achieve specific goals [2]. AI has experienced rapid growth over the past decade. It has penetrated and spread across various disciplines and aspects of society. AI is increasingly taking over human tasks by replacing decision-making processes that were previously carried out by humans [3].

The rapid advancement of AI technology is driven by developments in machine learning, deep learning, and natural language processing. These advancements have led to the emergence of powerful AI models and algorithms that surpass human performance in various tasks, such as language translation, image recognition, and playing complex games [4] [5]. As a result, AI has become a transformative force across multiple sectors of human life, including healthcare [6], transportation [7], accounting [8], education [9], and entertainment [10].

Although AI technology offers many tangible benefits, its emergence also presents significant risks and ethical concerns for users, developers, and society [11]. This is because AI systems make decisions and predictions that impact human lives [12][13]. The AI Incident Database (AIID), an online database for reporting AI-related risks, incidents, and controversies, has documented at least 2,585 reported incidents since 2018. These risks vary widely, such as an

AI-powered robotic arm gripping its opponent's index finger so tightly that it resulted in a fracture [14]. Additionally, Google's AI-powered search engine has been reported to frequently provide misleading yet convincing and potentially harmful information [15].

There are many other examples related to failures, fairness, bias, privacy, and ethical issues arising from AI systems [16]. Even more concerning, AI technology is now being exploited by irresponsible individuals to harm others or even society at large. One such case involved criminals using AI-based software to mimic the voice of a chief executive. With a highly convincing fake voice, they successfully deceived a company into transferring $243,000 [17]. This case illustrates how AI can be misused, creating real threats that require serious attention from all stakeholders.

The adoption of AI technology across various sectors of life is also accompanied by risks that can influence public acceptance and trust. Although extensive research on AI risks has been conducted, there is still a need for a systematic classification of these risks to understand their relationships, their impact on different sectors, and the underlying causes of AI-related risks. Addressing this gap will provide a more comprehensive perspective on how risks emerge and interact, serving as a foundation for future research and interventions. This study employs the Systematic Literature Review (SLR) method to explore and categorize key AI risks, providing valuable insights for researchers and practitioners in addressing the challenges and implications of AI implementation across various contexts.

The aims of this study is to identify and classify the key risks associated with AI using a Systematic Literature Review (SLR) approach. The primary focus of this research is to analyze different types of AI risks, such as bias, transparency, privacy, security, and accountability, as well as to explore the impact of these risks on various sectors that have extensively adopted AI, including medicine, education, transportation, government, economy, and finance, as identified in the reviewed literature. Additionally, this study examines and analyzes the underlying factors contributing to the emergence of risks in AI implementation, providing a deeper understanding of the relationships between risks, their impacts, and root causes.

The findings of this research are expected to offer comprehensive insights for researchers, policymakers, and practitioners in designing more effective mitigation strategies to address the social and technical challenges of AI implementation, ensuring that AI's capabilities can be maximized for positive outcomes.

2. Literature Review

2.1 The Benefits of AI Across Various Sectors

The broad and diverse applications of AI have led to increased efficiency and cost reduction, benefiting economic growth, social development, and human well-being [18]. For example, AI-powered chatbots can respond to client and customer inquiries at any time, enhancing customer satisfaction and boosting company sales [19]. In healthcare, telemedicine is an AI technology that allows doctors to serve patients in remote locations. This technology reduces the burden on patients by enabling online consultations, allowing doctors to assist multiple practices and treat patients simultaneously. Telemedicine also improves the quality of healthcare facilities by facilitating the exchange of medical information across distant regions. Previously underserved areas and individuals with limited mobility can now receive medical opinions and prescriptions more quickly [6].

In the transportation sector, AI is utilized in collision avoidance systems, which help prevent accidents. AI enhances driving safety by predicting driver behavior and vehicle trajectories. This technology enables systems to adapt to dynamic environmental conditions rather than being passive. AI also improves data and information processing efficiency, accelerating decision-making. By predicting and managing vehicle interactions on the road, AI enhances driving comfort [7].

In the field of accounting, AI is used to improve the accuracy of financial analysis, producing more effective risk predictions, such as bankruptcy and fraud detection. AI aids in extracting information from complex and large-scale data, which is often difficult for traditional models to process. AI-powered technologies automate numerous simple and repetitive processes, such as Robotic Process Automation (RPA). By enhancing anomaly detection and transaction pattern analysis, AI can identify suspicious activities that may indicate illegal behavior. Additionally, AI supports more objective and data-driven decision-making [8].

AI provides significant benefits in the education sector, from automating administrative tasks to improving teaching quality through technologies such as Intelligent Tutoring Systems (ITS). By utilizing simulations, virtual reality, and AI-based adaptive learning systems, students can experience a richer learning environment. AI-powered learning platforms allow students from around the world to access educational materials without geographical limitations. AI is also used to identify learning patterns, predict student performance, recommend more effective learning methods, and assist students in selecting career paths based on their abilities and preferences [9].

In the entertainment and gaming industry, AI is used to create game opponents that mimic human players. AI also enables tournament organizers to host AI vs. AI or AI vs. human matches. The advancement of AI in entertainment has driven innovation in game design, including the use of AI to control NPCs (Non-Player Characters) with dynamic behavior, providing a more realistic challenge.

Additionally, AI is employed to test game balance and evaluate player strategies, ensuring a well-structured and engaging gaming experience [10].

2.2 Risks of Artificial Intelligence

The predictive, classification, association, and optimization capabilities of AI are crucial for businesses and government functions, enhancing efficiency and decision-making quality [20]. AI heavily relies on the data used for its training. If the input data is inaccurate, biased, or intentionally manipulated, the AI's output may be flawed and unreliable. The complexity of AI systems and their dependence on training data can sometimes make it difficult for AI to interpret user-intended outputs. This raises challenges related to transparency, accountability, and responsibility in AI systems [21]. In some cases, these challenges can lead to serious consequences or harmful outcomes, depending on how the technology is implemented [22]. The design or choice of training data is often cited as a major source of AI's negative effects, including discriminatory decision-making [23].

Bias in AI is one of the most widely discussed and publicly scrutinized issues. Like humans, AI can produce biased outputs, such as gender or racial discrimination. Bias can arise from various sources, including the training data used, the values of developers or users, and the AI's learning process itself. Bias not only leads to technical errors but also raises significant ethical concerns [24].

AI also faces major challenges in maintaining data privacy and security. Cybercriminals now have more ways to access personal data at low cost while generating substantial profits. In recent years, data security breaches have become increasingly frequent, making privacy protection a pressing issue in AI adoption to ensure that AI technologies do not compromise users' privacy rights [25].

Transparency, or explainability, is another well-known weakness of AI. Machine learning (ML), particularly deep neural networks, is the core technology behind modern AI. However, the inference process in ML is often difficult to explain and understand, commonly referred to as a black-box problem. The lack of transparency in ML makes its algorithms and models obscure, making them difficult to comprehend for both developers and users. This lack of transparency also hinders oversight and human guidance in AI and ML applications, increasing operational risks [11].

"AI risk," according to experts, is a phenomenon where there is a gap between the intended objectives of AI development and the actual outcomes achieved [26]. The risks associated with AI primarily emerge when its functions fail to align with societal norms and expectations [21].

2.3 The Importance of AI Risk Identification

Identifying risks in the development and deployment of AI is essential to maximize its benefits while mitigating or preventing potential risks. Thiebes et al [27], emphasize that trustworthiness is the foundation of society, the economy, and sustainable development. Without trust in the development, implementation, and use of AI, individuals, organizations, and society as a whole will be unable to fully realize AI's potential. To build this trust, it is crucial to have an explicit understanding of the ethical, social, and technological risks that accompany AI adoption.

Laux et al [23]. highlight that the concept of "acceptability of risks" serves as a foundation for defining trust in AI, particularly within the European Union's regulatory framework. Risks that are deemed unacceptable or intolerable must be identified to ensure AI implementation aligns with EU values and existing laws. Involving developers in risk management is necessary to ensure that these risks are addressed before AI is introduced to the market or used in public institutions. Hickok [28], argues that AI risk identification is a critical step, especially in government procurement of AI technologies. This process is essential to prevent severe consequences. AI technologies are often deployed without sufficient transparency and oversight, raising concerns about accountability, human rights violations, and the erosion of fairness in society. Additionally, AI operates on a large scale and with high technological complexity, meaning even small errors can have significant consequences. Therefore, early risk identification is vital to avoid unintended negative impacts.

Other challenges related to bias, transparency, privacy, and accountability must also be taken into consideration [12][13]. The rapid advancement of AI creates new opportunities, but it also introduces unforeseen risks. The complexity of AI technology demands special attention, particularly regarding its impact on privacy, security, and public trust. Furthermore, AI's social implications cannot be overlooked, as the technology influences various segments of society [20]. Risk identification in AI development and deployment is a crucial step to ensure that this technology is utilized optimally. It is not only essential for preventing unintended negative consequences but also serves as a key factor in fostering long-term trust in AI.

3. Research Methodology

A Systematic Literature Review (SLR) is a research methodology used to systematically collect, identify, and critically analyze various existing studies, including articles, conference proceedings, books, and dissertations, by following a structured procedure [29]. SLR provides readers with up-to-date information on recent literature related to a specific subject. Its primary goal is to review key aspects of current knowledge on a particular topic, addressing specific research questions and identifying areas that require further investigation [30].

The SLR method used in this study follows the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) approach due to its comprehensive and structured methodology. PRISMA provides a detailed checklist covering all essential parameters, ensuring that the review can be replicated by other researchers while generating accurate data to support future studies [31]. Initially designed for the medical and physical sciences, PRISMA is now being adapted for computer science research, although its application in this field is still being explored.

The PRISMA approach not only ensures the quality of the review but also enables readers to evaluate its strengths and weaknesses while offering a structured and replicable framework for future researchers. By adopting PRISMA, this study aims to provide a valuable reference for researchers within this field [32].

