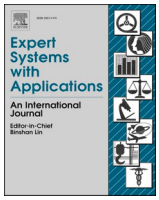
Expert Systems With Applications 240 (2024) 122442

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa



Review

Artificial Intelligence Trust, Risk and Security Management (AI TRiSM): Frameworks, applications, challenges and future research directions

Adib Habbal [a,*], Mohamed Khalif Ali [a,b,*], Mustafa Ali Abuzaraida [c]

[a] Department of Computer Engineering, Faculty of Engineering, Karabuk University, Karabuk, Türkiye
[b] Department of Computer Science, Faculty of Engineering and Computer Technology, Somali International University (SIU), Mogadishu, Somalia
[c] Department of Computer Science, Faculty of Information Technology, Misurata University, Misurata, Libya

ARTICLE INFO

Keywords:
AI TRiSM framework
AI TRiSM Orchestration
AI TRiSM Adaptive
ModelOps
Deepfake Technology and Adversarial attacks

ABSTRACT

Artificial Intelligence (AI) has become pervasive, enabling transformative advancements in various industries including smart city, smart healthcare, smart manufacturing, smart virtual world and the Metaverse. However, concerns related to risk, trust, and security are emerging with the increasing reliance on AI systems. One of the most beneficial and original solutions for ensuring the reliability and trustworthiness of AI systems is AI Trust, Risk and Security Management (AI TRiSM) framework. Despite being comparatively new to the market, the framework has demonstrated already its effectiveness in various products and AI models. It has successfully contributed to fostering innovation, building trust, and creating value for businesses and society. Due to the lack of systematic investigations in AI TRiSM, we carried out a comprehensive and detailed review to bridge the existing knowledge gaps and provide a better understanding of the framework from both theoretical and technical standpoints. This paper explores various applications of the AI TRiSM framework across different domains, including finance, healthcare, and the Metaverse. Futhermore, the paper discusses the obstacles related to implementing AI TRiSM framework, including adversarial attacks, the constantly changing landscape of threats, ensuring regulatory compliance, addressing skill gaps, and acquiring expertise in the field. Finally, it explores the future directions of AI TRiSM, emphasizing the importance of continual adaptation and collaboration among stakeholders to address emerging risks and promote ethical and enhanced overall security bearing for AI systems.

1. Introduction

The progress in technology has brought about the existence of Artificial Intelligence educators, or more broadly, automated teaching systems. AI has become a powerful influence that is altering different aspects of society, economy, and technology (Kim, 2022). The continuous advancements in computing capabilities have led to a substantial growth of AI technologies, spanning from machine learning to natural language processing and voice recognition (Sohn and Kwon, 2020; Abuzaraida et al., 2021) and AI-driven solutions are gaining more importance and exerting a greater impact on individuals' lives and our societies (Glikson and Woolley, 2020). These days, AI-based applications are dominant in practically every aspects of daily lives from product recommendation engines (Kim et al., 2022), smart city (Nikitas, 2020), education (Chen et al., 2020) and autonomous cars (Ma, 2020). Moreover, Its extensive adoption in crucial sectors like healthcare (Yu et al., 2018), finance (Sanz and Zhu, 2021), transportation (Bharadiya, 2023), and communications (Lv et al., 2020) has driven progress that offers exceptional advantages.

The widespread adoption of AI technologies raises significant concerns related to trust, risk, and security. On the other hand, the idea of trust in AI involves the assurance that users and involved parties have in the dependability, honesty, and ethical application of AI systems. On one hand, AI risks encompass the possible negative outcomes and uncertainties linked to AI algorithms and systems. These risks include biases, unintended results, violations of data privacy, and the possibility of harm caused by AI. To elaborate the solution of aforementioned problems, AI TRiSM is suitable framework to offer solutions providing a structured approach to evaluating the trustworthiness of AI systems by assessing their transparency, explainability, and accountability.

AI TRiSM is a comprehensive framework designed to handle the challenges related to Artificial Intelligence (AI) systems, enabling fairness, governance, efficacy, reliability and privacy. The AI TRiSM framework is designed to assist organizations developing a systematic

* Corresponding authors.
E-mail addresses: adibhabbal@karabuk.edu.tr (A. Habbal), khaliifcade@siu.edu.so (M.K. Ali), abuzaraida@it.misuratau.edu.ly (M.A. Abuzaraida).

https://doi.org/10.1016/j.eswa.2023.122442
Received 16 August 2023; Received in revised form 24 October 2023; Accepted 2 November 2023

approach to managing the risks associated with AI, including data privacy, risks related to security and ethical related concerns. The idea of AI TRiSM is a relatively recent concept that has gained significance attention as humans increasingly depend on AI-powered systems for complex tasks in today's society. A recent Gartner® report discussed the concept of AI TRiSM framework and highlighted the importance of organizations implementing its strategy that combines specialized tools to fully leverage the benefits it offers (Groombridge, 2022). Furthermore, AI TRiSM encompasses multiple elements such as transparency, responsibility, fairness, reliability, ethical considerations, and others. By embracing AI TRiSM framework, organizations can acquire a deeper understanding of the processes involved in designing, developing, and deploying AI models. AI TRiSM is expected to enable effective monitoring and mitigation of risks while ensuring the dependability and trustworthiness of the AI systems (Groombridge, 2022).

Current frameworks might not possess thorough the mechanisms essential to establish and maintain trust in AI systems, properly evaluate the risks linked with AI implementations, and efficiently handle security within the rapidly changing AI environment (Mahbooba, 2021). Moreover, the intricates and constantly changing characteristics of AI applications, coupled with the evolving landscape of potential threats, present distinct challenges that require customized strategies. Addressing these problems is crucial for advancing AI's trustworthy deployment, minimizing risks, and fortifying security, thereby fostering more reliable and secure AI ecosystem for broader applications (Elahi, 2021). The challenging question is "how can we manage trust, risk and security of AI-based solutions using comprehensive framework design which is based on transparency, responsibility, fairness, reliability and ethical considerations?" In addressing this inquiry, researchers have focused on understanding the subject of AI TRiSM to justify the AI-based solution reliability, transparency, explainability, accountability and trustworthy.

In this review, we conduct a thorough examination of academic literature by delving into the realm of AI Trust, Risk, and Security Management within AI systems. The review stands apart from previous research in four distinct manners: In particular, this study aims to: 1) evaluate and synthesize existing frameworks and methodologies intended to cultivate trust, mitigate risks, and enhance security in AI applications; 2) reveal and clarify five challenges and constraints encountered when implementing AI TRiSM frameworks efficiently.; 3) Concentrate on AI trust, risk, and security management and some areas that may be applicable; and 4) discus the potential advancements in the field of AI TRiSM framework.

We contribute to the literature on AI TRiSM within AI systems in four ways. Firstly, the review offers a consolidated and critical evaluation of existing frameworks, shedding light on their strengths and weaknesses in fostering trust, managing risks, and ensuring security in AI applications. Secondly, by analyzing real-world applications of AI TRiSM, the review provides insights into practical implementations and their impact on diverse sectors. Thirdly, the identification and delineation of prevalent challenges offer a foundational understanding necessary for devising effective strategies to overcome obstacles in the field. Lastly, the researchers proposed future research directions aiming to guide and inspire scholars, researchers, and practitioners in addressing emerging challenges and advancing AI TRiSM to meet the evolving demands of the technological landscape.

Our study extensively explores several crucial aspects of AI TRiSM, including frameworks, applications, challenges, and future directions. The paper is structured as follows: The initial sections provide an introduction to AI and AI TRiSM by balancing the current levels of trust, risks and security in AI and their consequences. Subsequently, we present how existing frameworks for AI Trust, Risk, and Security operate, AI TRiSM framework strategies and its application will be explored in the development of AI-based systems. The challenges and future directions of AI TRiSM will be deliberated in the last section. Finally, the research conclusions of this paper will be drawn.

2. Balancing AI trust, risk and security

Trust is identified as a key element for successful AI integration, with studies emphasizing the need for transparency, explainability, fairness, and accountability to establish trust among users and stakeholders (Lamsal, 2001). Additionally, risk and security considerations are paramount, as AI systems may introduce new vulnerabilities and threats. On the other hand, ensuring the ethical deployment and responsible of well trusted AI systems necessitates the crucial task of balancing AI trust, risk, and security. It involves considering various influences to establish and maintain trust, mitigate possible risks, and safeguard the security of AI systems. There is a high level of expectation for AI solutions to effectively tackle existing challenges. However, the increasing visibility of AI solutions that fail to deliver the promised advancements poses a risk of undermining user confidence in AI systems (Roski, 2021). Here are some key ideas to deliberate:

2.1. AI trust management conceptualizing

By definition, AI is distinct from natural intelligence, being a created and machine-based form of intelligence. In order to enhance decision-making within intricate and unpredictable systems, AI technology has been integrated, holding the promise to revolutionize medical approaches. AI models are engineered by humans to perform intricate tasks and process information akin to our own cognitive processes. The effects of initial AI systems are currently evident, presenting both challenges and prospects, and establishing the groundwork for the seamless integration of future AI advancements into social and economic sectors. (Kumar et al., 2023; Bedemariam and Wessel, 2023).

Over the past few years, the research community has shown significant interest in the domain of trust and reputation management (Zhang, 2020). The utilization of widely-accepted, interdisciplinary definitions to describe trust as a psychological state encompasses a willingness to embrace vulnerability driven by optimistic beliefs about the intentions or actions of another entity. Trusting convictions entail a trustor's view that a trustee possesses qualities that will be advantageous to the trustor (Cabiddu, 2022). Trust plays a critical role in the acceptance and adoption of AI technology, as individuals are more persuaded to use and depend on AI systems when they perceive them as reliable. Given the intricate algorithms employed by AI systems, privacy concerns arise for both individuals and organizations. Trust in AI involves a range of elements at the levels of society, organizations, and individuals that can enhance or diminish faith in AI technologies. Bais, discrimination, and privacy invasion are among the key factors concerning trust in AI.

Bias and Discrimination: The rise of AI techniques like machine learning is intensifying the issue of digital discrimination, with an increasing number of decisions being delegated to these systems (Ferrer, 2021). AI solutions propose that predictive algorithms can analyze data in a more precise and impartial manner compared to humans by foreseeing intricate individual behaviors. Their suggestions also indicate that the outcomes produced by predictive algorithms have the ability to reduce inherent human biases and offer additional advantages, such as uniformity, impartiality, and objectivity in assessing information related to an offender. On the flip side, significant concerns emerge regarding the court's utilization of biased and discriminatory data, including demographic and socio-economic factors, as inputs in the existing predictive AI system. Similarly, as they claim to generate biased and discriminatory results, these AI systems have a negative impact on the rights of individuals, principles of adjudication, and overall judicial integrity. This can lead to undesirable effects such as algorithmic bias, racial discrimination, and a surge in incarceration rates due to an elevated dependence on these predictive algorithms within a specific criminal justice system, rendering such automated risk-assessment systems deeply concerning from constitutional, technical, and ethical standpoints (Malek, 2022). Moreover, in 2017, Amazon discontinued its AI-driven candidate assessment recruitment tool due to evidence

indicating gender bias. The tool exhibited prejudiced behavior by giving lower ratings to resumes of female candidates, revealing a bias stemming from an insufficient representation of women in the training data used to develop the model (Mujtaba and Mahapatra, 2019). Lastly, the occurrence of biased results, whether by accident or due to systemic issues, fosters doubt and worry, impeding broad trust in AI's capability to act fairly and impartially. It is essential to tackle and reduce these adverse impacts to build confidence and guarantee the responsible and impartial use of AI across different sectors.

