A Study on AI TRISM for Responsible and Secure Systems

Saurabh Ghosh (A paper written under the guidance of Prof. Raj Jain) Download 

Abstract:

Artificial Intelligence is growing very fast and thereby changes how people and organizations work. Some people would argue that AI is bringing new opportunities, but at the same time, they fail to acknowledge that it is also the reason of different risks, such as data misuse, different kinds of security attacks and more importantly bias. TRISM can play a very important role here, since TRISM is a framework that is there to make AI systems more trustworthy and in a sense safer than many other contemporary technologies. TRISM also can help many institutions to ride the wave of AI more efficiently. This paper explains what TRISM means, why it is important, and what benefits it provides. Also, this work will take a closer look at the important research of TRISM, navigate around major risks, and presents some instances which will show how TRISM improves the reliability of AI systems.

Keywords:

AI TRISM, AI Risk, Responsible AI, Governance, Compliance, AI Security, Trustworthy AI

Table of Contents:

- 1. Introduction
- 2. Literature Review
  - 2.1 AI TRISM: what it is and why it's used
  - 2.2 Governance and standards
  - 2.3 Trustworthiness and responsibility
  - 2.4 New security risks: Shadow AI and generative AI
- 3. Importance of TRISM
  - 3.1 Trust in AI Systems
  - 3.2 Risk Management
  - 3.3 Security Management
  - 3.4 Growing Adoption and Standards
- 4. Benefits of TRISM
  - 4.1 Benefits for Organizations and Individuals
  - 4.2 Transparency and Security
  - 4.3 Overall Impact
- 5. Systematic Risks
  - 5.1 Bias in Data and Models

    - 5.2 Security Risks and Data Attacks
    - 5.3 Data Misuse and Privacy Concerns
    - 5.4 Lack of Transparency (Black-Box Models)
    - 5.5 Compliance and Legal Risks
    - 5.6 Lack of Human Oversight
- 6. Comparison
- 7. Conclusion
- 8. References
- 9. Acronyms

1. Introduction

It is needless to say that Artificial Intelligence is now part of everyday work. We see it in applications, schools, and hospitals. Although, it can save time and help people make decisions. But it also raises hard questions. Can we trust its choices? Is it fair to everyone? Is it safe from attacks? This paper looks at one answer that many groups are using: AI Trust, Risk, and Security Management (TRISM). TRISM is a simple idea with three parts that support each other. The first is Trust: People should understand what the system does and why. They should feel confident that they behave the same way each time. The second is Risk, which means that Teams should look for problems early, such as bias, bad data, or model drift, and fix them before harm happens. And the third is Security, which basically Models and data should be protected from leaks, tampering, and attacks. Now, the question comes, why do we need this now? AI systems are bigger and faster than before. They affect real life and money. A small error can spread quickly. For example, a loan model that favors one group by mistake can deny credit to many people. A hospital model that is not secure could leak private records. These cases show that performance alone is not enough. We also need guardrails. In this paper, I use TRISM as the guiding frame. I will explain why it matters, what it adds, and where it helps most. Figure 1 illustrates the features of AI TRISM.

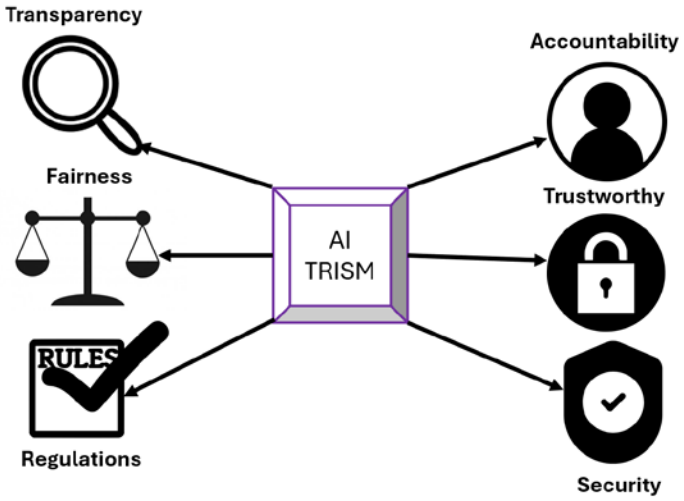


Figure 1: Features of AI TRISM.