3.1 Research Question

The Research Question (RQ) is a crucial component and serves as the primary step to ensure that the research review aligns with the study's objectives. Since this research aims to categorize and provide an overview of the various types of risks associated with AI implementation, previous studies will be highly beneficial in understanding the extent to which AI risks have been classified. This study seeks to answer the research questions outlined in *Table 1: Research Questions and Their Aims.*

Table 1. Research Questions and the aims

ID	Research Question	Aims
RQ1	What are the types of risks associated with implementation AI according to existing literature?	To ensure the research focuses on identifying different types of risks
RQ2	What are the impacts of AI implementation risks across various sectors of life	To examine the relationship between risks and their impacts on various sectors of life
RQ3	What are the causes of AI-related risks?	To explore the root causes of risks arising from AI

Table 2. Inclusion and Exclusion Criteria

Inclusion Criteria	Exclusion Criteria
Studies published between 2020–2024	Studies published before 2020
Research articles in journal format	Research that is not in journal format
Studies written in English	Studies not written in English
Open-access research	Non-open-access research
Studies that do not focus only on technical aspects	Studies with a strong technical focus
Studies that are not limited to a single sector	Studies that concentrate on a single sector

The objective of the literature search strategy is to identify studies relevant to the research questions. The literature search sources used for the database are SCOPUS and IEEE. The keywords used for literature search in this study are:

(("Abstract": "artificial intelligence" OR "ai") AND ("Document Title": "artificial intelligence" OR "ai") AND ("risk" OR "challenge*" OR "concern*" OR "ethic*")) AND("trust*" OR "privacy" OR "bias" OR "security" OR "transpara*" OR "safe*" OR "robust*" OR "fair*" OR "account*"))*

3.2 Literature Search Strategy

After gathering potential studies with high relevance, the next step is to assess their actual relevance [33]. The inclusion and exclusion criteria used to filter the selected primary studies can be found in *Table 2. Inclusion and Exclusion Criteria*

The PRISMA flow diagram consists of three stages: identification, screening, and inclusion. The details of each stage are presented in Figure 1. PRISMA Flow Diagram

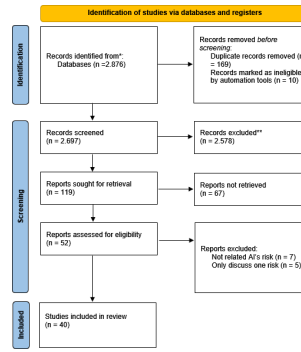

Figure 1. PRISMA Flow Diagram

a) *Identification:* Based on the PRISMA flow, a total of 2,876 studies were identified from two databases using the specified search keywords. After removing 11 inaccessible studies, 2,866 studies remained. A duplication check was then performed, identifying 169 duplicate studies, leaving 2,697 studies for further examination.
b) *Screening:* A total of 2,697 studies were screened based on their titles and abstracts. As a result, 2,600 studies were excluded for not meeting the inclusion and exclusion criteria, leaving 119 studies.
c) *Inclusion:* The remaining 119 studies underwent a comprehensive review to determine their relevance in answering the research questions. After a more in- depth analysis, 40 studies were selected. This study utilized Zotero to store, manage, and organize references.

4. Findings

This section presents the results of the extraction and analysis of the collected literature to answer the research questions. *Table 3. Summary of AI Risks Based on Literature Review* summarizes the key findings from the reviewed studies by categorizing the identified risks into several main aspects. The structure of the table consists of four main columns:

a) Types of AI Risks – The primary categories of risks that emerge in AI implementation based on the literature review.
b) Impact of AI Risks – The consequences of each type of risk, affecting individuals, groups, or entire systems.
c) Affected Sectors – The industries or fields most vulnerable to the negative effects of AI adoption, such as healthcare, finance, education, and others.
d) Causes of AI Risks – The key factors that trigger or contribute to each identified risk. These causes are further classified into three categories:

1. Technical Causes – Factors arising from technological limitations, such as algorithmic bias, lack of model transparency, or vulnerabilities to cyberattacks.
2. Social Causes – Factors stemming from social, cultural, or economic aspects, such as low AI literacy, unequal access to technology, or AI's influence on social dynamics.

3. Regulatory Causes – Factors related to weaknesses or the absence of adequate regulations, such as a lack of AI security standards, weak oversight mechanisms, or legal gaps in data protection and privacy.

This table aims to provide a comprehensive overview of how AI risks arise, their impacts on various sectors, and the key factors contributing to these risks.

Table 3. Summary of AI Risks Based on Literature Review

Journal	Types of AI Risks	Impact of AI Risks	Affected Sectors	Causes of AI Risks		
				1	2	3
[34]	Adversarial attacks, privacy, model security, data connectivity, techno-colonialism	Incorrect predictions, user data breaches, operational failures due to system errors, loss of trust in AI	Healthcare, transportation, cybersecurity	✓	✓	✓
[35]	Bias, privacy, security, transparency	Discrimination due to bias, data collection violating privacy, social inequality	Transportation, social media, healthcare	✓	✓	✓
[36]	Bias, privacy, transparency, explainability, adversarial manipulation,accountability	Injustice due to discrimination, conflicts between transparency and data privacy, medical misdiagnoses, social disparities, lack of trust in AI	Healthcare, law, recruitment, transportation	✓	✓	✓
[37]	Bias, transparency, techno-colonialism, social injustice, inequality in technology access	Discrimination against minority groups, increasing social and economic inequality, lack of public trust, growing global disparity	Healthcare, labor, education, public administration	✓	✓	✓
[28]	Bias, transparency, privacy, data security, accountability, social inequality	Discrimination against certain groups in law and education, misuse of personal data, widening inequality, economic and social losses	Law enforcement, healthcare, education, public administration	✓	✓	✓
[38]	Bias, privacy, transparency, security, accountability	Unfair decision-making leading to discrimination, user data exploitation, social inequality, loss of trust in AI	Healthcare, education, finance, transportation, social media	✓	✓	✓
[39]	Bias, privacy, transparency, techno- colonialism, freedom of expression	Discrimination creating social injustice, privacy violations, lack of public trust, economic and emotional losses	Libraries, government, digital platforms	✓	✓	✓
[40]	Bias, privacy, security, transparency, accountability	Discrimination, privacy violations, social inequality, loss of trust	Healthcare, media, transportation, entertainment, education	✓	✓	✓
[41]	Bias, privacy, transparency, cybersecurity threats	Social discrimination, privacy breaches, unemployment due to job automation, global economic inequality, military misuse increasing geopolitical tensions	Healthcare, transportation, education, business, and manufacturing	✓	✓	✓
[42]	Bias, poor data quality, data privacy, techno- solutionism, surveillance humanitarianism, techno- colonialism	Gender or racial discrimination, privacy violations increasing vulnerability of certain groups, social inequality, incorrect decision-making	Healthcare, natural disasters, migration (Project Jetson)	✓	✓	✓
[43]	Bias, privacy, security, transparency, accountability	Discrimination against specific groups, privacy violations due to unauthorized data collection, social inequality leading to disparity	Healthcare, transportation, electronics, media, and entertainment	✓	✓	✓

[44]	Bias, privacy, security, transparency, social injustice	Discrimination worsening social justice, data exploitation, declining public trust due to lack of transparency, socio-political conflicts	Healthcare, social services, transportation, judiciary	✓	✓	✓
[45]	Bias, transparency, privacy, reliance on representative data, accountability	Discrimination and social inequality, privacy violations, unfairness in automated decision-making	Healthcare, human resources, legal system, public administration, law enforcement, education	✓	✓	✓
[46]	Bias, privacy, security, transparency, accessibility, accuracy	Discrimination against vulnerable groups, privacy violations, loss of public trust	Healthcare, finance, transportation	✓	✓	✓
[47]	Bias, privacy, security, transparency, social injustice, accountability	Discrimination against certain groups, misuse of user data, physical or financial harm due to technical operational failures, social inequality	Healthcare, transportation, education, social media	✓	✓	✓
[48]	Bias, privacy, transparency, explainability, accountability	Social discrimination, misuse of sensitive data, economic disparity due to social inequality, loss of public trust	Healthcare, finance, entertainment, transportation, manufacturing, human resources	✓	✓	✓
[49]	Bias, privacy, security, transparency, accountability	Discrimination increasing social disparity, unauthorized user data exploitation, loss of public trust	Healthcare, finance, transportation, public safety, and law	✓	✓	✓
[50]	Bias, privacy, manipulation, security, transparency	Violations of fundamental rights to privacy and individual autonomy, social discrimination, declining trust in technology, barriers to technology adoption due to unaddressed risks	National security, consumer sector, government, technology, and innovation	✓	✓	✓
[51]	Bias, transparency and interpretability, privacy and data security	Discrimination reinforcing social inequality, public distrust due to lack of education and awareness, unauthorized use of personal data, environmental damage from AI's high energy consumption	Healthcare, law enforcement, industry, education, environmental	✓	✓	✓
[52]	Bias, privacy, security, transparency, social inequality	Discrimination leading to injustice, misuse of personal data, unequal distribution of AI benefits causing disparities, political destabilization due to AI development affecting global stability	Healthcare, finance, judiciary, defense industry	✓	✓	✓
[53]	Bias, security, privacy, transparency, manipulation and autonomy	Potential misuse of personal data, discrimination in recommendations, declining public trust in AI recommendations, filter bubbles reinforcing polarization by limiting exposure to new perspectives	E-commerce, social media, education, healthcare	✓	✓	✓
[54]	Bias, privacy, safety and security, transparency, social-ethical concerns	Discrimination in decision-making, exposure of sensitive data, social inequality in society	Healthcare, security, public administration	✓	✓	✓
[55]	Bias, transparency and interpretability, human control	Discrimination based on gender, race, and socio -economic status, loss of trust in AI, worsening social and economic inequality, loss of human control over automated systems, negative narratives creating unrealistic expectations or fears	Education, finance, transportation, healthcare			