Privacy Invasion: As AI advances, it gains the capability to draw conclusions based on intricate data patterns that may elude human perception. AI systems typically depend on extensive data for effective training and functioning, which can pose a risk to privacy if sensitive data is mishandled or used inappropriately. Consequently, individuals may not even realize that their personal data is being used to form decisions that will affect them. Creating human trust in AI-based systems holds enormous importance in various domains, such as critical medical care decisions that can be a matter of life and death, as well as significant financial and transportation choices (Asan et al., 2020; Nicodeme, 2020). As AI systems increasingly take on these crucial decision-making roles, it becomes imperative for developers to design systems that embody transparency, reliability, and accountability. Moreover, Health-related data, preserving an individual's privacy concerning health data is a fundamental ethical value, given its direct link to personal well-being and identity. Safeguarding patients' confidentiality is crucial to prevent any unauthorized or secondary utilization of their data. Failing to meet patients' privacy requirements could result in psychological and reputational damage to them. The potential for data breaches heightens concerns about AI models that share personal health data. The worry is that AI processes could reidentify supposedly anonymized data, amplifying fears of privacy infringement and data breaches (Esmaeilzadeh, 2020).

Moreover, effective communication with users regarding the inner workings and limitations of AI systems is also vital to foster trust. Additionally, the implementation of regulations and standards plays a significant role in building trust by ensuring responsible and ethical development and usage of AI systems. Conversely, when comparing trust in interpersonal connections, where both parties are human, to relationships involving technology or machines, the trust can be placed either in the technology itself or the organization providing it. Trust in technology and the trust in the provider are interdependent, meaning that an increase in trust in one component will also lead to an increase in trust in the other (Wazan, 2017).

2.2. AI risk management conceptualizing

AI risk involves identifying possible threats and risks associated with AI systems. It encompasses examining the competences, constraints, and possible failure modes of AI technologies. In order to detect vulnerabilities, techniques like adversarial testing (Park, 2022) and verification data (Benmoussa, 2022) may also be employed. AI is proving to have both positive and negative implications, posing challenges for organizations and individuals alike. Many are currently struggling with the issues tied to AI systems, which give rise to unintended risks. These risks, at times, can result in severe consequences, including bias (Nelson, 2019), privacy violations (Zhu, 2020), discrimination (Van Bekkum and Borgesius, 2023); accidents (Hadj-Mabrouk, 2019), and political manipulation systems (Peters, 2022). Here are some considerations when conceptualizing risks in AI.

***Society Manipulation*:** One of an unintended risk outcome of the recent AI advancement is the manipulation of social dynamics (King, 2020). Social media has become omnipresent in modern society, serving a multitude of functions, including entertainment, spreading information, engaging in political discourse, and promoting businesses (Marr, 2018). Utilization of social media has brought to the forefront significant societal, cultural, and moral concerns, encompassing topics like privacy, cyberbullying, filter bubbles, and the propagation of misinformation. Specifically, allegations have been made against social media platforms for employing algorithms to control the content that users encounter in their feeds, with the aim of advancing specific political or commercial agendas. Personalized search algorithms strive to enhance the search process for users by presenting results that align with their interests and requirements, aiming for a more efficient experience. Nevertheless, these algorithms have faced criticism for their potential to uphold current biases, restrict information diversity, and foster filter bubbles. A significant hurdle associated with personalized search algorithms is their dependence on extensive user data to operate efficiently. This data enables the creation of comprehensive user profiles, encompassing interests, preferences, and behavioral trends. While this can enhance the relevance of search outcomes, it also opens the door to user targeting for advertising and potential manipulation.

Deepfake Technology: Another possible concern is the utilization of deepfake technology, a form of AI employed to produce convincing counterfeit visuals, videos, and audio clips that give the impression of authenticity. Deepfake technology operates by employing machine learning algorithms to assess and modify various data, including images, videos, and audio recordings, to generate fresh, artificial content (Ienca, 2023; Westerlund, 2019). Deepfakes are predominantly directed towards social media platforms, where it's effortless for rumors, misinformation, and conspiracies to proliferate, given that individuals often follow popular opinions. Simultaneously, a persistent 'infopocalypse' is causing individuals to believe in the reliability of information only if it originates from their social circles, encompassing family, close friends, or acquaintances, and aligns with their preexisting beliefs. In reality, a considerable number of individuals are receptive to content that validates their established perspectives, even when they harbor suspicions about its authenticity (Westerlund, 2019). A significant risk concerning deepfake incidents in India during April 2018 was highlighted. In this instance, a video quickly circulated on WhatsApp, a widely used mobile instant messaging platform. The video, appearing to be from a surveillance camera, depicted two individuals on a motorcycle purportedly kidnapping a young child before swiftly escaping. This video, falsely portraying a kidnapping, triggered extensive bewilderment and fear, leading to an eight-week period of mob violence that tragically claimed the lives of at least nine innocent individuals (Vaccari and Chadwick, 2020).

Lethal Autonomous Weapons Systems (LAWS): In this context, the term 'autonomous' pertains to any result generated by a machine or software without human involvement. LAWS are a distinctive category of weapon systems that employ sensor arrays and computer algorithms to detect and attack a target without direct human intervention in the system's operation. The delegation of decision-making to automated weapons inevitably raises various concerns, including accountability, appropriateness, potential unintended escalation due to imminent accidents, ethical quandaries, and additional aspects (Pedron and J.d.A. da Cruz, 2020). Employing LAWS could pose significant risks. In addition to getting ready for a future featuring super-intelligent machines (Wogu, 2018), it's important to acknowledge that AI programmed like autonomous weapons (Surber, 2018) to do something dangerous can already present risks. In essence, LAWS have the potential to alter the way humans exert authority over the deployment of force and its aftermath. Moreover, humans might lose the ability to foresee which individuals or entities could become the focus of an assault, or even elucidate the rationale behind a specific target selection made by a LAWS (Marr, 2018; de Ágreda, 2020). It is essential to recognize that AI risk management is a multi-layered and evolving different domains. As AI technology develops and becomes more widespread, it is imperative to proactively identify and address possible risks to guarantee the secure and accountable development and utilization of AI systems.

2.3. AI security management conceptualizing

With the rising integration of AI technology across different fields, safeguarding the AI systems security together with the sensitive information they manage becomes critically important. AI security management involves the adoption of practices and measures aimed at protecting AI systems and the data they process from unauthorized access, breaches, and malicious activities. AI security management encompasses different key aspects like Threat Identification (Kumar and Kumar, 2023), Access control (Song, 2020), Security Awareness and training (Solomon, 2022) and also Privacy (Schiliro et al., 2020). Analyzing possible attack vectors such data breaches, unauthorized access, adversarial attacks, and insider threats is one way to spot potential risks and vulnerabilities that could jeopardize the security of AI systems. Ensuring a secure and responsible deployment of AI involves effectively managing the progress of AI technology while proactively addressing the adverse social, organizational, and individual implications. Here are some considerations when conceptualizing security in AI.

Malicious Use of AI: The remarkable technological advancements in AI have garnered acclaim, presenting enhanced possibilities across various aspects of our daily lives. Despite that, AI and machine learning (ML) are transforming the risk landscape concerning security for individuals, organizations, and nations. The Malicious utilization of AI has the potential to endanger digital security (Schneier, 2015), physical security (Weingart, 1987), and political security (Brundage, et al., 1802; Dhiman and Toshniwal, 2022). International law enforcement entities grapple with a variety of risks linked to the Malevolent Utilization of AI. Interpol and the Center for Al and Robotics at the United Nations Interregional Crime and Justice Research Institute (UNICRI) have raised concerns about "political assaults," notably employing deepfakes, and physical assaults carried out by criminals, such as utilizing combat drones integrated with facial recognition algorithms. AI can be employed to perpetrate a crime directly or subvert another AI system by tampering with the data (Bazarkina and Pashentsev, 2020). However, addressing and mitigating the potential harms stemming from the malicious use of AI is a critical concern in the development and deployment of AI technologies.

Insufficient Security Measures: The progress of technology has been significant across various sectors (Khan, 2020). The impact of AI on cybersecurity has been dual-sided, with positive and negative aspects. AI-driven automation through machine learning algorithms has effectively thwarted attackers from employing traditional attack methods on systems. Demonstrating that machine learning algorithms outperform humans in delivering security, AI integration in cybersecurity is instrumental in error prevention (Ansari, 2022). However, guaranteeing the durability and strength of AI models against evolving adversarial attacks is an ongoing worry. Malicious entities can take advantage of weaknesses in AI algorithms to alter results, potentially resulting in tangible real-life impacts. Additionally, it's vital to prioritize safeguarding privacy and handling data responsibly, particularly given AI's significant data needs. Balancing the extraction of valuable insights with privacy maintenance is a delicate task. Finally, the swift progress of AI technology surpasses existing regulatory frameworks, highlighting the need for flexible and adjustable policies to ensure AI systems comply with ethical and legal guidelines (Di Vaio, 2020).

Furthermore, guaranteeing adherence to privacy regulations and upholding user data privacy constitutes a vital component of AI security management. Incorporating privacy-by-design principles and employing anonymization techniques can aid in safeguarding sensitive personal information. Equally important is the promotion of security alertness among administrators, developers, and users of AI systems through training programs and educational initiatives. This serves to cultivate a security-conscious culture and minimize the potential risks associated with human errors and social engineering attacks. Moreover, Table 1 illustrates the balancing of AI trust, risk, and security with respect to threat types and damages signifies the relationship between these items in the context of probable threats, the resulting damages. By visualizing the relationship between threat types and damages the table depicts the importance of understanding and addressing possible threats, the resulting damages in the context of AI trust, risk, and security. It highlights the need for a comprehensive and proactive method to protect AI systems and their users from harm and maintain the desired balance between these critical elements.

3. AI TRiSM framework

The widespread acceptance and effective integration of AI into different facets of society heavily depend on establishing trust in it. Given its diverse and multifaceted characteristics, it's essential to take into account a multitude of factors when assessing this trust, which is currently absent in existing static models. In healthcare environments, certain concerns pertain to enhancing trust, improving transparency of AI-driven systems, and minimizing bias in medical use cases. A key lesson drawn from these critical and delicate sectors is the significance

Table 1
The balancing of AI trust, risk, and security with respect to threat types and damages.

Aspect	Threat Vector Types	Types of Damages
AI Trust Management	1. **Bias and Discrimination** Dissemination of misleading information and biased narratives to shape negative perceptions of AI's capabilities and intentions.	Destruction of public trust, hindrance to AI adoption, and impeding societal progress by fostering fear, skepticism, and reluctance towards leveraging AI systems.
	2. **Privacy Invasion** Adversarial Attacks utilizing manipulated training data to deceive AI systems.	Erosion of user trust, compromised sensitive data, and potential for discriminatory or harmful decision-making.
AI Risk Management	1. **Society Manipulation** Synchronized AI-driven misinformation campaigns intended at distorting public perceptions and influencing social, political, or economic outcomes.	Dispersion of misleading or fostering social division, and creating an environment susceptible to misinformation through AI-driven manipulation.
	2. **Deepfake Technology:** Fabrication of realistic audiovisual content depicting AI systems making harmful decisions, perpetuating mistrust in AI's reliability	Damaging reputations, and undermining public trust by generating deceptive content that is difficult to distinguish from reality, Discouragement the credibility of AI systems.
	3. **Lethal Autonomous Weapons Systems (LAWS)** Humans might lose the ability to foresee, cyberattacks targeting the communication, control, or decision-making mechanisms LAWS.	misuse, and loss of human oversight, ethical norms, raising significant concerns about the uncontrolled use of AI in warfare.
AI Security Management	1. **Malicious Use of AI** Data theft, or unauthorized access, exploiting vulnerabilities in AI systems.	Breach of sensitive data, compromised system integrity, potential AI model poisoning, resulting in security breaches and loss of trust in AI-powered technologies.
	2. **Insufficient Security Measures** Mistreatment of weak authentication, encryption, or access control in AI systems.	Unauthorized access to sensitive information, and potential misuse of AI systems, leading to compromised privacy and loss of trust in AI technologies.

of incorporating human involvement within AI processes. This can be observed when AI defers a classification decision to a human when uncertain about a specific case. The human-in-the-loop strategy in the medical field shows potential in alleviating bias, enhancing transparency, and building trust in AI-powered systems (Lukyanenko et al., 2022; Rehman, 2022). On the other hand, AI methods for risk management are increasingly extending into novel domains encompassing the examination of extensive document repositories, the automation of repetitive tasks, and the identification of money laundering, which necessitates the analysis of sizable datasets. As AI systems continually advance, making it difficult for risk management frameworks to preserve pace with the latest developments and potential risks. Leveraging individual data for evaluating risks is governed by the General Data Protection Regulation (GDPR) (Kingston, 2017; Mitrou, 2018). Companies employing AI must adhere to this additional responsibility and strategize on aligning with GDPR guidelines. However, if a small or medium-sized enterprise (SME) attempts to create an in-house AI risk assessment system, it would be highly inefficient and too costly for the resources of a single SME to manage (Žigienė et al., 2019). Additionally, Numerous ongoing initiatives and suggestions are dedicated to addressing AI security frameworks, aiming to mitigate security risks that lead to a loss of control over AI application behavior and decision-making due to security challenges. These frameworks can act as a guide to enhance the safety and security features of AI while encouraging its well-organized and safe progression (Jing, 2021).

The existing frameworks for AI Trust, Risk, and Security operate independently and separately, lacking cohesion and alignment (Chauhan and Gullapalli, 2021; Wickramasinghe, et al., 2020; Sobb et al., 2020). These isolated frameworks often do not effectively cooperate or synchronize their actions, resulting in a fragmented strategy for managing AI. The absence of seamless integration and coordination among these essential dimensions creates areas where trust may be established without a complete understanding of the associated risks and security consequences. The AI TRiSM Framework seeks to bridge this divide by presenting a unified approach that brings together AI trust, risk evaluation, and security protocols. It amalgamates key components from individual frameworks, promoting improved collaboration and synergy in these critical aspects of AI governance and administration. This unified approach ensures a comprehensive strategy to tackle challenges related to AI while promoting a stronger and more dependable AI system.

The AI TRISM framework emphasizes the importance of trust, risk, and security considerations throughout the entire life cycle of AI systems, encompassing design, development, deployment, and operation stages. This framework provides a comprehensive approach to manage the risks associated with AI systems. Furthermore, it can assist businesses in formulating and implementing AI strategies that align with their objectives and values. These guidelines cover the operational aspects, safety considerations, and potential ramifications related to AI TRiSM. Their purpose is to offer explicit direction to developers and the AI community on how to effectively deploy secure and innovative platforms based on AI.

It is important for both AI model developers and users need to be reminded the necessity to integrate safeguards into their frameworks and models. This step is crucial for instilling trust in AI systems, mitigating associated risks, and upholding their security. A comprehensive AI TRiSM framework is indispensable for achieving these objectives. According to the current frameworks that organizations should accept include solutions, approaches, and measures that enable model explanation and interpretability (De, 2020), ensure smooth model operations, protect privacy and enhance resistance to adversarial attacks. These measures are aimed at benefiting both the enterprise and its customers. Fig. 1 illustrates the four core principles of AI TRiSM: Model Monitoring, ModelOps, AI Application Security, and Model Privacy (Groombridge, 2022). Together, these pillars establish a comprehensive framework for responsible, trust and secure AI implementation. Here's an overview of the components of the current AI TRISM framework:

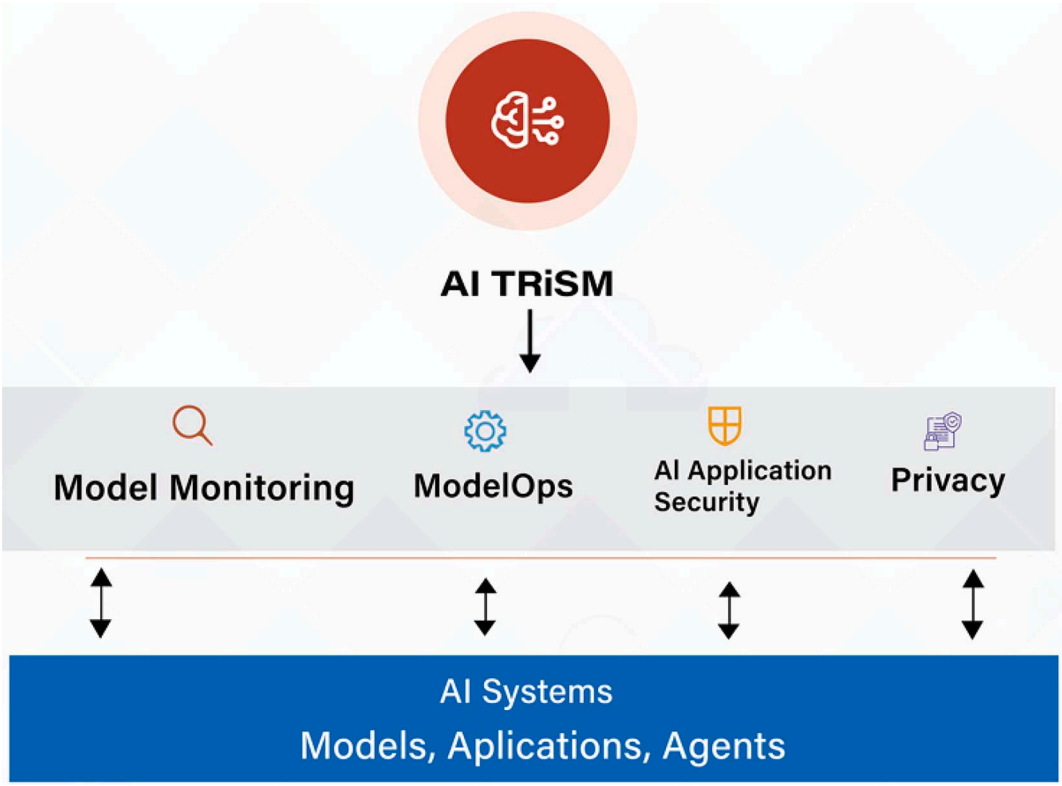


Fig. 1. The four AI TRiSM pillars to deliver managed trust, risk and security for AI systems.

3.1. Model monitoring

One of the significant challenges faced by AI models today is the lack of trust among users, primarily stemming from issues related to transparency and ethics. Many individuals feel uncomfortable interacting with machines instead of human counterparts, to which they are accustomed. This concern becomes more pronounced when the decision-making process within the "black box" of AI models is complex or inscrutable, leaving users without explanations or reassurance. It is crucial to enable the integration of AI into real-world applications in a manner that prioritizes fairness, accountability, and transparency, utilizing governance frameworks. The absence of transparency in AI systems can pose accountability challenges and make it difficult to assign responsibility for their actions. Therefore, it is vital to address these concerns to ensure that AI systems are transparent and accountable (Smith, 2021). While the application of AI in healthcare has great promise for transformative advancements, it also introduces ethical dilemmas that necessitate careful consideration and resolution. In order to mitigate the risks associated with relinquishing control over AI systems, as well as potential issues like human deskilling and widespread surveillance, it is essential to conduct thorough ethical analyses (Gerke et al., 2020; Taddeo, 2019).

However, by implementing model monitoring and explainability it is making sure that AI models are functioning properly and do not add biases. This contributes to the understanding of how AI models operate and reach defensible conclusions and promoting transparency and building trust in the AI system. Fig. 2 shows the illustration and description of interpretability AI TRiSM models monitoring operation's aim to achieve and how it provides transparent, trust and information to the user.

3.2. AI ModelOps

Despite the potential demonstrated by AI in various application areas, its integration into enterprises is still at a nascent stage. One possible explanation for this is the lack of suitable tools and methodologies to facilitate the complete development lifecycle of AI solutions. This encompasses crucial tasks like data preparation, model design and training, application development, quality assurance, deployment, monitoring, feedback, and ensuring reproducibility and auditability throughout the process. It is evident that a more structured and efficient approach is required for the development and management of the lifecycle of AI applications. An essential aspect of ModelOps involves employing a domain-specific language that prioritizes core elements in AI solutions. These encompass datasets, model specifications, trained models, applications, monitoring events, and the algorithms and platforms utilized for data processing, model training, and application deployment (Hummer, et al., 2019). An essential part of an AI TRiSM framework is the ModelOps procedure, which is the operationalization of AI models, involves managing the lifecycle and governance of all AI models as well as responsibility of managing the foundational infrastructure and environment, such as cloud resources, to ensure the optimal performance of the models. Meanwhile, Fig. 3 illustrates the comprehensive ModelOps procedure, encompassing the key stages of model design, deployment, operations, and monitoring. At the beginning, the model design phase comprises collection of requirement and data availability and prioritizing carefully and also data preprocessing techniques to ensure optimal model performance and accuracy. Following design, the deployment phase entails selecting appropriate platforms and infrastructure to seamlessly integrate the model into the desired environment. During operations, the model is actively utilized in real-world scenarios, necessitating continuous performance monitoring, error handling, and fine-tuning for maintaining optimal functionality.