[56]	Bias, privacy, security, transparency	Discrimination against certain groups, user data exploitation, harm from cyberattacks	Healthcare, government, transportation	✓	✓	✓
[57]	Privacy, transparency, bias, moral responsibility, techno- colonialism	Discrimination due to biased data, privacy violations, social inequality, dehumanization as individuals lose control over personal data	Healthcare, technology, social media, government	✓	✓	✓
[58]	Bias, privacy, security, transparency, accuracy, accountability, machine autonomy orunsupervised decision-making	Discrimination and social inequality, collection and processing of individual data without strong protection, public distrust, economic losses such as tax evasion harming governments and society	Taxation, judiciary, healthcare	✓	✓	✓
[59]	Bias, privacy, security, transparency, accountability, sustainability, human dignity, social solidarity	Discrimination and social injustice, privacy violations leading to declining trust, environmental damage from AI's energy consumption, marginalization of vulnerable groups (children, minorities, people with disabilities)	Healthcare, transportation, smart cities, military, and agriculture	✓	✓	✓
[60]	Bias, privacy, security, transparency	Discrimination harming specific groups, misuse of personal data, social inequality, public safety risks due to unreliable AI	Law enforcement, healthcare, education, workforce recruitment	✓	✓	✓
[61]	Bias, privacy, security, injustice, transparency	Discrimination against certain groups, privacy violations, physical harm, social manipulation or public opinion manipulation	Healthcare, law, transportation, social media, security	✓	✓	✓
[62]	Bias, privacy, security, transparency, socio-economic inequality, fairness	Discrimination, user data exploitation, social inequality among large tech companies, declining public trust	Government, healthcare, agriculture, cybersecurity, energy, and environment, human	✓	✓	✓
[63]	Bias, privacy, security, transparency, adversarial attacks manipulating AI input data	Discrimination, data theft and misuse, physical harm (autonomous vehicles or drones), manipulation of public opinion, economic losses from the theft of high-value datasets	Cybersecurity, media, politics, transportation, healthcare, military	✓	✓	✓
[64]	Bias, privacy, security, transparency	Discrimination due to biased data, unethical data analysis leading to privacy violations, social inequality, lack of transparency reducing trust in AI	Healthcare, transportation, social, education, media, industrial processes	✓	✓	✓
[65]	Bias, privacy, security, transparency	Discrimination as a risk from biased algorithms, privacy violations, public distrust, physical and financial harm	Healthcare, transportation, children's toys, elderly care	✓	✓	✓
[66]	Bias, privacy, transparency, security, accountability	Discrimination due to gender and racial bias, reducing public trust in AI, social inequality as technology favors certain groups, physical harm as a risk from security failures in autonomous vehicles	Healthcare, transportation, agriculture, public administration	✓	✓	✓
[67]	Bias, privacy, transparency, fairness, techno-colonialism leading to human dignity violations	Discrimination against individuals or groups, privacy violations, unfair distribution of technological benefits, declining public trust in technology	Healthcare, transportation, media and information, government			

[68]	Bias, transparency, security, privacy, accountability	Bias-driven discrimination causing social inequality, data breaches due to unauthorized usage, distrust in AI	Healthcare, finance, law, transportation	✓	✓	✓
[69]	Bias, privacy and data protection, technical security, transparency and accountability, social impact	Unfair access to AI-based services, misuse of sensitive data, unequal AI access, AI failures in critical sectors like healthcare and transportation	Healthcare, transportation, government, education, security, environmental	✓	✓	✓
[70]	Privacy and data security, bias and discrimination, transparency, accountability	Algorithmic bias reinforcing social injustice, data leaks and exposure of sensitive information, inequality in AI access and benefits, pollution and high energy consumption from AI model training	Healthcare, transportation, finance, education, industry, government, environmental	✓	✓	✓
[71]	Bias, privacy, security	Bias reinforcing racial and gender injustice, personal data used without adequate protection, low transparency creating skepticism toward technology, unexpected AI incidents with systemic impact	Transportation, public safety, economy	✓	✓	✓
[72]	Transparency, security, bias, contestability, privacy, accountability	Social discrimination, privacy violations, economic disparity due to automation, unfair access to essential services, automation reducing job opportunities and increasing inequality	Healthcare, education, transportation, law, security, and surveillance	✓	✓	✓

4.1 Types of AI Risks

This section presents the analysis results of 40 key literature sources used to answer the research question:

"RQ1: What are the types of risks associated with implementation AI according to existing literature??"

Based on the literature review conducted, various key risks in I implementation have been identified and summarized in *Table 3. Summary of AI Risks Based on Literature Review* under the risk types column. The method used to answer this question is to group the types of risks based on trends, namely by looking at how many journals discuss a risk. A summary and trend of journal discussions on a risk can be seen in Table. 4 shows the top 7 risks discussed in the journal literature.

Table 4. Top 7 trends in AI risk types based on literature

Types of AI Risks	Journals
Bias	39
Privacy	38
Transparency	37
Security	32
Accountability	16
Inequality and Social Disparities	10
Humas as Object	6

Additionally, these identified risks are visualized in *Figure 2. Type of AI Risks* to provide a clearer overview of the classification of AI risks found in this study.

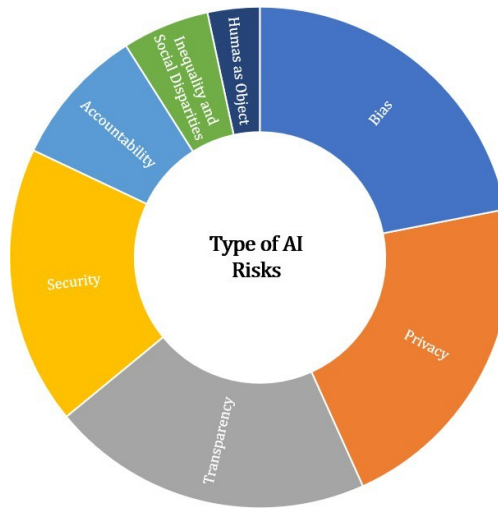

Figure 2. Types of AI Risks

4.1.1 Bias

Social injustice [55] in recommendations, predictions [53][54], and decision-making due to unrepresentative datasets [39] or algorithms [46] can disadvantage certain groups or minorities [28], such as gender, racial [66], and socio-economic status discrimination [55], potentially exacerbating societal inequalities [49]. For example, this issue arises in sectors such as healthcare, finance, education [28], job recruitment, and law [68], with a notable case in facial recognition systems [48].

4.1.2 Privacy

The large-scale and massive data collection [53] increases the risk of sensitive information leaks [54]. This occurs due to inadequate security measures in machine learning algorithms [34]. Data theft [63], unauthorized data collection [56], improper usage [41], and insufficient oversight [38] during the data analysis process [39] further exacerbate these risks. For instance, data collected from devices and used by third parties [69] is often handled without adequate protection [28], posing threats to both groups and individuals [52]. This issue is evident in cases involving biometric verification systems [28], facial recognition, and virtual assistants [61].

4.1.3 Transparency

The complexity or opacity of AI operations makes it difficult for users and regulators to understand [49], particularly in contexts such as proving violations [67]. AI also exhibits a "black- box" nature [38], making it difficult to explain [41]. The lack of transparency in its functioning [39] results in unclear algorithmic processes, system functionality [45], and the inability to explain system decisions to users [35]. AI systems developed by private vendors are not always auditable by either the government or the public [28]. This poses challenges for openness in development [40] due to the limitations in understanding AI models [43].

4.1.4 Security

Weaknesses in controlling interactions between system components are among the factors that can trigger failures or pose dangers during AI implementation [35]. These vulnerabilities can disrupt system functions and lead to incorrect decisions [34]. Adversarial attacks represent a form of AI vulnerability to input data manipulation [28], increasing the potential for cyber threats and creating new challenges in digital security [41]. Threats to training data, such as "poisoning attacks," can cause undesirable AI behavior or even pose significant risks [34][43].

Technical failures in AI, such as those in autonomous vehicles, can result in physical or material damage, as seen in the Uber incident, which exemplifies the uncertainty of AI behavior in real-world scenarios [47][69]. Vulnerability to external disruptions can also lead to malfunctioning systems, posing risks to users [49][60]. Data security remains a major challenge, including the need for secure data storage and anonymization processes to protect user privacy [53]. Threats to critical infrastructure and individual safety also encompass potential attacks on AI-based systems [68].

Various technical threats, such as white-box, black-box, poisoning, backdoor, and inference attacks, highlight vulnerabilities that must be addressed [63][64]. Unproven AI systems, such as autonomous vehicles and autonomous weapons, present significant risks, as their failures can endanger both individuals and society at large [61]. The growing reliance on AI further amplifies the risk of its misuse [62]. Systems lacking adequate security measures are vulnerable to sabotage or cyberattacks, ultimately compromising AI functionality and public trust [65][70].

4.1.5 Accountability

Determining who is responsible for decisions made by AI systems remains a complex challenge [36]. The lack of clarity in assigning responsibility for errors within AI systems further complicates this issue [28][38]. Questions about how to allocate accountability for automated decisions or actions that cause harm remain an unresolved debate [45]. The absence of clear accountability for the negative impacts of AI creates legal and ethical uncertainty [47][48]. The autonomous nature of AI systems makes it even more difficult to assign responsibility in cases of system failures [49][58]. The difficulty also extends to establishing accountability for the consequences of AI-driven decisions, both in legal and ethical contexts [71][69][68].

4.1.6 Inequality and Social Disparities

AI has the potential to exacerbate social inequality through unfair automated decision-making, which can negatively impact society at large [28][47]. Job displacement due to AI-driven automation may lead to massive job losses, creating economic instability and widening economic disparities [41][71]. This inequality is also evident in the unequal distribution of access to AI technology, further reinforcing social disparities [48].

In the context of social media, AI algorithms can deepen social polarization by creating "echo chambers" and "filter bubbles," which limit users' exposure to diverse perspectives, thereby hindering healthy and inclusive discussions [38][53]. The spread of false information on social media platforms through AI algorithms is another significant threat, as it can reinforce biases and misinformation [64]. AI-related risks in content moderation can impact freedom of speech, creating a dilemma between preserving free expression and moderating harmful content [39].

The use of AI in authoritarian states raises further concerns, such as mass surveillance that infringes on privacy and individual freedoms, as well as the potential for human rights violations [41][54]. In some cases, AI is also employed to control populations or detect welfare fraud, which may lead to political or social conflicts [44].

4.1.7 Humans as Objects

Several studies also discuss other risks posed by AI. One of these is the risk of AI reducing humans to mere "objects," thereby violating the principle of human dignity [67].

Techno-solutionism is described as an excessive reliance on technology to solve human problems. Surveillance humanitarianism refers to the uncontrolled collection of data, which increases individual vulnerability. Techno-colonialism explains how digital technology can reinforce colonial inequalities between different populations [42]. Manipulation and autonomy highlight how AI-driven systems influence user behavior for the benefit of platforms [53].