3.3. AI security application

AI security applications employ sophisticated machine learning algorithms and methodologies to promptly identify and address weaknesses, unauthorized access, and harmful actions. These applications have the capability to observe network patterns, assess user actions, and pinpoint irregularities that could signal a breach in security (Jain, et al., 2020; Gopalan et al., 2020). The utilization of AI technology requires huge amount of data, and ensuring the protection of that data is of paramount importance. Within the context of AI TRiSM, data security holds particular significance in heavily regulated sectors such as healthcare (Norori, 2021) and finance (Giudici and Raffinetti, 2023). Furthermore, AI TRiSM, data protection frameworks like synthetic data, differential privacy (Meden, 2023), as well protocols such as Full Homomorphic Encryption (FHE) (Kadykov et al., 2021) and Secure Multi-Party Computation (SMPC) are applied which are essential for proting

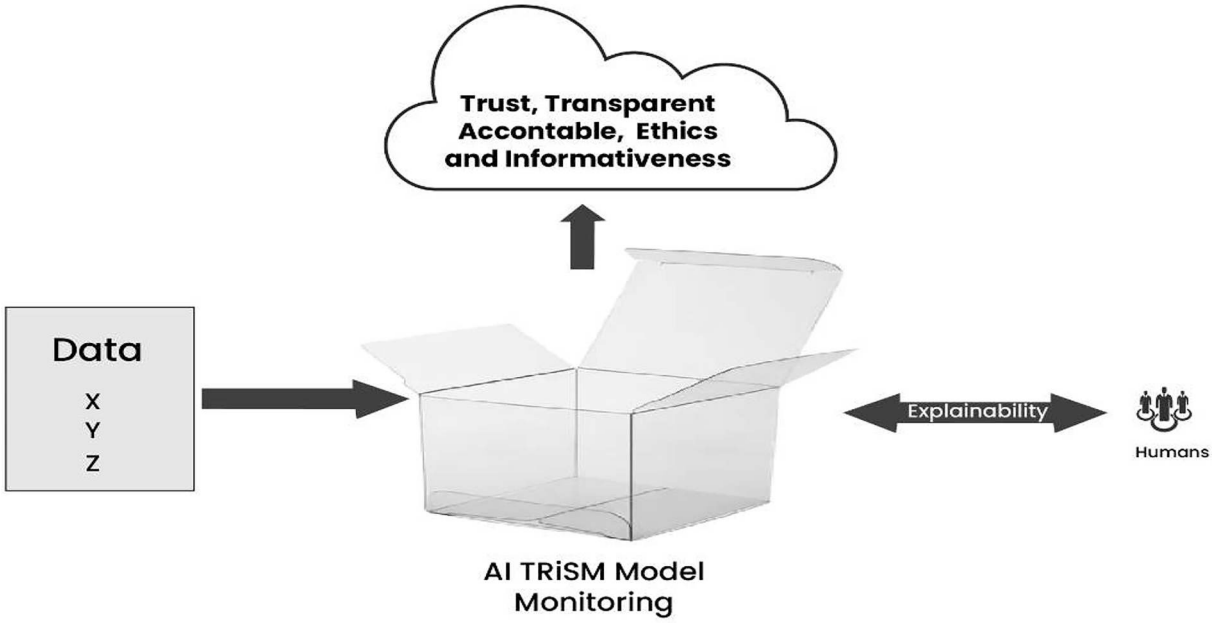


Fig. 2. AI TRiSM models monitoring functionality.

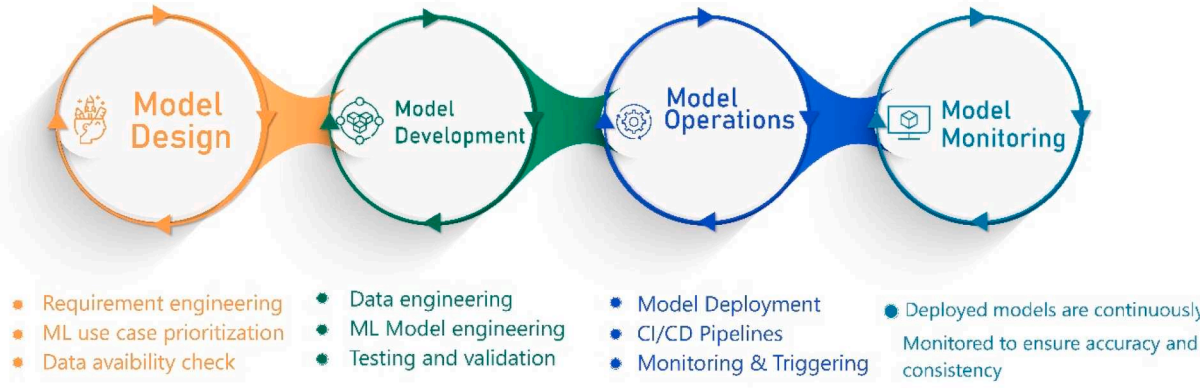


Fig. 3. Visually delineates how these interconnected properties contribute to an effective ModelOps lifecycle.

the great quantities of data required for AI technology, and ensuring trust, security, and mitigating risks in AI systems. Sensitive or confidential information can be hidden while maintaining the general statistical properties of the original data by employing synthetic data for data protection implementation. This encourages data analysis and sharing without running the danger of disclosing private information.

The implementation of SMPC in data privacy maintains the confidentiality of sensitive data by avoiding plaintext sharing among parties and using cryptographic techniques to protect data. It becomes challenging for individuals with malicious intent to manipulate or intercept data. Delivering inference services for the trained tree-based model presents numerous obstacles. In practical scenarios, the computational or communication capacity of the model owner might be restricted and inadequate for managing extensive queries. Entrusting the tree-based model directly to a high-performance server, even though it's a fundamental digital asset for enterprises or institutions, would greatly jeopardize its privacy and interests (Zhao, 2023). This approach ensures that data accuracy is preserved while privacy is maintained.

FHE allows data to be processed without ever being exposed in plaintext, providing a high level of data protection. FHE can be used for various applications, including data analysis and machine learning, where sensitive data needs to be processed while maintaining its confidentiality (Tonyali, 2018). It allows multiple parties to perform computations jointly without sharing their private data. AI security application ensures the protection and security of models from cyber threats, enabling organizations to utilize TRiSM's framework to establish security protocols and implement measures that safeguard against unauthorized access or tampering.

3.4. Model privacy

Data privacy involve confirming that AI systems gather, store, and process personal and sensitive data while adhering to privacy regulations and established best practices. This entails gaining proper consent, implementing data anonymization techniques, and employing secure data handling practices to protect individuals' privacy rights. Privacy safeguards are necessary for the data used in training or testing AI models. Differential privacy techniques can be employed to safeguard the anonymity of individuals within datasets. This approach introduces noise or randomization to the dataset, thereby preventing the identification of individuals (Hassan et al., 2019). Privacy corresponds to system concepts such as confidentiality, concealment, individual entitlements, and possession (Habbal, 2017). The core idea is to increase the complexity for potential intruders to determine the presence of specific individuals in a dataset, while still enabling accurate analysis of the data. By integrating privacy considerations into AI TRiSM, organizations can bolster user trust, adhere to privacy regulations, and mitigate potential privacy risks associated with AI systems. Privacy protection is vital for upholding ethical practices and ensuring responsible and trustworthy deployments of AI. The AI TRiSM framework aids organizations in establishing guidelines and protocols for the ethical collection, storage, and utilization of data, which is particularly critical in fields like healthcare, where various AI models are employed to process sensitive patient information.

Moreover, Table 2 showcases a comparison of key parameters of the trust, risk and security in AI systems challenges and potential improvements before and after the implementation of the AI TRiSM Framework. The table focuses the key aspect like model monitoring, AI modelOps, AI security application and model privacy.

Table 2
AI systems challenges and potential improvements before and after the implementation of AI TRiSM Framework.

Aspect	Before and After AI TRiSM Framework	
	Challenges	Potential Improvements
Model Monitoring	Limited transparency Uncertain model behavior (Vollmer, et al., 1812).	Improved explainability through model introspection. Validation against expected behavior.
	Lack of accountability for AI system behavior	Clear governance frameworks for increased accountability,
	Potential biases and discriminatory outcomes (Smith, 2021).	Fairness-aware algorithms and bias detection/mitigation techniques
	Limited security measures for model monitoring (Tan, 2021).	Enhanced security protocols and measures for model monitoring.
AI ModelOps	Limited of suitable tools and AI system development methodologies (Hummer, et al., 2019).	AI system lifecycle managing, Governance of all AI model infrastructure and environment. Robust AI system design Methodology.
	Insufficient consideration of ethical implications.	Ethical guidelines and regulatory frameworks to address ethical concerns.
AI Security Application	Limited protection against AI model poisoning and adversarial attacks (Eluwole and Akande, 2022).	security protocols and implement measures that safeguard against unauthorized access or tampering.
	Physical harm, accidents, or unintended consequences (Norori, 2021).	Robust system design, fail-safe mechanisms, rigorous testing, and human oversight
Model Privacy	Limited data privacy and handling AI models sensitive information of individuals (Hassan et al., 2019).	Enhanced trust with clear communication about privacy measures.
	Increased risk of personal data misuse, unauthorized access, and privacy violations (Habbal, 2017).	Adhering regulations, and mitigate potential privacy risks associated with AI systems.

4. Automation of trust risk and security management

Trust, risk, and security management are critical characteristics of operational organization, as they comprise preserving sensitive information, protecting assets, and confirming compliance with regulations and industry standards. Traditionally, these responsibilities have been manually made by expert teams, which can be time-consuming, causing lot of errors, and problematic to scale. With the progression of technology, automation has become a valuable tool in addressing AI based system challenges. In the following sections will elaborate how AI TRiSM will automate and manage trust, risk and security of AI systems.

4.1. Automation of trust management

Recent research has revealed that AI-based systems are susceptible to unreliability, untrustworthiness (González-Gonzalo, 2022) and the presence of bias in AI systems can undermine trust among individuals and organizational clients, leading to discriminatory outcomes, economic disturbances, and the potential for misuse by malicious entities. This is primarily due to the fact that the presence of bias within an AI system is influenced by the bias found in the training data it relies on. Consequently, if the training data itself contains bias, the resulting AI system will also exhibit bias. This can lead to discriminatory decisions that unfairly affect individuals based on factors like race, gender, or socioeconomic status. Additionally, AI technology has the capability to generate realistic counterfeit images and videos, which can be exploited to spread misinformation or manipulate public opinion. Furthermore, AI can be utilized to create highly sophisticated phishing attacks, deceiving individuals into divulging sensitive information or clicking on malicious links (Desolda, 2023).

To establish trust, it is crucial to automate the management of trust within AI trust and risk framework, ensuring that AI systems are governed in a manner that prioritizes transparency, accountability, and fairness. This necessitates designing and implementing AI systems in a way that mitigates bias and discrimination while adhering to ethical principles and standards. Moreover, AI systems must be reliable, resilient, and capable of operating securely and safely. Managing trust involves ensuring that these requirements are fulfilled to instill confidence and trust in AI technology (Kong, 2023).

In order to avoid bias and discrimination through the automated management of trust in AI TRiSM, it is important to train AI systems on datasets that are diverse and representative. This means incorporating data from various sources and populations to prevent biases stemming from underrepresented or skewed data. Additionally, involving a range of stakeholders, including domain experts and ethicists, in the design and development of AI systems can help identify and address potential biases and ensure protection against discrimination.

4.2. Automation of risk and security management

Developing and deploying AI systems without taking ethical considerations into account can give rise to a range of potential risks and consequences. Some of the most significant risks are observed judicial evaluation systems (Pechegin, 2022), facial autonomous vehicles systems (Kumar, 2022) and recognition systems (Pantic and Rothkrantz, 2000). For instance, facial recognition systems often exhibit higher error rates for individuals with darker skin tones or who are female. Similarly, hiring algorithms have been found to demonstrate discriminatory behavior against specific groups based on factors like age, ethnicity, or gender. Moreover, Automation enhances the efficiency of risk assessment, allowing organizations to precisely measure the risks linked to AI implementations. Through the utilization of AI algorithms, organizations can pinpoint potential aspects of worry, including privacy concerns, model biases, or software vulnerabilities, and categorize them according to their seriousness and likely consequences. Additionally, incorporating automation into the management of risks and security facilitates the prompt and effective deployment of security protocols, reinforcing the overall robustness of AI systems within a constantly changing threat environment.

4.3. Features of automation of trust, risk and security management

The automation of overseeing trust, risk, and security in AI TRiSM presents numerous features, transforming how organizations address safety and dependability. Firstly, automation boosts effectiveness and accuracy in assessing the reliability and possible risks linked to AI systems. Sophisticated algorithms can consistently observe and scrutinize extensive data, quickly detecting any irregularities or deviations from anticipated behavior. This immediate monitoring enables prompt reactions and mitigation of security risks, reducing potential harm. Moreover, automation simplifies evaluating trust by utilizing advanced trust models and measurements. It has the capability to monitor performance, dependability, and compliance with predetermined ethical and regulatory criteria (Zhang, 2019; Jacobsson et al., 2016). This comprehensive and precise trust assessment offers organizations a proactive strategy, enabling improved decision-making and strategic foresight. Ultimately, this approach cultivates increased trust and acceptance of AI technologies both within the organization and across the broader community. Furthermore, automation trust, risk, and security in AI TRiSM considerably lessens the workload on human resources by automating repetitive and time-intensive security and risk management responsibilities. This enables security specialists and risk managers to concentrate on more intricate and strategic aspects of security, like devising novel strategies to combat emerging threats and enhancing the system's overall resilience (Chen, 2021). Ultimately, this reassignment of tasks results in a more effective and productive security and risk management procedure, enhancing the organization's overall security stance and facilitating quicker and well-informed decisions in a swiftly evolving technological environment.

However, it is important to note that when AI systems does not provide responsibility and transparency, it can lead to reduced adoption and increased distrust among users. Fig. 4 illustrates the features provided by AI TRiSM framework to deliver trust, security and transparency in AI systems.

5. AI TRiSM applications

The following secion will provide five illustrative scenarios that highlight the effectiveness and possibilities offered by AI TRiSM. These instances demonstrate how organizations leverage AI TRiSM to foster innovation, enhance results, and generate value for both businesses and society.

Use case 1: Fair, Financial Transparent, and Accountable AI Models.

An example of AI TRiSM in action is demonstrated by the Danish Business Authority (DBA). The DBA aligns its ethical principles with concrete actions, conducts fairness tests to validate model predictions, and establishes a robust monitoring structure. They have successfully implemented and managed 16 AI models that oversee financial transactions valued in billions of euros. This strategy not only assisted DBA in ensuring the morality of its AI models, but it also aided in increasing customer and stakeholder trust.

Use case 2: cause-and-effect relationships interpretable AI models generator.

Abzu, a Danish startup, has developed an AI product that constructs mathematically interpretable models for identifying cause-and-effect relationships Abzu (Sebastian and Peter, 2022). Their clients leverage these models to validate outcomes with efficiency, resulting in the successful development of effective drugs for breast cancer treatment. Abzu's product has the capability to analyze vast volumes of data and uncover patterns and connections that may not be readily discernible to humans. This enables their clients to make well-informed decisions and advance the development of enhanced treatments for patients. The

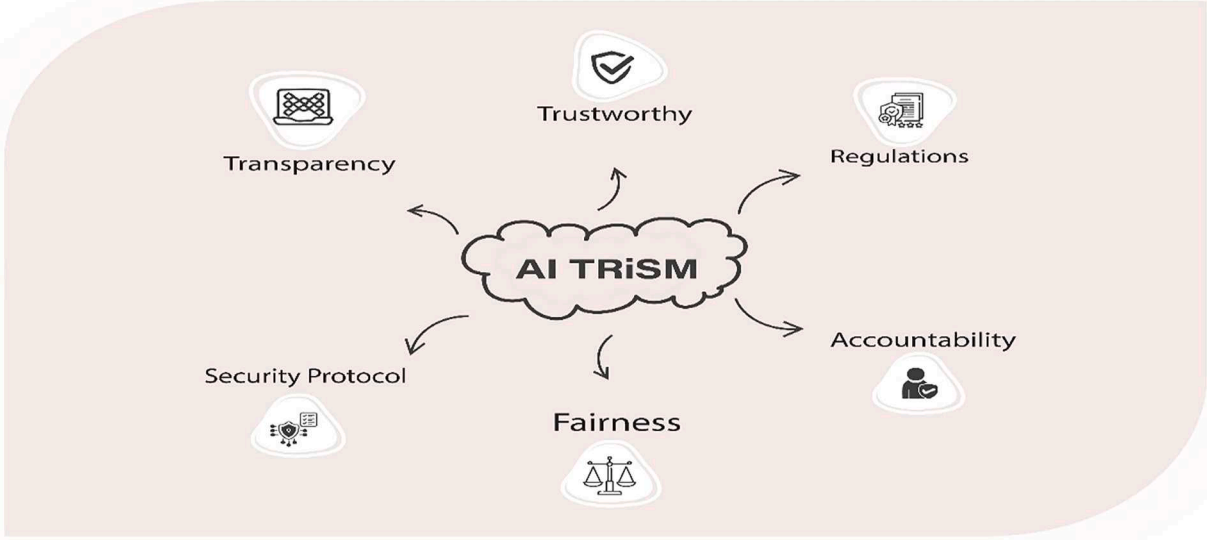


Fig. 4. The Features provided by AI TRiSM.

transparent models produced by Abzu's AI product play a role in fostering trust between patients and healthcare providers. These models offer a transparent view of the AI's decision-making process, providing a clear comprehension of how it reaches its conclusions.

Case 3. *The concept of sharing medical data in the context of smart healthcare.*

AI models smart healthcare assisting professionals in a multitude of responsibilities, ranging from managing administrative tasks and documenting clinical information to reaching out to patients. Additionally, they offer specialized assistance in areas like analyzing medical images, automating medical devices, and monitoring patient health (Bohr and Memarzadeh, 2020). AI TRiSM may help safeguard patient data by implementing robust security measures, access controls, and encryption techniques. It ensures that sensitive health information is stored, transmitted, and processed securely, reducing the risk of data breaches and unauthorized access.

Moreover, ATRSM can involve validating and auditing AI models used in healthcare applications. It ensures that the AI algorithms are accurate, reliable, and compliant with regulations and standards. This validation process minimizes the risk of biased or incorrect predictions, particularly in critical areas like diagnostics or treatment planning.

Case 4. *Secure smart cities framework*

AI applications play a significant role in transforming smart cities by optimizing resource allocation, enhancing efficiency, and improving the quality of life for citizens. The concept of a smart city stands as a highly favored implementation, embodying a wide array of pervasive services collaborating and orchestrating their efforts to enhance the living standards and overall life experience for urban inhabitants (Habbal et al., 2019). As cities become more interconnected and reliant on AI technologies, ensuring the trustworthiness, mitigating risks, and ensuring the security of these systems become paramount. Smart cities produce substantial volumes of data originating from diverse channels, including sensors, cameras, and interactions among citizens. AI systems must be designed to handle this data securely and protect individual privacy (Veselov, 2021; Kamruzzaman et al., 2022).

To ensure the privacy and safeguarding of data, methods like data anonymization, encryption, and access controls can be utilized through AI TRiSM framework. By addressing privacy concerns, managing risks, and implementing robust security measures, cities can build trust among citizens and stakeholders, fostering a safe and secure environment for their residents, so AI TRisM framework is fundamental in ensuring the successful deployment and operation of AI systems applied in smart cities.

Case 5. *AI TRiSM in Metaverse: Securing the Virtual Frontier*

The use of AI in the metaverse has the potential impact and improve various features of virtual space like virtual economy and marketplace and also education (Cheng, 2022; Huynh-The, 2023) content generation (Lin, 2023), AI can assist in generating vast amounts of virtual content within the metaverse. Through methods such as generative adversarial networks (GANs) and procedural generation, AI algorithms have the ability to independently produce virtual assets such as landscapes, buildings, objects, and other elements within the metaverse. This capability aids in creation the metaverse with diverse and realistic content, minimizing the need for manual content creation (Yang, 2022; Goodfellow, 2020).

As the metaverse expands and becomes more complex, ensuring the security and trustworthiness becomes paramount. AI TRiSM in the metaverse refers to the application of trust, risk, and security management framework in the context of AI within the virtual space. Through the integration of the AI TRiSM framework into the metaverse, developers and platform providers have the ability to establish a secure, trustworthy, and ethically sound virtual environment. It assists to protect user data, mitigate risks, and build self-confidence in the AI systems that power the metaverse experience.

6. Challenges of AI TRiSM

In spite of the current existence of the AI TRiSM framework and its important, there are challenges that must be addressed before it is widely implemented in the near future. The reason behind this is that AI TRiSM framework its self requires reliability and significant advancement in AI systems. In the following sections, we examine a number of these obstacles that must be undertaken to ensure successful implementation, along with some potential paths for future directions. These challenges encompass numerous aspects such as the ever-changing threat landscape, adversarial attacks, compliance with regulations and skilled expertise gap.

6.1. Monitoring AI models

Monitoring AI models is a crucial aspect of ensuring trust, risk and security in their deployment and may presents several challenges that organizations developers need to address effectively. These may include,

but not limited to, data drift because AI models depend on the training data they receive, and if the characteristics of the incoming data change over time, it can result in a phenomenon known as data drift. Continuous monitoring and timely model updates are necessary to detect and resolve data drift effectively (Zhao, 2021). Moreover, bias and fairness monitoring in AI models are susceptible to biases, which have the potential to result in unjust or discriminatory results. To ensure fair treatment of individuals, it is crucial to meticulously choose suitable fairness metrics and continuously evaluate the model's performance (Liang, 2022). To tackle these challenges, it requires a combination of methods, including ongoing data collection, monitoring pipelines, versioning of models, automated alerts, and establishing feedback loops involving human experts (Asan et al., 2020; Liyanage, 2023). Applying robust monitoring practices ensures the trustworthiness, risk mitigation, and security of AI models throughout their lifecycle.

6.2. Regulatory compliance

According to existing Controlling bodies and governments, the connection between AI TRiSM and like the GDPR (General Data Protection Regulation) (van Mil and Quintais, 2022), which is of significant in policy spheres, and the corporate segment, is another important aspect to consider. Navigating through regulatory landscapes and meeting necessary requirements can be a complex task for organizations as they struggle to obey to applicable regulations and compliance frameworks. These include data protection laws of GDPR, industry-specific regulations in sectors such as healthcare or finance, and ethical guidelines. It is crucial for organizations to ensure that their AI systems comply with these regulations and frameworks. AI TRiSM serves as a framework designed to handle the risks linked to AI systems and facilitate governance, reliability, fairness, efficacy, and privacy. Nevertheless, the establishment of AI model governance and trustworthiness is a complex task due to the lack of agreement regarding the appropriate definition of controllers.

6.3. Adversarial attacks

Adversarial attacks often aim to exploit weaknesses in AI models. Ensuring the robustness of AI models against such attacks is a significant challenge. Adversarial examples can be crafted to mislead the model's predictions or bypass security measures, potentially leading to false positives or negatives in risk detection. Another challenge for Adversaries can manipulate training data or inject malicious samples into the dataset used to train AI models. This can result in the model learning incorrect patterns or making biased decisions, compromising the integrity and accuracy of the AI TRiSM system. The tactics used by adversaries are continuously changing and growing more sophisticated. They may employ advanced algorithms, use gradient-based optimization methods, or explore novel attack strategies. Keeping up with the evolving landscape of adversarial attacks requires continuous research, monitoring, and adaptation. Addressing these challenges requires a multi-faceted approach, including robust model training, ongoing monitoring for adversarial behavior, implementing defense mechanisms like anomaly detection and input validation, conducting regular vulnerability assessments and penetration testing (Karim, 2022). Organizations must continually update and improve their AI TRiSM systems to stay ahead of adversarial threats and ensure the security and trustworthiness of their AI deployments.