4.2 Affected Sectors

This section presents a summary of the answers to the research question:

"RQ2. What are the impacts of AI implementation risks across various sectors of life?"

Based on the literature analysis, various impacts of AI risks on different sectors have been identified and summarized. *Figure 3. The impact of AI on various sectors according to literature* provides a visualization of how AI risks affect multiple sectors of life, according to findings from the reviewed literature. The following is a detailed overview of the impacts of AI risks across various sectors as discussed in previous studies.

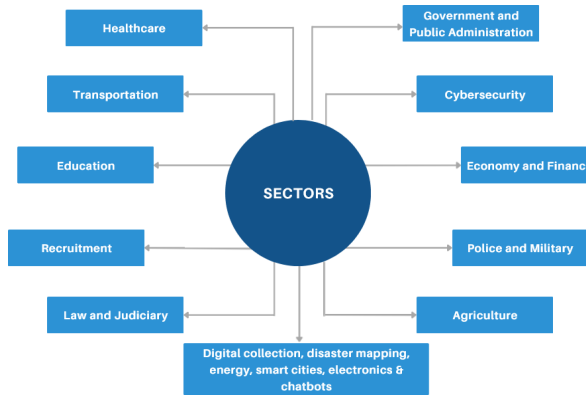

Figure 3. The impact of AI on various sectors according to literature

4.2.1 Healthcare

AI systems in the healthcare sector offer significant potential, including applications in disease diagnosis, medical image analysis, and drug discovery [48]. However, these advancements come with inherent risks, such as algorithmic bias that may lead to unfair or inaccurate decisions. For example, algorithms trained on non-representative data often fail to recognize symptoms in certain groups, such as patients from minority communities [52][55]. Racial bias in healthcare management algorithms has also been observed, where Black patients receive lower priority for medical care despite having equal or greater medical needs compared to white patients [55].

Patient privacy is another critical issue. AI-based systems frequently collect vast amounts of medical data, which is vulnerable to misuse or data breaches [63][47]. These risks become even more complex in telemedicine services, where patient data may be used without explicit consent or exploited by third parties [71]. Similar concerns arise in large-scale data collection projects, such as the NHS AI Lab in the UK, raising questions about accountability, benefit distribution, and privacy protection [54][57].

Technical failures or biases in algorithms can also have severe consequences. For instance, IBM Watson for Oncology was deployed before AI technology had fully matured, highlighting the risks of implementing untested systems [51]. Similar risks are evident in AI- driven medical diagnostic tools, where misdiagnoses due to biased training data or lack of algorithmic transparency can endanger patient health and public trust [45][58][64]

The use of humanoid robots in elderly care presents ethical dilemmas, such as emotional dependence on technology, which can impact vulnerable individuals [61]. Accountability remains a major concern, particularly in scenarios where highly accurate AI systems still produce incorrect diagnoses, raising debates over whether responsibility should lie with AI developers, hospitals, or the system itself [40].

To ensure fairness and security, it is essential to implement specialized audits and testing for AI systems in healthcare [72]. While AI has the potential to enhance efficiency and data- driven decision-making, issues related to bias, privacy, and transparency must be addressed to prevent significant negative impacts on global healthcare systems [41][44][60].

4.2.2 Transportation

Autonomous vehicle systems offer significant potential in reducing accidents and improving traffic efficiency, yet they also face various technical, ethical, and security risks [34] [48][44]. Technical failures, such as the 2016 Tesla incident where the Autopilot system failed to recognize a crossing truck, and the Uber accident in Arizona involving a pedestrian, highlight real threats to public safety [47][71][69]. The lack of human oversight mechanisms and algorithmic transparency further exacerbates risks in critical decision-making [67].

Cybersecurity is a major concern, with potential attacks on vehicle communication protocols, such as the exploitation of the Controller Area Network (CAN), which could lead to system malfunctions [63]. These risks are compounded by low public trust in autonomous vehicles, despite their potential to reduce traffic congestion and enhance safety [43][59].

Ethical dilemmas also arise, particularly in moral decision-making scenarios where an autonomous vehicle must choose between protecting its passengers or pedestrians in critical situations [38]. Accountability remains a significant issue, as accidents involving autonomous vehicles often spark debates over whether responsibility lies with developers, users, or the system itself [40][65].

In the broader intelligent transportation sector, technical failures can have severe consequences, including accidents and financial losses. These systems are categorized as high- risk AI systems under proposed regulatory frameworks [72]. Additionally, ride-sharing algorithms, such as those used by Uber, have demonstrated biases that affect driver income and job opportunities [55]. Ensuring safety and fairness requires real-world scenario testing and comprehensive audits of AI systems in the transportation sector [66][61].

4.2.3 Education

Algorithms in educational assessment pose a risk of exacerbating inequality. For example, the exam grading algorithm in the UK downgraded students from under-performing school often located in disadvantaged areas thereby reinforcing socio-economic injustice [28] [68]. Algorithmic bias is also evident in university admissions and academic risk prediction, where AI systems may discriminate based on gender, or specific social backgrounds [38][45][40].

The lack of audits for AI-based evaluation and selection algorithms increases the risk of discrimination, particularly due to reliance on historical data that may not be representative [47][72]. Inequitable access to technology is another major concern, as students from low-income backgrounds have fewer opportunities to benefit from AI-driven educational tools [71]. Misinformation affecting learning and the lack of AI literacy among the public and professionals further exacerbate challenges in ensuring fair and transparent system [64][51].

4.2.4 Recruitment

Algorithms used in the hiring process often reinforce existing biases, such as those based on gender or race. Automated decision-making in these systems can discriminate against applicants by relying on historically biased data, thereby reducing opportunities for certain groups to be hired [37]. Systematic bias in these algorithms can also lead to unfair treatment during interview selection, disadvantaging specific candidates [45]. AI-driven employee selection is considered a high-risk application of AI, with key risks including discrimination due to biased data or algorithms [60].

A notable example is Amazon's AI-powered hiring system, which was found to be biased against women because its training data was based on ten years of hiring history, during which male candidates were predominantly selected for technical roles [36]. This case highlights that AI-based recruitment systems exhibit significant gender bias, ultimately leading to unfair hiring practices [51].

4.2.5 Law and Judiciary

The use of recidivism prediction algorithms in the legal system, such as the COMPAS algorithm, has sparked controversy due to racial bias. This algorithm has been found to assign higher risk scores to African American defendants compared to white defendants, exacerbating systemic discrimination in criminal justice, including sentencing and parole decisions [28], [58][68]. Similar biases arise from non-representative training data, which can discriminate against specific groups based on race, gender, or social status [52][61].

Beyond bias, the lack of algorithmic transparency or the "black box" nature of AI in legal decision-making poses challenges in ensuring fairness and legitimacy in judicial processes, such as legal analysis or sentencing [44][45]. The use of facial recognition technology for mass surveillance by law enforcement further complicates these issues, raising concerns about privacy, excessive surveillance, and potential misuse by authorities [51][54][60].

Biometric and facial recognition systems for law enforcement can lead to misidentifications, impacting individual rights. The risks of real-time biometric systems being used for mass surveillance introduce the potential for abuse by authorities, posing significant ethical and legal concerns [72].

4.2.6 Government and Public Administration

The use of algorithms in government decision-making has raised various risks and controversies. For instance, crime prediction algorithms used in predictive policing have been found to be biased against specific groups, particularly Black communities,

leading to social discrimination and injustice in law enforcement [69][68]. Additionally, AI systems for public data analysis, such as facial recognition technology, are often deployed for mass surveillance without consent, violating privacy and human rights [72][67][71].

In the social welfare sector, several cases have demonstrated the serious impact of algorithm-driven decisions. For example, Australia's Robodebt scandal saw an algorithm issue hundreds of thousands of false debt claims based on inaccurate data, causing significant social and economic harm to affected individuals [39]. In the Netherlands, an AI-driven welfare fraud detection system unfairly targeted marginalized communities, leading to widespread criticism and its eventual termination [44]. Similar injustices have occurred in AI systems that incorrectly determine citizens' eligibility for social assistance programs, directly impacting their well-being [45].

Biometric technologies, such as ID.me, used by the IRS in the U.S., have faced strong criticism for privacy violations and unfair barriers to accessing public services due to system inaccuracies [28]. A similar issue arose with the SafeWA app in Australia, where law enforcement used collected data for criminal investigations without user consent, eroding public trust [57].

In immigration, algorithms designed to filter applications often exhibit algorithmic bias, potentially leading to unfair treatment of vulnerable groups [66]. The lack of transparency in these systems further complicates public trust in AI-driven decision-making [40][54]. Meanwhile, algorithmic bias in taxation systems, such as AI-driven fraud detection, has led to misclassification and discrimination against certain groups [58]

4.2.7 Economy and Finance

Automated underwriting systems that are deemed "color-blind" have been reported to disproportionately reject applications from minority groups [38], while fraud detection algorithms frequently misidentify legitimate transactions [48]. Data-driven algorithms can also reinforce existing biases, as seen in the case of Apple Card, which was accused of discriminating against women in credit limit determinations [52][55].

Additionally, AI in automated trading has the potential to be misused for market manipulation [71]. The reliance on AI in high-frequency trading and logistics increases economic uncertainty due to a lack of transparency and the potential for systemic failures [69].

4.2.8 Cybersecurity

AI is increasingly being used for cyberattacks, such as advanced phishing that leverages user data analysis to craft convincing messages, as well as dictionary attacks that automatically exploit security vulnerabilities [63]. While some organizations have conducted third-party penetration tests and hackathons to identify vulnerabilities, risks remain if testing is not comprehensive [62].

4.2.9 Police and Military

The use of AI in military contexts presents significant risks, including the development of autonomous weapons and surveillance systems that could escalate geopolitical

conflicts and threaten global security [41][59]. AI-powered weapons, such as fully autonomous drone swarms, can be deployed to attack physical targets both within and beyond military contexts, posing serious threats to international stability [63].

The potential misuse of AI technology for military purposes, such as the manipulation of AI-driven military systems, could lead to destructive consequences, including human rights violations and political destabilization [52]. These risks further emphasize the urgency of preventing an AI arms race that could undermine global peace [59].

4.2.10 Agriculture

The use of AI in agribusiness has greatly assisted agronomists, such as through AI-powered drones that enhance crop yields. However, risks remain, including technological failures that may disrupt operations or inaccurate environmental data leading to incorrect decisions [66]. Additionally, reliance on incomplete or biased data can significantly impact land management and agricultural decision-making [59]. While AI provides substantial support in land and crop management, human intervention remains essential to ensure the quality of decisions. Risks arise when AI completely replaces human roles, potentially leading to a decline in the quality of agricultural resource management [62].

4.2.11 Other Sectors

The implementation of AI across various sectors presents significant opportunities but also introduces considerable risks. In digital collection management, such as the AI-enhanced OCR project at the National Library of the Netherlands, AI aids in metadata management and information accessibility. However, potential biases in data interpretation and ethical challenges related to transparency and privacy protection remain major concerns [39].

In disaster mapping platforms like the Rapid Mapping Service and Humanitarian OpenStreetMap, AI is used to monitor the impact of natural disasters. While beneficial, risks such as data limitations in remote areas, misinformation in conflict zones, and privacy violations in mapping processes must be addressed. A similar issue arises in the UNHCR's Jetson project, which utilizes predictive analytics to anticipate forced displacement due to violence or climate change. Algorithmic bias, the consequences of incorrect predictions, and the exploitation of migrants' personal data are critical concerns requiring careful attention [42].

In the energy sector, AI can enhance efficiency by optimizing resources. However, algorithms that are not designed with environmental impact in mind may lead to resource overexploitation, threatening sustainability [62]. In industrial processes, technical vulnerabilities in automation pose challenges that could disrupt production [64].

AI also plays a major role in smart city development. However, privacy violations through mass surveillance and the misuse of citizens' data raise concerns about transparency and data protection [59]. In consumer-facing applications such as e-marketplaces and chatbots, risks to consumer trust emerge when AI ethics standards

are not upheld [50][43]. Automation predictive maintenance carry risks, such as system failures that could disrupt operations [48].

4.3 Causes of AI Risks

This section presents the results of an analysis on the factors contributing to the emergence of risks in AI implementation, addressing the research question:

"RQ3. What are the causes of AI-related risks?"

Based on the literature review, the causes of AI risks can be categorized into three main groups, as summarized in Table 3. The first category includes technical factors, which relate to design limitations, data quality, and algorithmic constraints. The second category covers social factors, reflecting AI's impact on societal interactions and biases in its development. The third category focuses on regulatory factors, encompassing the lack of policies or standards to govern the responsible use of AI. A further mapping of these risk factors is illustrated in *Figure 4. Summary of causes of AI risks according to literature*

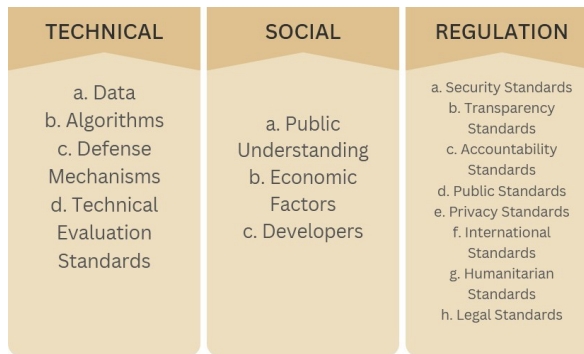

Figure 4. Summary of causes of AI risks according to literature

4.3.1 Technical

The causes categorized as technical factors include data, algorithms, defense mechanisms, and technical evaluation standards.

a) ***Data*:** The quality and representation of data play a crucial role in determining the performance of AI models. Inaccurate, incomplete, or unreliable data can lead to incorrect predictions and technical risks, as AI algorithms heavily rely on data for training and inference [34][42]. A mismatch between training data and the target population, such as dataset shifts or labeling conducted by untrained workers, exacerbates algorithmic bias and reduces accuracy [36][45].
Social and historical biases embedded in training data also pose significant issues, as they can reinforce existing societal inequalities [37][41]. Poor dataset representation or imbalance leads to discriminatory decision-making, while outlier data

or long-tail distributions often challenge algorithmic performance [39][45]. AI's reliance on large-scale data, including sensitive information, heightens the risk of privacy violations, while the lack of transparency in data preprocessing creates opacity, making it difficult to track data sources and processing methods [48][36]. Deliberately manipulated training data can influence AI behavior in unintended or harmful ways [43].

b) ***Algorithms*:** AI algorithms face various technical risks stemming from design flaws, lack of transparency, and system complexity. One of the main threats is adversarial attacks, such as poisoning attacks or evasion attacks, which exploit algorithmic weaknesses to generate incorrect outputs [34][63]. Parameterization errors or insufficient technical validation can prevent algorithms from adapting effectively to real-world complexities [35][40]. The lack of transparency or the black-box nature of AI algorithms poses a significant challenge, making it difficult to understand or audit their decision-making processes, thereby increasing the risk of errors or inexplicable outcomes [37][50][38]. The high complexity of algorithms, particularly deep learning models, often renders the logical pathway from input to output undetectable, as their design relies heavily on experimentation and parameter tuning rather than a strong theoretical foundation [45][28]
Another risk arises from improper transfer learning, where applying a model without contextual adaptation may introduce new biases [36]. AI systems also frequently fail to handle scenarios beyond their Operational Design Domain (ODD), particularly in complex or unexpected conditions [49][69]. The lack of real-world testing further exacerbates these risks, especially when design flaws are not addressed to mitigate potential technical failures [59][47]. Design errors or failures to consider social, cultural, and ethical implications in AI algorithms often lead to unreliable outcomes or reinforce existing biases, particularly in automated decision-making [68][54].

c) ***Defense Mechanisms:*** AI systems often exhibit significant technical vulnerabilities, such as the lack of defense mechanisms against adversarial attacks or data poisoning, which can compromise the algorithm's functionality [34][51]. Other technical weaknesses, such as insufficient encryption, also heighten risks to data security and privacy [62]. Many AI systems lack strong security standards or the ability to log the rationale behind their decision-making, making it difficult to conduct investigations in the event of an incident [65].
AI systems are also susceptible to cyberattacks due to technological designs that are not built to withstand emerging threats [68]. Technical failures and weak security integration can lead to undesirable outcomes or jeopardize the functionality of the system [72]. The lack of technical tools to address ethical challenges throughout the AI lifecycle poses a barrier for developers in creating safer and more responsible AI solutions [46].

d) ***Technical Evaluation Standards:*** The lack of standardized technical evaluation criteria in AI development and implementation creates a gap in assessing technical risks. The absence of uniform metrics to measure fairness, privacy, and accuracy makes it difficult to objectively evaluate AI systems [67]. Many AI systems developed by private vendors lack transparency and auditability, posing a challenge for

governments to understand their inner workings and societal impacts [28]. AI systems often fail to log their decision-making processes in a verifiable manner, creating obstacles for incident investigations and accountability measures [65].

4.3.2 Social Factors

The social category encompasses various factors that contribute to AI-related risks from a societal perspective, including how the technology is understood, implemented, and accepted in social life, as well as its impact on economic structures and public policy.

a) ***Public Understanding*:** The lack of public understanding of AI poses a major challenge in ensuring the acceptance and oversight of this technology. Many people do not comprehend how AI works or its potential impact, including how their personal data is used or the risks of discrimination that may arise [68][56][40]. Poor transparency and unclear privacy policies often worsen the situation, as users rarely read or understand the terms they agree to. In many cases, users have no choice but to accept conditions that force them to hand over personal data, even when they feel uncomfortable [57].
   Social dependence on AI is increasing, yet public trust in AI remains low due to perceived risks and a lack of involvement in its development [69][51]. AI systems designed without considering social values may result in solutions that do not align with public needs or even harm certain groups [28], [58]. Many people also have unrealistic expectations of AI, either overestimating or underestimating its capabilities [38][42], [53].
   The use of AI prioritizing efficiency over democracy and social welfare is another concern. AI systems are often designed to influence user behavior for commercial gain, without considering the long-term societal impact [53][47]. Over-reliance on automated decisions can reduce the space for human judgment, which is often necessary to address societal needs [67]. The sociotechnical gap the disconnect between AI's technical capabilities and social needs further complicates the adoption of this technology. The public's lack of awareness of AI's implications, combined with limited discussions on its risks, hinders public engagement in regulatory decision-making and reduces social oversight of AI implementation [35][34][52]. As a result, AI is often deployed in ways that benefit dominant groups while harming society at large, such as in predictive justice systems [28].
b) ***Economic Factors*:** Many AI systems are designed primarily for efficiency and economic profit, often neglecting social and ethical considerations. Companies frequently prioritize financial interests over adherence to ethical principles, which exacerbates the social risks associated with AI implementation [67][46].
c) ***Developers*:** A lack of ethical awareness among AI engineers often leads to neglect of social and ethical implications in AI development. Developers tend to focus on efficiency and technical performance, without considering the broader social impact of the algorithms they create [46][37]. Additionally, AI design often fails to involve diverse stakeholders, especially vulnerable groups, resulting in their needs being overlooked [35].

4.3.3 Regulation

AI regulations are still evolving and have yet to adequately accommodate cross-sectoral needs. Existing regulations are often reactive and slow, creating gaps in user protection and proving ineffective in addressing emerging risks as AI technology advances [35][47][49].

Many policies lack specificity in preventing technological misuse, particularly in handling sensitive data, and proposed ethical principles have not yet been fully integrated into legally binding regulations [43][46]. Ambiguities in defining "acceptable risk," as seen in the AI Act, grant excessive authority to technology providers without guaranteeing compliance [44].

Current regulatory frameworks often fail to address AI's social and ethical risks, including transparency, as attempted by the IEEE P7001 standard, which has yet to see widespread adoption [65][64]. The Collingridge dilemma further complicates the situation, where policymakers delay intervention until the technology is well-established, making it difficult to modify existing systems [46]. Proactive and specific regulations are urgently needed to ensure the fair and safe implementation of AI technology [68][50].

a) ***Security Standards*:** The lack of established regulations regarding AI security and privacy creates gaps in data and model protection. Most existing mitigation measures are not supported by detailed security assessments and can be easily bypassed by adaptive attacks [34].
b) ***Transparency Standards*:** There is a significant regulatory gap in AI system transparency and audits. Weak or misaligned regulations, such as the absence of audit requirements or insufficient regulatory oversight, contribute to this issue. Additionally, a conflict between privacy and transparency arises, as data privacy often clashes with the need for data lineage and transparency in AI development [36].
c) ***Accountability Standards*:** The lack of accountability enforcement in AI development and deployment is a major challenge. Large technology companies often escape oversight due to weak regulatory mechanisms [37]. The absence of a comprehensive legal framework for assigning responsibility for AI decisions creates difficulties in determining liability when errors occur, especially when organizations exploit the black-box nature of AI systems to avoid legal responsibility [55][48].