6.4. Skill gap and expertise

Recent studies demonstrate a significant surge in the need for AI expertise. The demand for roles necessitating AI skills has grown by a factor of ten between 2010 and 2019 and has quadrupled in terms of the overall job market share. Although a substantial portion of the demand for AI skills is concentrated in the Information Technology, Professional Services, Finance, and Manufacturing sectors (Alekseeva, 2021). Moreover, for successful implementation of AI TRiSM, it is essential to form a collaborative cross-functional team consisting of experts from various disciplines such as legal, security, compliance, computer science, and data analytics. These professionals collectively contribute their skills and knowledge to effectively manage and address the challenges associated with AI TRiSM. In order to maximize outcomes, it is advisable to form a dedicated team, or alternatively, a task force if necessary. It is of utmost importance to ensure that each AI project has adequate representation from the business side. This ensures that the team possesses the necessary expertise and perspectives to effectively drive the project towards success, so the scarcity of individuals with these combined skills can present challenges in establishing effective AI TRiSM practices.

6.5. Rapidly evolving threat landscape

The domain of AI TRiSM operates within a dynamic and ever-changing environment of potential threats. Constantly, new vulnerabilities, attack methods, and privacy concerns emerge, posing ongoing challenges for organizations. Staying well-informed about the latest security practices and effectively addressing emerging risks becomes a continuous task. To tackle the challenges presented by this rapidly evolving threat landscape, organizations must adopt a proactive and adaptable approach to cybersecurity. This entails staying updated on the most recent threats and vulnerabilities, implementing strong security measures, regularly assessing risks, educating employees about security best practices, and fostering a culture of cybersecurity awareness and attentiveness (Jang-Jaccard and Nepal, 2014). Collaboration between organizations, sharing information, and investing in research and development are also vital in effectively countering the ever-evolving threat landscape.

7. Transformation features

In addition to the obstacles mentioned above, there are numerous transformation features and characteristics that will make AI TRiSM framework very attractive in future and may include, providing enterprise risk management and security orchestration in AI systems, offering a comprehensive unified approach for responsible and sustainable development and deployment of AI systems, adaptive strategies to address challenges, market coalition and guaranteeing adherence to regulations and ethical concerns. In the following sections we discussed in detail the AI TRiSM market directions that are expected to progress in the future.

7.1. Elimination fragmentation

In September 2023, The European Commission claims that the EU is well-placed to promote artificial intelligence that is focused on humanity, sustainability, security, inclusivity, and trustworthiness. The report anticipates the emergence of new generations of integrated features over time, ultimately leading to the availability of comprehensive governance or framework AI systems (Francisco and Linnér, 2023).

By tackling a range of challenges and concerns surrounding trust, risk, and security in the realm of AI, AI TRiSM plays a crucial role in mitigating the fragmentation of AI systems. This is because AI providers typically do not offer a comprehensive set of features to effectively and consistently manage trust, risk, and security in AI. Consequently, users are compelled to select from a variety of suppliers specializing in different categories of AI TRiSM solutions to meet their specific requirements.

7.2. Market coalition

AI TRiSM also brings the potential for enhancing and coordinating AI systems by incorporating alerts and remediation processes of ModelOps.

These aspects will be seamlessly integrated into the existing enterprise risk management and security orchestration systems, offering a comprehensive and unified approach. To accomplish this, the platform administration of ModelOps will incorporate the enterprise's utilization of third-party models. Additionally, alerts for adversarial attacks and corresponding corrective measures will be integrated into the current Security Orchestration systems (Sethi, 2021; Zhang, 2022). The integration of AI TRiSM into the core offerings and operations will occur through a collaborative coalition of platform-neutral vendors in the market of ModelOps. These vendors aim to provide organizations with transparent, fair, and secure solutions for operationalizing models. By utilizing this approach, organizations can confidently implement AI models while efficiently mitigating risks and guaranteeing adherence to regulations and ethical concerns. As a result, the number of vendors exclusively dedicated to ModelOps is expected to decrease, as integrated platforms encompass these capabilities. These integrated platforms will coexist with advanced solutions that enhance composite AI and generative AI functionalities. The methods utilized for safeguarding data outside of AI applications will continue to evolve to address data protection concerns pertaining specifically to AI model data.

7.3. AI TRiSM adaptivity

According to recent studies, organizations that prioritize AI transparency, trust, and security in their operations are projected to witness a substantial 50 % improvement in the adoption, successful attainment of business objectives, and user acceptance of their AI models by the year 2026. The predictions also indicate that by 2028, AI-driven machines will account for approximately 20 % of the global workforce and contribute to 40 % of overall economic productivity (Groombridge, 2022).

In broad terms, the AI TRiSM framework will provide support for a system of checks and promote a high level of transparency in documentation. A strong documentation structure will be implemented, with a particular focus on AI training data, to ensure trustworthiness and assist in conducting technical audits in case of any issues. The documentation system will have automated functionalities and the ability to identify incomplete, missing, or abnormal records. It is crucial for the documentation system to preserve reliability and user-friendliness to effectively support AI TRiSM and enable the utilization of AI technology. Lastly, by enabling non-technical users to understand the data gathering process and the decision-making system, businesses can enhance AI accountability and transparency.

8. Discussions and future research directions

Our review highlight's comprehensive structure that integrates trust, risk assessment, and security management within the AI landscape, promoting a holistic strategy to tackle the increasing concerns related to AI implementation. This AI TRiSM framework we primarily categorized into four fundamental pillars that can foster trust among its AI developers and users base while leveraging the forthcoming advancements in AI technologies and also addressing unique security challenges posed by AI, such as adversarial attacks, model explainability, and privacy preservation. The framework is intended to provide a structured approach, offering guidance and strategies to establish trust, mitigate risks, and enhance security in AI applications. Moreover, our research also demonstrates real-world applications of the proposed frameworks and methodologies in diverse domains to showcase innovative applications of AI TRiSM in areas like healthcare, finance, smart cities, and metaverse to highlight the effectiveness and possibilities of trust enhancements, risk mitigations and privacy protection offered by AI TRiSM.

To advance knowledge and technology in this field, numerous potential directions for future research should be examined. In the realm of AI TRiSM, there are abundant upcoming opportunities that demand attention and are anticipated to influence its trajectory, including the Internet of Things (IoT), Quantum-Safe AI TRiSM, Federated Learning and Edge Computing for Security and other domains. Here are a few specific aspects that require exploration.

8.1. IoT security and privacy

The Internet of Things (IoT) comprises gadgets that produce, analyze, and share extensive quantities of data related to security, safety, and privacy-sensitive details, making them attractive to diverse cyber threats (Dorri, 2017; Askar, 2023). AI TRiSM is set to have a crucial impact on enhancing the security of the Internet of Things (IoT). As the IoT network grows, involving various interconnected devices and systems, prioritizing trust and security becomes crucial. AI TRiSM framework utilizes AI to enhance reliability by utilizing machine learning algorithms to recognize typical patterns in device behavior and promptly identify deviations that could suggest potential security risks. This proactive strategy facilitates timely evaluation of risks and their management, contributing to a stronger and more adaptable IoT infrastructure. Moreover, AI has the potential to automate incident response protocols, guaranteeing swift and effective actions in the face of a security breach. By promoting a trustworthy environment through inclusive security measures, AI TRiSM aims to inspire trust among users, manufacturers, and stakeholders within the IoT ecosystem. This multifaceted strategy, which merges AI and security oversight, is critical in constructing a durable and dependable IoT setting, imperative for the smooth assimilation and realization of the potential advantages offered by IoT technology.

8.2. Quantum-Safe in AI TRiSM

Quantum computing technology presents fundamentally distinct approaches to computational challenges, allowing for superior problem-solving efficiency compared to conventional classical computations (Rieffel and Polak, 2000; Gyongyosi and Imre, 2019). AI TRiSM, within the realm of Quantum computing will addresses the changing landscape of AI merging with quantum computing advancements. The progress in quantum computing presents a notable risk to conventional cryptographic systems, making existing security measures potentially ineffective. AI TRiSM is dedicated to securing and fostering trust in quantum computing by preemptively managing risks associated with quantum technology. Moreover, AI TRiSM in the Quantum-Safe context prioritizes evaluating risks and implementing strategies to minimize them, taking into account the possible weaknesses that may arise from quantum computing. By promoting a culture of ongoing vigilance, adjustment, and effective response, Quantum-Safe AI TRiSM guarantees the resilience and reliability of quantum technologies amid the changing threat environment presented by quantum computing.

8.3. AI TRiSM in Federated learning

AI TRiSM is pivotal in the realm of Federated Learning, an approach to machine learning that is decentralized, involving local model training on edge devices or servers. Only aggregated updates are then shared with a central server. AI TRiSM frameworks will play a crucial role in establishing trust by incorporating strong security measures like encryption, secure aggregation protocols, and access controls. Safeguarding the privacy and confidentiality of sensitive data is a cornerstone for building trust among participants, fostering broader acceptance of Federated Learning across diverse sectors such as healthcare, finance, and smart devices. Nonetheless, despite the advantages of Federated Learning (Zhang, 2021; Fedorchenko et al., 2022), it brings certain risks. Effectively handling these risks within the AI TRiSM framework entails recognizing possible threats, weaknesses, and consequences connected to the federated learning setup. Moreover, AI TRiSM should incorporate systems for overseeing and examining

federated learning procedures to guarantee adherence to regulatory guidelines and industry benchmarks. This will fortify the general security readiness and reliability of the federated learning setting.

8.4. Multidisciplinary research workforce

In our study, researchers emphasize the necessity of interdisciplinary collaboration, bringing together experts from AI, cybersecurity, ethics, law, and policy domains to devise holistic strategies for AI TRiSM. Furthermore, there is a call to delve deeper into specific application domains, such as healthcare, finance, and autonomous systems, to tailor AI TRiSM frameworks to the unique challenges and requirements of each sector. Within the AI TRiSM resource scheduling domain are anticipated to gain increased focus in the future. Specifically, some of the challenges and transformation features previously discussed are advocated for and proposed for integration into resource scheduling within the AI TRiSM context.

- Monitoring AI models for bias and fairness is sensitive to biases that can lead to unfair or discriminatory outcomes. Confirming equitable treatment demands careful selection of appropriate fairness metrics and consistent evaluation of model performance. Addressing these obstacles necessitates a blend of strategies, encompassing continuous data collection, monitoring pipelines, model versioning, automated alerts, and fostering feedback loops with human experts. The implementation of robust monitoring practices guarantees the reliability, risk management, and security of AI models over their lifecycle.
- Exploring the unique intricacies of AI trust, risk, and security within specific applications is imperative. Customized research should investigate how these elements manifest in diverse sectors like healthcare, finance, transportation, and more. Grasping the distinct challenges and needs of each sector will enable the creation of targeted solutions and recommendations to guarantee reliable AI implementation, while minimizing risks and bolstering security.
- To Mitigation Adversarial Attack, potent AI algorithms and methods to counter adversarial attacks, ensuring the durability of AI systems against both deliberate and inadvertent manipulation or misuse should be Investigated. The strategies employed by adversaries are in a constant state of evolution and growing in complexity. They might utilize advanced algorithms, employ optimization methods based on gradients, or experiment with innovative attack approaches. Staying abreast of the changing landscape of adversarial attacks necessitates continual research, vigilance, and adjustment. Tackling these hurdles requires a comprehensive strategy, including rigorous model training, ongoing vigilance for adversarial activities, implementation of defensive measures like anomaly detection and input validation, regular vulnerability evaluations, and penetration testing.
- Given the constantly changing nature of AI technologies and the evolving threats they face, it's essential for continuous research to consistently assess and enhance the suggested frameworks and approaches. Subsequent investigations should prioritize real-time adjustment and perpetual enhancement, incorporating AI TRiSM principles into the AI systems' development cycle. Moreover, collaborative efforts involving interdisciplinary teams comprising researchers, practitioners, policymakers, and ethicists are vital to cultivate a comprehensive grasp of AI TRiSM and propel its seamless integration into society.