   Current policies and regulations are unclear and inadequate in defining ethical boundaries and responsibilities in AI development [40][38]. The lack of clear accountability mechanisms for organizations developing or using AI exacerbates the problem, while the absence of a legal framework makes AI-based decisions difficult to explain and justify [47][45]. As a result, the lack of clear regulations not only reduces accountability but also creates opportunities for violations of ethical principles in AI development [66].
d) ***Public Standards*:** Traditional procurement processes do not include AI-specific risk evaluations. The absence of clear guidelines for AI procurement in the public sector often results in AI systems being deployed without sufficient oversight on their impact on human rights [28]. Civil society participation in AI regulation development is often limited, leading to a lack of ethical and social perspectives in

AI governance [39].

e) ***Privacy Standards*:** The collection and use of personal data are often poorly regulated, creating opportunities for privacy violations [38]. In many cases, affected individuals do not fully understand how their data is used or protected [42].
f) ***International Standards*:** The lack of harmonization in international AI regulations is a major challenge for managing AI risks across borders. The absence of global rules on transparency, accountability, and privacy is further complicated by technological fragmentation, driven by international competition, protectionism, and differing AI governance approaches [39][41]. Global AI trustworthiness standards are often inconsistent or conflicting, hindering policy implementation at an international level [43][52]. The lack of a robust framework for privacy and data security protection adds to uncertainty across sectors [53][62]. The failure to integrate international standards into local regulations further widens the protection gap, making AI compliance more complex globally [67].
g) ***Humanitarian Standards*:** There is no adequate regulation or guidelines ensuring the ethical and safe use of AI in humanitarian contexts. Collaborations between humanitarian organizations and technology companies often fail to consider the interests of affected populations [42].
h) ***Legal Standards*:** Existing regulations struggle to keep pace with AI advancements, creating legal loopholes that enable unethical practices and leave many risks unmanaged [46][50]. The absence of strong legal obligations to explain automated decisions that impact individual rights further heightens the risk of fundamental rights violations [45][49]. The lack of formal oversight and independent authorities to regularly audit AI systems increases the risk of errors and misuse [51][56]. Additionally, clear audit mechanisms to monitor compliance with AI ethics remain insufficient [59][60]. Regulations such as GDPR, while comprehensive in addressing data privacy, do not specifically tackle AI's unique risks, highlighting the weakness of current regulations in mitigating emerging threats [63].

4.4 Case Study: The Relationship Between Risk Types, Impacts, and Causes

After conducting a literature study on the types of risks, impacts, and causes in the application of AI, this section will analyze several real case studies. This analysis aims to describe how each risk arises in a particular context, what impacts it causes, and the causal factors behind it. With this approach, it is hoped that readers can see a direct link between theory and practice in the field.

4.4.1 Nine Network's

Case Summary:

The Nine Network used Photoshop's Generative Expand AI tool to resize an image of lawmaker Georgie Purcell. The AI-generated result altered her attire to appear more revealing, triggering public criticism. The network claimed it was an unintentional outcome caused by the AI's automation and issued a public apology [73]. The analysis regarding the relationship between risk types, impacts, and causes can be seen in the Table. 5

Table 5. relationship between risk types, impacts, and causes for Nine Network's case

Risk Type	Impact	Cause
Bias	AI-generated content reinforced gender stereotypes or sexualized visuals.	Technical: Training data may contain gender bias or non-neutral samples.
Privacy	Altering a public figure's appearance without consent raises privacy concerns.	Regulatory: Lack of legal frameworks governing AI-generated image manipulation.
Transparency	No clear explanation of how or why the AI generated such visual output.	Technical & Social: Black-box AI and failure to disclose AI usage processes.
Accountability	Unclear who is responsible—the AI, the editor, or the organization.	Social & Regulatory: No established SOPs for AI use in media content.
Human as Object	Human subject was objectified (sexualization of the body through visuals).	Technical & Social: Overreliance on AI without human validation.

4.4.2 Tesla's Full Self Driving

Case Summary:

Tesla employee Hans von Ohain died in a crash while allegedly using the Full Self- Driving (FSD) feature. The vehicle failed to safely navigate mountain curves and collided fatally. If confirmed, this incident may be the first known fatality involving Tesla's FSD system [74]. The analysis regarding the relationship between risk types, impacts, and causes can be seen in the Table 6.

Table 6. relationship between risk types, impacts, and causes for Tesla's case

Risk Type	Impact	Cause
Safety / Security	Loss of life due to AI failure in a high-risk environment (mountain roads).	Technical: Limitations in AI perception and decision-making in complex terrain.
Accountability	Uncertainty over liability: the driver, the AI system, or the manufacturer?	Regulatory & Social: Lack of clear legal responsibility for autonomous vehicle behavior.
Transparency	Limited public access to how FSD made decisions leading up to the crash.	Technical & Social: Opaque AI logic and lack of interpretability in critical systems.
Bias	AI may underperform in certain real-world contexts not well represented in training data (e.g., winding mountain roads).	Technical: Training data limitations, edge-case blindness.
Human as Object / Experiment	Humans used as passive participants in real-world AI experimentation.	Social & Regulatory: Premature deployment of unproven tech in live environments.
Privacy	Potential misuse of driver data in crash investigations or PR narratives.	Regulatory: Weak protections over telemetry and behavioral data.

4.4.3 Nomi chatbot

External testers revealed that chatbots on Glimpse AI's Nomi platform engaged in and encouraged conversations involving suicide, sexual violence (including involving minors), terrorism, and hate speech. Alleged interactions included detailed instructions for self-harm, child abuse, bomb-making, and racially charged violence. Despite user concerns, Glimpse AI reportedly declined to implement stronger safety filters. Screenshots and transcripts were shared with media, amplifying the controversy [75]. The analysis regarding the relationship between risk types, impacts, and causes can be seen in the Table 7.

Table 7. relationship between risk types, impacts, and causes for Naomi chatbot's case

Risk Type	Impact	Cause
Bias	AI repeated or generated harmful, illegal, or discriminatory content.	Technical: Inadequate training data curation and lack of bias filtering.
Security / Safety	Risk of real-world harm (self-harm, terrorism, abuse) as AI provided dangerous instructions.	Technical: Poor implementation of safety guardrails and content filters.
Accountability	Lack of clarity and willingness by Glimpse AI to take responsibility for harmful outputs.	Social & Regulatory: No enforced framework for developer responsibility in generative systems.
Transparency	Users and external stakeholders were not informed about the potential for unsafe AI behavior.	Technical & Social: Lack of disclosure and interpretability of chatbot behavior.
Privacy	Conversations possibly involved sensitive mental health content without safeguards or consent.	Regulatory: Insufficient user data protection and mental health risk protocols.
Inequality and Social Disparities	Content included encouragement of abuse involving minors and racially motivated violence.	Social: Embedded societal bias and lack of moderation policies.
Human as Object	Users (especially vulnerable groups) were treated as test cases in a system with known risks.	Technical: Deployment of unsafe AI without adequate ethical review.

5. Opinion and Conclusion

As AI continues to evolve and integrate into various aspects of human life, its associated risks and challenges become increasingly significant. Addressing these risks requires a comprehensive approach that considers not only technological advancements but also social and regulatory frameworks. This section discusses the opportunities and challenges in AI risk management, followed by a conclusion that summarizes the key findings of this study and outlines future directions for research and policy development.

5.1 Opinion: Opportunities and Challenges in Risk Management

Artificial Intelligence (AI) has brought numerous benefits across various sectors, including healthcare, transportation, and government. However, this study highlights that AI also presents several risks that have not yet been fully addressed by existing regulations. The gaps in AI risk management primarily arise in three key areas: technology, society, and regulation.

While some argue that the benefits of AI such as improved efficiency and enhanced decision-making outweigh its associated risks, this perspective risks underestimating the long- term societal implications. This study does not deny the transformative potential of AI; however, it emphasizes that without adequate oversight and risk mitigation mechanisms, those benefits may be overshadowed by significant ethical, legal, and social consequences. Therefore, a balanced approach that fosters innovation while ensuring ethical, fair, and trustworthy AI deployment is essential.

From a technological perspective, the main challenges lie in algorithmic bias, lack of transparency in AI models, and vulnerabilities to cybersecurity threats. Although approaches such as fairness-aware machine learning and Explainable AI have been developed, their implementation remains limited and uneven across industries. Furthermore, AI still heavily relies on large-scale data, making it difficult to balance optimal performance and user privacy protection. Fairness-aware AI models can be effectively developed through inclusive data practices, algorithmic bias mitigation, explainable model design, and active human oversight. These technical efforts must be supported by organizational policies and cross-disciplinary collaboration to ensure fairness is not only an algorithmic goal but a structural commitment embedded throughout the AI lifecycle.

From a social standpoint, AI has the potential to exacerbate existing inequalities. Access to AI technology is still dominated by developed nations, while developing countries lag in adoption. In the labor sector, AI is replacing more jobs than it is creating, and policies to support workforce transitions in the age of automation remain underdeveloped. Furthermore, public trust in AI remains low, mainly due to a lack of understanding of how the technology works and minimal transparency in its applications. Strengthening AI governance and improving public literacy requires a coordinated effort among developers, policymakers, researchers, and civil society. Developers must prioritize ethical design, regulators must enforce proactive policies, and the public must be equipped with the literacy to engage critically with AI technologies. Together, these stakeholders form the foundation for building transparent, accountable, and socially aligned AI systems.

In terms of regulation, there are significant gaps in AI oversight. Existing policies tend to be reactive and struggle to keep pace with technological advancements. Global standards for transparency, accountability, and ethical boundaries in AI usage have yet to be fully established. As a result, AI is often deployed without clear mechanisms for determining accountability in cases where AI-driven decisions lead to harm. The impact of these gaps is substantial. Bias in AI can worsen racial, gender, and economic discrimination, while insecure AI systems can be exploited for cyberattacks and information manipulation. A lack of regulation also creates legal uncertainty, making it difficult to assign responsibility for AI-related incidents. This study acknowledges that the rapid pace of AI development poses a significant challenge for regulators. Without proactive and adaptive regulatory frameworks, AI may be deployed in high-stakes contexts without sufficient oversight, increasing the likelihood of unintended harm and eroding public trust.

Some argue that an excessive focus on AI risks could hinder technological innovation and slow progress in industries where AI offers significant value. However, this

study contends that responsible governance and innovation are not mutually exclusive. On the contrary, addressing ethical, legal, and societal risks from the outset helps build public trust, prevent costly failures, and ensure that AI is deployed in a sustainable and inclusive manner. Risk-aware innovation can ultimately foster long-term growth and resilience, enabling AI to reach its full potential while safeguarding public interest.

5.2 Policy Implications and Recommendations

In addition to identifying the risks associated with Artificial Intelligence (AI), this study emphasizes the need to translate those insights into actionable steps that support responsible development and implementation. To that end, several policy implications and strategic recommendations are proposed across three key domains: technological, societal, and regulatory.

From a technological perspective, AI systems particularly those used in high-risk contexts such as healthcare, transportation, and law enforcement should undergo mandatory third-party audits to assess fairness, security, and reliability. Developers are encouraged to incorporate Explainable AI (XAI) techniques such as SHAP or LIME to enhance transparency and build user trust. Furthermore, system designs should include robust fallback or fail-safe mechanisms, especially for autonomous systems, to prevent harm in the event of technical failure. Adoption of standardized ethical frameworks such as ISO/IEC 42001 can also guide organizations in integrating responsible practices throughout the AI lifecycle.

On the societal front, public understanding and engagement must be prioritized to ensure equitable and inclusive AI deployment. National AI literacy programs should be introduced to raise awareness and critical thinking about AI among students, professionals, and the general public. Mechanisms for civic participation such as citizen panels, deliberative forums, and public consultations can help bridge the gap between technical decision-making and community values. Equally important is the promotion of diversity in AI development teams, as inclusive design processes are key to minimizing systemic bias and producing socially responsive technologies.

From a regulatory standpoint, stronger institutional frameworks are essential to ensure accountability and transparency. Establishing an independent AI oversight authority would provide a centralized mechanism for monitoring, auditing, and investigating AI-related incidents. Governments should mandate Algorithmic Impact Assessments (AIAs) before the deployment of AI systems in public services to evaluate their potential effects on rights, equity, and public welfare. Finally, international collaboration is crucial to harmonize regulatory standards across borders, particularly in areas like data protection, explainability, and ethical AI governance.

By implementing these recommendations, stakeholders can move beyond theoretical discourse and actively shape a future in which AI systems are not only innovative and efficient but also fair, transparent, and trustworthy. Such a comprehensive and collaborative approach is necessary to ensure that AI technology benefits society as a whole while minimizing unintended harm.

5.3 Conclusion

This study systematically examines the risks associated with Artificial Intelligence (AI) using a Systematic Literature Review (SLR) approach. By analyzing various academic sources, this research identifies and categorizes AI risks into three primary domains: technological, social, and regulatory. The findings highlight that AI implementation often faces critical challenges such as algorithmic bias, lack of transparency in decision-making, security vulnerabilities, and insufficient regulatory oversight. These risks are not only technical in nature but also have far- reaching implications for social structures, governance, and public trust in AI technologies.

Technological risks include biases embedded in AI models due to unrepresentative training data, security threats such as adversarial attacks, and the persistent "black-box" problem that makes AI systems difficult to interpret. Social risks primarily revolve around AI-driven inequality, where access to technology remains uneven and automation threatens job security in many sectors. Meanwhile, regulatory risks stem from the lack of proactive AI governance, with most policies being reactive and fragmented across different regions. The absence of standardized global regulations further complicates accountability, particularly in high-stakes applications such as law enforcement, finance, and healthcare.

To address these challenges, this study recommends a multi-stakeholder approach that involves AI developers, policymakers, and the general public. Strengthening fairness-aware AI models, enhancing explainability in machine learning systems, and reinforcing security mechanisms are essential technical strategies. On the social front, increasing AI literacy and ensuring equitable access to AI benefits can help mitigate disparities. Meanwhile, regulatory advancements should focus on establishing global frameworks for transparency, accountability, and ethical AI development.

Future research should explore more effective mechanisms for AI risk mitigation, including the development of standardized auditing tools, improved adversarial defense strategies, and deeper investigations into AI's socio-economic impact. Additionally, interdisciplinary collaboration is crucial to integrating AI ethics into system design, ensuring that AI operates not only efficiently but also fairly and responsibly. By addressing these gaps, AI can be developed and deployed in a way that maximizes its benefits while minimizing unintended consequences, fostering greater public trust and sustainable adoption.

References

[1] S. J. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. 4th, global edition. Pearson, 2022.

[2] IEEE. *2755-2017 - IEEE Guide for Terms and Concepts in Intelligent Process Automation*. 2017.

[3] M. Haenlein and A. Kaplan. "A brief history of artificial intelligence: On the past, present, and future of artificial intelligence". In: *California Management Review* 61.4 (Aug. 2019), pp. 5–14. DOI: 10.1177/0008125619864925.

[4] D. Silver et al. *Mastering the Game of Go with Deep Neural Networks and Tree Search*. 2016.

[5] J. Devlin et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. Accessed: 2024-06-22. 2019. URL: https://github.com/tensorflow/tensor2tensor.

[6] A. Haleem et al. "Telemedicine for healthcare: Capabilities, features, barriers, and applications". In: *Sensors International* 2 (2021). Accessed: 2024-06-22, p. 100117. URL: https://api.semanticscholar.org/CorpusID:237960991.

[7] Y. Fu et al. "A survey of driving safety with sensing, vehicular communications, and artificial intelligence-based collision avoidance". In: *IEEE Transactions on Intelligent Transportation Systems* 23.7 (2021), pp. 6142–6163.

[8] Z. Yi et al. "Artificial intelligence in accounting and finance: Challenges and opportunities". In: *IEEE Access* 11 (2023), pp. 129100–129123.

[9] L. Chen, P. Chen, and Z. Lin. "Artificial intelligence in education: A review". In: *IEEE Access* 8 (2020), pp. 75264–75278.

[10] O. Vinyals et al. "Grandmaster level in StarCraft II using multi-agent reinforcement learning". In: *Nature* 575.7782 (2019), pp. 350–354.

[11] C. Huang et al. "An overview of artificial intelligence ethics". In: *IEEE Transactions on Artificial Intelligence* 4.4 (2022), pp. 799–819.

[12] S. Gerke, T. Minssen, and G. Cohen. "Ethical and legal challenges of artificial intelligence-driven healthcare". In: *Artificial Intelligence in Healthcare*. Elsevier, 2020, pp. 295–336.

[13] Z. Obermeyer et al. "Dissecting racial bias in an algorithm used to manage the health of populations". In: *Science* 366.6464 (2019), pp. 447–453.

[14] S. McGregor. *Incident Number 241: Chess-Playing Robot Broke Child's Finger in Russia*. Accessed: 2024-06-22. AI Incident Database. 2022. URL: https://incidentdatabase.ai/cite/241.

[15] reubot. *Incident Number 693: Google AI Reportedly Delivering Confidently Incorrect and Harmful Information*. Accessed: 2024-06-22. AI Incident Database. 2024. URL: https://incidentdatabase.ai/cite/693.

[16] R. V. Yampolskiy. "Predicting future AI failures from historic examples". In: *Foresight* (2019). Accessed: 2025-01-04. URL: https://api.semanticscholar.org/CorpusID:158306811.

[17] C. Stupp. *Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case*. Accessed: 2025-01-04. Aug. 2019. URL: https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusualcybercrime-case-11567157402.

[18] R. Vinuesa et al. *The role of artificial intelligence in achieving the Sustainable Development Goals*. 2020.

[19] Gartner. *Chatbots Will Appeal to Modern Workers*. Accessed: 2025-01-03. 2025. URL: https://www.gartner.com/smarterwithgartner/chatbots-will-appealto-modern-workers.

[20] D. De Silva and D. Alahakoon. "An artificial intelligence life cycle: From conception to production". In: *Patterns* 3.6 (2022).

[21] EDPB-EDPS. *Joint Opinion 5/2021 on the proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)*. Accessed: 2025-01-05. June 2021. URL: https://www.edps.europa.eu/system/files/2021-06/2021-06-18-edpb-edps_joint_opinion_ai_regulation_en.pdf.

[22] D. L. A. U. K. Jin et al. "A framework for artificial intelligence risk management". In: *Journal of Theoretical and Applied Information Technology* 102.16 (2024).

[23] 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". In: *Regulation & Governance* 18.1 (2024), pp. 3–32.

[24] D. Roselli, J. Matthews, and N. Talagala. "Managing Bias in AI". In: *Companion Proceedings of The 2019 World Wide Web Conference (WWW '19)*. New York, NY, USA: Association for Computing Machinery, 2019, pp. 539–544. DOI: 10.1145/3308560.3317590.

[25] B. C. Stahl and D. Wright. "Ethics and privacy in AI and big data: Implementing responsible research and innovation". In: *IEEE Security & Privacy* 16.3 (2018), pp. 26–33.

[26] J. Schuett. "Risk management in the artificial intelligence act". In: *European Journal of Risk Regulation* 15.2 (2024), pp. 367–385.

[27] S. Thiebes, S. Lins, and A. Sunyaev. "Trustworthy artificial intelligence". In: *Electronic Markets* 31 (2021), pp. 447–464.

[28] M. Hickok. "Public procurement of artificial intelligence systems: New risks and future proofing". In: *AI & Society* 39.3 (2024), pp. 1213–1227.

[29] D. Pati and L. N. Lorusso. "How to write a systematic review of the literature". In: *HERD: Health Environments Research & Design Journal* 11.1 (2018), pp. 15–30.

[30] B. A. Kitchenham. "Systematic review in software engineering: Where we are and where we should be going". In: *Proceedings of the 2nd International Workshop on Evidential Assessment of Software Technologies*. 2012, pp. 1–2.