Looking ahead, future research directions should prioritize developing AI TRiSM frameworks that are adaptive and scalable to accommodate evolving AI technologies and applications. Continuous research on explainability, fairness, and bias detection in AI systems is essential to ensure equitable outcomes and enhance trust. Moreover, interdisciplinary collaborations involving experts from AI, ethics, law, cybersecurity, and psychology are vital to address the complex and multifaceted challenges associated with AI TRiSM. Finally, there is a need for public education and awareness initiatives aimed at enlightening users about the advantages, potential risks, and optimal approaches linked to AI, enabling them to make informed decisions and participate in shaping AI policies and regulations.

9. Conclusion

AI TRiSM is a critical area that requires attention to ensure the responsible and secure deployment of AI systems. In this comprehensive study, we have conducted an extensive examination of the utilization of AI TRiSM, which is a quite new. The significance of AI systems has grown as society increasingly depends on it to carry out complex tasks in the contemporary world. AI TRiSM plays a vital role for organizations in ensuring the proper regulation to deploy AI models and effective management of potential risks. By embracing AI TRiSM, organizations can acquire valuable understanding of the design, development, and distribution of AI models, enabling them to efficiently monitor and mitigate risks while maintaining reliability and credibility. In our study, we have emphasized the importance of trust, risks, and security in existing AI-driven systems. We have provided an extensive review of prior research of AI system trust and security issues and explored the advantages of integrating AI TRiSM. Additionally, we have projected potential challenges that may face by current AI TRiSM framework in different dimensions. This study serves as a valuable resource for researchers seeking a comprehensive understanding of the current frameworks, strategies, challenges, and potential future directions in the realm of AI TRiSM. However, certain barriers to the implementation of AI TRiSM in AI systems still remain. Future research activities will be focused on addressing these issues while leveraging the latest technologies.

CRediT authorship contribution statement

Adib Habbal: Conceptualization, Supervision, Writing – original draft, Writing – review & editing, Investigation. **Mohamed Khalif Ali:** Conceptualization, Writing – original draft, Writing – review & editing, Visualization. **Mustafa Ali Abuzaraida:** Conceptualization, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

References

Kim, J., et al. (2022). Perceived credibility of an AI instructor in online education: The role of social presence and voice features. *Computers in Human Behavior, 136*, Article 107383.

Sohn, K., & Kwon, O. (2020). Technology acceptance theories and factors influencing artificial Intelligence-based intelligent products. *Telematics and Informatics, 47*, Article 101324.

Abuzaraida, M. A., Elmehrek, M., & Elsomadi, E. (2021). Online handwriting Arabic recognition system using k-nearest neighbors classifier and DCT features. *International Journal of Electrical and Computer Engineering, 11*(4), 3584.

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals, 14*(2), 627–660.

Kim, J., Kang, S., & Bae, J. (2022). The effects of customer consumption goals on artificial intelligence driven recommendation agents: Evidence from Stitch Fix. *International Journal of Advertising, 41*(6), 997–1016.

Nikitas, A., et al. (2020). Artificial intelligence, transport and the smart city: Definitions and dimensions of a new mobility era. *Sustainability, 12*(7), 2789.

Chen, L., Chen, P., & Lin, Z. (2020). Artificial intelligence in education: A review. *Ieee Access, 8*, 75264–75278.

Ma, Y., et al. (2020). Artificial intelligence applications in the development of autonomous vehicles: A survey. *IEEE/CAA Journal of Automatica Sinica, 7*(2), 315–329.

Yu, K.-H., Beam, A. L., & Kohane, I. S. (2018). Artificial intelligence in healthcare. *Nature Biomedical Engineering, 2*(10), 719–731.

Sanz, J.L. and Y. Zhu. *Toward scalable artificial intelligence in finance*. in *2021 IEEE International Conference on Services Computing (SCC)*. 2021. IEEE.

Bharadiya, J. (2023). Artificial Intelligence in Transportation Systems A Critical Review. *American Journal of Computing and Engineering, 6*(1), 34–45.

Lv, Z., Lou, R., & Singh, A. K. (2020). AI empowered communication systems for intelligent transportation systems. *IEEE Transactions on Intelligent Transportation Systems, 22*(7), 4579–4587.

Mahbooba, B., et al. (2021). Explainable artificial intelligence (XAI) to enhance trust management in intrusion detection systems using decision tree model. *Complexity, 2021*, 1–11.

Elahi, H., et al. (2021). On the Characterization and Risk Assessment of AI-Powered Mobile Cloud Applications. *Computer Standards & Interfaces, 78*, Article 103538.

Lamsal, P., *Understanding trust and security*. Department of Computer Science, University of Helsinki, Finland, 2001.

Roski, J., et al. (2021). Enhancing trust in AI through industry self-governance. *Journal of the American Medical Informatics Association, 28*(7), 1582–1590.

Kumar, P., Chauhan, S., & Awasthi, L. K. (2023). Artificial intelligence in healthcare: Review, ethics, trust challenges & future research directions. *Engineering Applications of Artificial Intelligence, 120*, Article 105894.

Bedemariam, R., & Wessel, J. L. (2023). The roles of outcome and race on applicant reactions to AI systems. *Computers in Human Behavior, 148*, Article 107869.

Zhang, C., et al. (2020). AIT: An AI-enabled trust management system for vehicular networks using blockchain technology. *IEEE Internet of Things Journal, 8*(5), 3157–3169.

Cabiddu, F., et al. (2022). Why do users trust algorithms? A review and conceptualization of initial trust and trust over time. *European Management Journal, 40*(5), 685–706.

Ferrer, X., et al. (2021). Bias and discrimination in AI: A cross-disciplinary perspective. *IEEE Technology and Society Magazine, 40*(2), 72–80.

Malek, M. A. (2022). Criminal courts' artificial intelligence: The way it reinforces bias and discrimination. *AI and Ethics, 2*(1), 233–245.

Mujtaba, D.F. and N.R. Mahapatra. *Ethical considerations in AI-based recruitment*. in *2019 IEEE International Symposium on Technology and Society (ISTAS)*. 2019. IEEE.

Asan, O., Bayrak, A. E., & Choudhury, A. (2020). Artificial intelligence and human trust in healthcare: Focus on clinicians. *Journal of Medical Internet Research, 22*(6), e15154.

Nicodeme, C. *Build confidence and acceptance of AI-based decision support systems-Explainable and liable AI*. in *2020 13th International Conference on Human System Interaction (HSI)*. 2020. IEEE.

Esmaeilzadeh, P. (2020). Use of AI-based tools for healthcare purposes: A survey study from consumers' perspectives. *BMC Medical Informatics and Decision Making, 20*(1), 1–19.

Wazan, A.S., et al., *Trust management for public key infrastructures: Implementing the X. 509 trust broker*. Security and Communication Networks, 2017. **2017**.

Park, C., et al. (2022). An enhanced ai-based network intrusion detection system using generative adversarial networks. *IEEE Internet of Things Journal, 10*(3), 2330–2345.

Benmoussa, A., et al. (2022). Interest Flooding Attacks in Named Data Networking: Survey of Existing Solutions, Open Issues, Requirements, and Future Directions. *ACM Computing Surveys, 55*(7), 1–37.

Nelson, G. S. (2019). Bias in artificial intelligence. *North Carolina Medical Journal, 80*(4), 220–222.

Zhu, T., et al. (2020). More than privacy: Applying differential privacy in key areas of artificial intelligence. *IEEE Transactions on Knowledge and Data Engineering, 34*(6), 2824–2843.

Van Bekkum, M., & Borgesius, F. Z. (2023). Using sensitive data to prevent discrimination by artificial intelligence: Does the GDPR need a new exception? *Computer Law & Security Review, 48*, Article 105770.

Hadj-Mabrouk, H. (2019). Contribution of artificial intelligence to risk assessment of railway accidents. *Urban Rail Transit, 5*(2), 104–122.

Peters, U. (2022). Algorithmic political bias in artificial intelligence systems. *Philosophy & Technology, 35*(2), 25.

King, T. C., et al. (2020). Artificial intelligence crime: An interdisciplinary analysis of foreseeable threats and solutions. *Science and Engineering Ethics, 26*, 89–120.

Marr, B. (2018). Is Artificial Intelligence dangerous? 6 AI risks everyone should know about. *Forbes*.

Ienca, M. (2023). On Artificial Intelligence and Manipulation. *Topoi*, 1–10.

Westerlund, M. (2019). The emergence of deepfake technology: A review. *Technology Innovation Management Review, 9*(11).

Vaccari, C. and A. Chadwick, *Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news*. Social Media+ Society, 2020. **6**(1): p. 2056305120903408.

Pedron, S. M., & da Cruz, J.d. A. (2020). The future of wars: Artificial intelligence (ai) and lethal autonomous weapon systems (laws). *International Journal of Security Studies, 2*(1), 2.

Wogu, I., et al. (2018). Super-Intelligent Machine Operations in Twenty-First-Century Manufacturing Industries: A Boost or Doom to Political and Human Development? *Towards Extensible and Adaptable Methods in Computing*, 209–224.

Surber, R., *Artificial intelligence: autonomous technology (AT), lethal autonomous weapons systems (LAWS) and peace time threats*. ICT4Peace Foundation and the Zurich Hub for Ethics and Technology (ZHET) p, 2018. **1**: p. 21.

de Ágreda, Á. G. (2020). Ethics of autonomous weapons systems and its applicability to any AI systems. *Telecommunications Policy, 44*(6), Article 101953.

Kumar, D. and K.P. Kumar. *Artificial Intelligence based Cyber Security Threats Identification in Financial Institutions Using Machine Learning Approach*. in *2023 2nd International Conference for Innovation in Technology (INOCON)*. 2023. IEEE.

Song, F., et al. (2020). Smart collaborative balancing for dependable network components in cyber-physical systems. *IEEE Transactions on Industrial Informatics, 17*(10), 6916–6924.

Solomon, A., et al. (2022). Contextual security awareness: A context-based approach for assessing the security awareness of users. *Knowledge-Based Systems, 246*, Article 108709.

Schiliro, F., Moustafa, N., & Beheshti, A. (2020). Cognitive privacy: AI-enabled privacy using EEG signals in the internet of things. *in 2020 IEEE 6th International Conference on Dependability in Sensor, Cloud and Big Data Systems and Application (DependSys)*. IEEE.

Schneier, B. (2015). *Secrets and lies: Digital security in a networked world*. John Wiley & Sons.

Weingart, S. H. (1987). Physical security for the μABYSS system. *in 1987 IEEE Symposium on Security and Privacy*. IEEE.

Brundage, M., et al., *The malicious use of artificial intelligence: Forecasting, prevention, and mitigation*. arXiv preprint arXiv:1802.07228, 2018.

Dhiman, A., & Toshniwal, D. (2022). AI-based Twitter framework for assessing the involvement of government schemes in electoral campaigns. *Expert Systems with Applications, 203*, Article 117338.

Bazarkina, D., & Pashentsev, E. (2020). Malicious use of artificial intelligence. *Russia in Global Affairs, 18*(4), 154–177.

Khan, M. K. (2020). *Technological advancements and 2020* (pp. 1–2). Springer.

Ansari, M. F., et al. (2022). The Impact and Limitations of Artificial Intelligence in Cybersecurity: A Literature Review. *International Journal of Advanced Research in Computer and Communication Engineering*.

Di Vaio, A., et al. (2020). Artificial intelligence and business models in the sustainable development goals perspective: A systematic literature review. *Journal of Business Research, 121*, 283–314.

Lukyanenko, R., Maass, W., & Storey, V. C. (2022). Trust in artificial intelligence: From a Foundational Trust Framework to emerging research opportunities. *Electronic Markets, 32*(4), 1993–2020.

Rehman, A., et al. (2022). CTMF: Context-aware trust management framework for internet of vehicles. *IEEE Access, 10*, 73685–73701.

Kingston, J. (2017). Using artificial intelligence to support compliance with the general data protection regulation. *Artificial Intelligence and Law, 25*(4), 429–443.

Mitrou, L., *Data protection, artificial intelligence and cognitive services: is the general data protection regulation (GDPR)'artificial intelligence-proof'?* Artificial Intelligence and Cognitive Services: Is the General Data Protection Regulation (GDPR)'Artificial Intelligence-Proof, 2018.

Žigienė, G., Rybakovas, E., & Alzbutas, R. (2019). Artificial intelligence based commercial risk management framework for SMEs. *Sustainability, 11*(16), 4501.

Jing, H., et al. (2021). An Artificial Intelligence Security Framework. *Journal of Physics: Conference Series*. IOP Publishing.

Chauhan, C., & Gullapalli, R. R. (2021). Ethics of AI in pathology: Current paradigms and emerging issues. *The American Journal of Pathology, 191*(10), 1673–1683.

Wickramasinghe, C.S., et al. *Trustworthy AI development guidelines for human system interaction*. in *2020 13th International Conference on Human System Interaction (HSI)*. 2020. IEEE.

Sobb, T., Turnbull, B., & Moustafa, N. (2020). Supply chain 4.0: A survey of cyber security challenges, solutions and future directions. *Electronics, 9*(11), 1864.

De, T., et al. (2020). Explainable AI: A hybrid approach to generate human-interpretable explanation for deep learning prediction. *Procedia Computer Science, 168*, 40–48.

Smith, H. (2021). Clinical AI: Opacity, accountability, responsibility and liability. *Ai & Society, 36*(2), 535–545.

Gerke, S., Minssen, T., & Cohen, G. (2020). Ethical and legal challenges of artificial intelligence-driven healthcare. In *Artificial intelligence in healthcare* (pp. 295–336). Elsevier.

Taddeo, M. (2019). Three ethical challenges of applications of artificial intelligence in cybersecurity. *Minds and machines, 29*, 187–191.

Hummer, W., et al. *Modelops: Cloud-based lifecycle management for reliable and trusted ai*. in *2019 IEEE International Conference on Cloud Engineering (IC2E)*. 2019. IEEE.

Jain, H., et al. *Weapon detection using artificial intelligence and deep learning for security applications*. in *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)*. 2020. IEEE.

Gopalan, S.S., A. Raza, and W. Almobaideen. *IoT security in healthcare using AI: A survey*. in *2020 International Conference on Communications, Signal Processing, and their Applications (ICCSPA)*. 2021. IEEE.

Norori, N., et al. (2021). Addressing bias in big data and AI for health care: A call for open science. *Patterns, 2*(10), Article 100347.

Giudici, P., & Raffinetti, E. (2023). SAFE artificial intelligence in finance. *Finance Research Letters*, Article 104088.

Meden, B., et al. (2023). Face deidentification with controllable privacy protection. *Image and Vision Computing*, Article 104678.

Kadykov, V., Levina, A., & Voznesensky, A. (2021). Homomorphic encryption within lattice-based encryption system. *Procedia Computer Science, 186*, 309–315.

Zhao, J., et al. (2023). Efficient and privacy-preserving tree-based inference via additive homomorphic encryption. *Information Sciences, 650*, Article 119480.

Tonyali, S., et al. (2018). Privacy-preserving protocols for secure and reliable data aggregation in IoT-enabled smart metering systems. *Future Generation Computer Systems, 78*, 547–557.

Hassan, M. U., Rehmani, M. H., & Chen, J. (2019). Differential privacy techniques for cyber physical systems: A survey. *IEEE Communications Surveys & Tutorials, 22*(1), 746–789.
Habbal, A., et al. (2017). Assessing experimental private cloud using web of system performance model. *International Journal of Grid and High Performance Computing (IJGHPC), 9*(2), 21–35.
Vollmer, S., et al., *Machine learning and AI research for patient benefit: 20 critical questions on transparency, replicability, ethics and effectiveness.* arXiv preprint arXiv: 1812.10404, 2018.
Tan, L., et al. (2021). Secure and resilient artificial intelligence of things: A HoneyNet approach for threat detection and situational awareness. *IEEE Consumer Electronics Magazine, 11*(3), 69–78.
Eluwole, O.T. and S. Akande. *Artificial Intelligence in Finance: Possibilities and Threats*. in *2022 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT)*. 2022. IEEE.
González-Gonzalo, C., et al. (2022). Trustworthy AI: Closing the gap between development and integration of AI systems in ophthalmic practice. *Progress in Retinal and Eye Research, 90*, Article 101034.
Desolda, G., et al. (2023). Explanations in warning dialogs to help users defend against phishing attacks. *International Journal of Human-Computer Studies, 176*, Article 103056.
Kong, H., et al. (2023). The impact of trust in AI on career sustainability: The role of employee–AI collaboration and protean career orientation. *Journal of Vocational Behavior*, Article 103928.
Pechegin, D. (2022). Judicial evaluation of data from artificial intelligence systems and other innovative technologies in transport. *Transportation Research Procedia, 63*, 86–91.
Kumar, G., et al. (2022). Investigation and analysis of implementation challenges for autonomous vehicles in developing countries using hybrid structural modeling. *Technological Forecasting and Social Change, 185*, Article 122080.
Pantic, M., & Rothkrantz, L. J. M. (2000). Automatic analysis of facial expressions: The state of the art. *IEEE Transactions on pattern analysis and machine intelligence, 22*(12), 1424–1445.
Zhang, T., et al. (2019). The roles of initial trust and perceived risk in public's acceptance of automated vehicles. *Transportation research part C: emerging technologies, 98*, 207–220.
Jacobsson, A., Boldt, M., & Carlsson, B. (2016). A risk analysis of a smart home automation system. *Future Generation Computer Systems, 56*, 719–733.
Chen, Y., et al. (2021). Trust calibration of automated security IT artifacts: A multi-domain study of phishing-website detection tools. *Information & Management, 58*(1), Article 103394.
Sebastian, A. M., & Peter, D. (2022). Artificial intelligence in cancer research: Trends, challenges and future directions. *Life, 12*(12), 1991.
Bohr, A., & Memarzadeh, K. (2020). The rise of artificial intelligence in healthcare applications. In *Artificial Intelligence in healthcare* (pp. 25–60). Elsevier.
Habbal, A., Goudar, S. I., & Hassan, S. (2019). A Context-aware Radio Access Technology selection mechanism in 5G mobile network for smart city applications. *Journal of Network and Computer Applications, 135*, 97–107.
Veselov, G., et al. (2021). Applications of artificial intelligence in evolution of smart cities and societies. *Informatica, 45*(5).
Kamruzzaman, M., Alrashdi, I., & Alqazzaz, A. (2022). *New opportunities, challenges, and applications of edge-AI for connected healthcare in internet of medical things for smart cities. Journal of Healthcare Engineering*.
Cheng, S., et al. (2022). *Roadmap toward the metaverse: An AI perspective. The Innovation, 3*(5).
Huynh-The, T., et al. (2023). Artificial intelligence for the metaverse: A survey. *Engineering Applications of Artificial Intelligence, 117*, Article 105581.
Lin, Y., et al. (2023). Blockchain-aided secure semantic communication for ai-generated content in metaverse. *IEEE Open Journal of the Computer Society, 4*, 72–83.
Yang, Q., et al. (2022). Fusing blockchain and AI with metaverse: A survey. *IEEE Open Journal of the Computer Society, 3*, 122–136.
Goodfellow, I., et al. (2020). Generative adversarial networks. *Communications of the ACM, 63*(11), 139–144.
Zhao, Z., et al. (2021). Challenges and opportunities of AI-enabled monitoring, diagnosis & prognosis: A review. *Chinese Journal of Mechanical Engineering, 34*(1), 1–29.
Liang, W., et al. (2022). Advances, challenges and opportunities in creating data for trustworthy AI. *Nature Machine Intelligence, 4*(8), 669–677.
Liyanage, M., et al. (2023). Open RAN security: Challenges and opportunities. *Journal of Network and Computer Applications, 214*, Article 103621.
van Mil, J., & Quintais, J. P. (2022). A Matter of (Joint) control? Virtual assistants and the general data protection regulation. *Computer Law & Security Review, 45*, Article 105689.
Karim, S. M., et al. (2022). Architecture, protocols, and security in IoV: Taxonomy, analysis, challenges, and solutions. *Security and Communication Networks*.
Alekseeva, L., et al. (2021). The demand for AI skills in the labor market. *Labour economics, 71*, Article 102002.
Jang-Jaccard, J., & Nepal, S. (2014). A survey of emerging threats in cybersecurity. *Journal of Computer and System Sciences, 80*(5), 973–993.
Francisco, M., & Linnér, B.-O. (2023). AI and the governance of sustainable development. An idea analysis of the European Union, the United Nations, and the World Economic Forum. *Environmental Science & Policy, 150*, Article 103590.
Sethi, K., et al. (2021). Attention based multi-agent intrusion detection systems using reinforcement learning. *Journal of Information Security and Applications, 61*, Article 102923.
Zhang, D., et al. (2022). Orchestrating artificial intelligence for urban sustainability. *Government Information Quarterly, 39*(4), Article 101720.
Dorri, A., et al. Blockchain for IoT security and privacy: The case study of a smart home. in 2017 IEEE international conference on pervasive computing and communications workshops (PerCom workshops). 2017. IEEE.
Askar, N., et al. (2023). *Forwarding Strategies for Named Data Networking based IOT: Requirements, Taxonomy, and Open Research Challenges. IEEE Access*.
Rieffel, E., & Polak, W. (2000). An introduction to quantum computing for non-physicists. *ACM Computing Surveys (CSUR), 32*(3), 300–335.
Groombridge, D. (2022). *Gartner Top 10 Strategic Technology Trends for 2023. Gartner*. https://www.gartner.com/en/articles/gartner-top-10-strategic-technology-trends-for-2023.
Gyongyosi, L., & Imre, S. (2019). A survey on quantum computing technology. *Computer Science Review, 31*, 51–71.
Zhang, C., et al. (2021). A survey on federated learning. *Knowledge-Based Systems, 216*, Article 106775.
Fedorchenko, E., Novikova, E., & Shulepov, A. (2022). Comparative review of the intrusion detection systems based on federated learning: Advantages and open challenges. *Algorithms, 15*(7), 247.