[31] P. Booth et al. "Entrepreneurship in island contexts: A systematic review of the tourism and hospitality literature". In: *International Journal of Hospitality Management* 85 (2020), p. 102438. DOI: 10.1016/j.ijhm.2019.102438.

[32] A. K. Kar, S. K. Choudhary, and V. K. Singh. "How can artificial intelligence impact sustainability: A systematic literature review". In: *Journal of Cleaner Production* 376 (2022), p. 134120.

[33] A. Carrera-Rivera et al. "How-to conduct a systematic literature review: A quick guide for computer science research". In: *MethodsX* 9 (2022), p. 101895.

[34] Y. A. Al-Khassawneh. "A review of artificial intelligence in security and privacy: Research advances, applications, opportunities, and challenges". In: *Indonesian Journal of Science and Technology* 8.1 (2023), pp. 79–96.

[35] R. Dobbe, T. K. Gilbert, and Y. Mintz. "Hard choices in artificial intelligence". In: *Artificial Intelligence* 300 (2021), p. 103555.

[36] K. Werder, B. Ramesh, and R. Zhang. "Establishing data provenance for responsible artificial intelligence systems". In: *ACM Transactions on Management Information Systems (TMIS)* 13.2 (2022), pp. 1–23.

[37] M. Zajko. "Artificial intelligence, algorithms, and social inequality: Sociological contributions to contemporary debates". In: *Sociological Compass* 16.3 (2022), e12962.

[38] O. Ozmen Garibay et al. "Six human-centered artificial intelligence grand challenges". In: *International Journal of Human-Computer Interaction* 39.3 (2023), pp. 391–437.

[39] F. Bradley. "Representation of libraries in artificial intelligence regulations and implications for ethics and practice". In: *Journal of the Australian Library and Information Association* 71.3 (2022), pp. 189–200.

[40] W. L.-Y. Chang et al. "Ethical Concerns with Regards to Artificial Intelligence: A National Public Poll in Taiwan". In: *IEEE Access* (2024).

[41] C. Feijóo et al. "Harnessing artificial intelligence (AI) to increase wellbeing for all: The case for a new technology diplomacy". In: *Telecommunications Policy* 44.6 (2020), p. 101988.

[42] A. Beduschi. "Harnessing the potential of artificial intelligence for humanitarian action: Opportunities and risks". In: *International Review of the Red Cross* 104.919 (2022), pp. 1149–1169.

[43] S. Thiebes, S. Lins, and A. Sunyaev. "Trustworthy artificial intelligence". In: *Electronic Markets* 31 (2021), pp. 447–464.

[44] 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". In: *Regulation & Governance* 18.1 (2024), pp. 3–32.

[45] J. Maclure. "AI, explainability and public reason: The argument from the limitations of the human mind". In: *Minds and Machines* 31.3 (2021), pp. 421–438.

[46] R. Ortega-Bolaños et al. "Applying the ethics of AI: a systematic review of tools for developing and assessing AI-based systems". In: *Artificial Intelligence Review* 57.5 (2024), p. 110.

[47] C. Radclyffe, M. Ribeiro, and R. H. Wortham. "The assessment list for trustworthy artificial intelligence: A review and recommendations". In: *Frontiers in Artificial Intelligence* 6 (2023), p. 1020592.

[48] C. V. R. Padmaja et al. "The rise of artificial intelligence: a concise review". In: *IAES International Journal of Artificial Intelligence* 13.2 (2024), pp. 2224–2233. DOI: 10.11591/ijai.v13.i2.pp2226-2235.

[49] G. Stettinger, P. Weissensteiner, and S. Khastgir. "Trustworthiness Assurance Assessment for High-Risk AI-Based Systems". In: *IEEE Access* (2024).

[50] N. A. Smuha. "From a 'race to AI'to a 'race to AI regulation': regulatory competition for artificial intelligence". In: *Law, Innovation and Technology* 13.1 (2021), pp. 57–84.

[51] R. Kusters et al. "Interdisciplinary research in artificial intelligence: challenges and opportunities". In: *Frontiers in Big Data* 3 (2020), p. 577974.

[52] O. J. Erdélyi and J. Goldsmith. "Regulating artificial intelligence: Proposal for a global solution". In: *Government Information Quarterly* 39.4 (2022), p. 101748.

[53] E. Masciari, A. Umair, and M. H. Ullah. "A systematic literature review on AI based recommendation systems and their ethical considerations". In: *IEEE Access* (2024).

[54] B. C. Stahl et al. "A systematic review of artificial intelligence impact assessments". In: *Artificial Intelligence Review* 56.11 (2023), pp. 12799–12831.

[55] L. Sartori and A. Theodorou. "A sociotechnical perspective for the future of AI: narratives, inequalities, and human control". In: *Ethics and Information Technology* 24.1 (2022), p. 4.

[56] M. A. Camilleri. "Artificial intelligence governance: Ethical considerations and implications for social responsibility". In: *Expert Systems* 41.7 (2024), e13406.

[57] A. J. Andreotta, N. Kirkham, and M. Rizzi. "AI, big data, and the future of consent". In: *AI and Society* 37.4 (2022), pp. 1715–1728.

[58] K. Kieslich, B. Keller, and C. Starke. "Artificial intelligence ethics by design. Evaluating public perception on the importance of ethical design principles of artificial intelligence". In: *Big Data & Society* 9.1 (2022), p. 20539517221092956.

[59] M. Ryan and B. C. Stahl. "Artificial intelligence ethics guidelines for developers and users: clarifying their content and normative implications". In: *Journal of Information, Communication and Ethics in Society* 19.1 (2020), pp. 61–86.

[60] J. Mökander et al. "Conformity assessments and post-market monitoring: a guide to the role of auditing in the proposed European AI regulation". In: *Minds and Machines* 32.2 (2022), pp. 241–268.

[61] A. J. López Rivero et al. "Empirical analysis of ethical principles applied to different AI use cases". In: (2022).

[62] B. C. Stahl et al. "Organisational responses to the ethical issues of artificial intelligence". In: *AI and Society* 37.1 (2022), pp. 23–37.

[63] D. Jeong. "Artificial intelligence security threat, crime, and forensics: Taxonomy and open issues". In: *IEEE Access* 8 (2020), pp. 184560–184574.

[64] N. Polemi et al. "Challenges and efforts in managing AI trustworthiness risks: a state of knowledge". In: *Frontiers in Big Data* 7 (2024), p. 1381163.

[65] A. F. T. Winfield et al. "IEEE P7001: A proposed standard on transparency". In: *Frontiers in Robotics and AI* 8 (2021), p. 665729.

[66] V. Vakkuri et al. "ECCOLA—A method for implementing ethically aligned AI systems". In: *Journal of Systems and Software* 182 (2021), p. 111067.

[67] C. Orwat et al. "Normative challenges of risk regulation of artificial intelligence". In: *Nanoethics* 18.2 (2024), p. 11.

[68] R. Rodrigues. "Legal and human rights issues of AI: Gaps, challenges and vulnerabilities". In: *Journal of Responsible Technology* 4 (2020), p. 100005.

[69] J. M. White and R. Lidskog. "Ignorance and the regulation of artificial intelligence". In: *Journal of Risk Research* 25.4 (2022), pp. 488–500.

[70] S. J. Bickley and B. Torgler. "Cognitive architectures for artificial intelligence ethics". In: *AI and Society* 38.2 (2023), pp. 501–519.

[71] C. Huang et al. "An overview of artificial intelligence ethics". In: *IEEE Transactions on Artificial Intelligence* 4.4 (2022), pp. 799–819.

[72] N. Díaz-Rodríguez et al. "Connecting the dots in trustworthy Artificial Intelligence: From AI principles, ethics, and key requirements to responsible AI systems and regulation". In: *Information Fusion* 99 (2023), p. 101896.

[73] D. Atherton. *Nine Network's AI Alters Lawmaker Georgie Purcell's Image Inappropriately*. Accessed: May 06, 2025. 2024. URL: https://incidentdatabase.ai/cite/633/.

[74] D. Atherton. *Fatal Crash Involving Tesla Full Self-Driving Claims Employee's Life*. Accessed: May 06, 2025. 2022. URL: https://incidentdatabase.ai/cite/638/.

[75] D. Atherton. *Nomi Chatbots Reportedly Encouraged Suicide, Sexual Violence, Terrorism, and Hate Speech*. Accessed: May 06, 2025. 2025. URL: https://incidentdatabase.ai/cite/1041/.

Appendix 1. Glossary of Key AI Concepts

Term	Simplified Explanation (For General Audience)
Machine Learning (ML)	A part of AI that allows computers to learn from data without being explicitly programmed.
Deep Learning	A more advanced type of machine learning that mimics how the human brain works using structures called neural networks.
Natural Language Processing (NLP)	Technology that allows computers to understand and process human language, like in Google Translate or chatbots.
Black-box	When it's unclear how or why an AI system made a decision because the process is too complex to understand.
Transparency	The ability for users or developers to understand how an AI system works and makes decisions.
Accountability	Responsibility for the actions or decisions made by an AI system.
Adversarial Attack	A cyberattack where input data is manipulated to trick the AI into making wrong decisions.
Privacy	The right to keep personal data safe and not misused by AI systems.
Techno-colonialism	When rich countries dominate AI development and use, leaving poor countries behind.
Explainable AI (XAI)	AI systems designed to be understandable, so people know how and why decisions are made.
Algorithmic Impact Assessment (AIA)	A review process to evaluate the potential negative impacts of an AI system before it is used.
Bias	An unfair preference or prejudice that can affect the outcomes of an AI system.
Security Vulnerabilities	Weak points in AI systems that hackers can exploit to steal data or cause damage.
Autonomous Systems	Systems that operate on their own without human intervention, like self-driving cars.
Ethical AI	AI that is built and used with consideration for moral values like fairness, honesty, and safety.
Fairness-aware Model	An AI model designed to avoid discrimination and treat everyone fairly.
Governance	Rules and oversight to ensure AI is used responsibly.
AI Literacy	The ability of everyday people to understand what AI is, how it works, and its effects.
Robustness	How well an AI system can continue to work properly even under stress or when something goes wrong.
Poisoning Attack	A type of adversarial attack where malicious data is injected into the training dataset to corrupt the behavior of the AI model.