2. Literature Review

This section talks about what others have said about TRISM, AI governance, trustworthiness, and new security risks.

2.1 AI TRISM: what it is and why it's used

Now-a-days many organizations use AI TRISM to make AI systems safer. Gartner introduced TRISM as one approach that brings trust, risk, and security together for the whole AI life cycle [Litan24]. Research supports this view. Habbal et al. describe TRISM as a structured way to manage AI risks while building trust with users and teams [Habbal24]. They also report early signs that TRISM can help create value across areas like finance and healthcare by keeping systems reliable and compliant [Habbal24] and industrial settings [AS25]. TRISM is not just a concept. It asks teams to do practical things: reduce bias, test models often, and harden security before deployment [Mathew24]. It also asks for clear methods to build trust (through transparency and accountability), manage risks (finding bias, errors, misuse), and enforce security (against cyber threats and adversarial attacks) [Mathew24]. Narula et al. show how TRISM fits inside broader AI governance, linking ethics, regulation, and reliability into one model [Narula25]. In short, TRISM bundles explainability, robustness, and compliance checks under one umbrella [Narula25]. That said, there are gaps. Narula et al. note that real-world case studies are still limited, and some steps for technical rollout are not fully clear yet [Narula25]. Even so, examples are starting to appear. Rupani and Rao describe TRISM ideas inside cloud DevSecOps to automate threat detection while shipping software fast [Rupani25]. Tasatanattakool shows a TRISM -style setup for digital asset management to track AI tool usage and reduce unmanaged information risk [T24]. Overall, TRISM connects trust, risk, and security into one playbook [Litan24][Habbal24]. It has promise, but we still need more hands-on case studies and tools [Narula25]. Early uses in cloud and asset management look encouraging [Rupani25][T24].

2.2 Governance and standards

Policy and standards groups - two of the most influencial writing groups are writing rules for responsible AI. The EU AI Act proposes a new approach with strong requirements for high-risk uses, such as clear documentation and human oversight [EC21]. However, in practice, this means that organizations should be planning for transparency, risk control, and also accountability. Two major standards help with this work. NIST AI RMF 1.0 sets a process to identify, measure, and manage AI risks across the life cycle [NIST23]. It highlights traits of trustworthy AI such as validity, fairness, privacy, security, and explainability [NIST23]. ISO/IEC 23894:2023 gives international guidance for risk controls and assurance across the AI life cycle [ISO23]. These however goes well with TRISM goals. Yet, many teams struggle with "who does what" in governance. Batool et al. reviewed the field and found gaps in assigning accountability, defining what is governed (data, models, outputs), and deciding when and how governance happens [AB25]. The author, named Wang adds that governance must also protect human judgment and autonomy, not just automate everything [Wang24]. Anand and Wan Mohamed show that when AI helps make public decisions, we need clear rules for oversight, fairness, and public trust [Anand25]. Overall, Laws and standards (EU Act, NIST, ISO) give a structure for responsible AI [EC21][NIST23][ISO23]. But many details on roles, timing, and methods are still evolving

[AB25]. Good governance should keep humans in the loop and protect human agency [Wang24][Anand25].

2.3 Trustworthiness and responsibility

Trust comes from systems that deserve trust. Lahusen et al. explain the difference between people trusting a system and the system being trustworthy in fact [Li23]. The goal is to meet clear criteria—reliability, transparency, fairness—rather than only persuading users to trust [Li23]. Explainability is a key part of trust. Barredo Arrieta et al. review XAI methods that help open the "black box" so we can see why a model gave a result [BA20]. But there are trade-offs. Simple models are easier to explain but may be less accurate. Complex models may perform better but are hard to interpret [BA20]. Another author, named Nastoska and her peers list some trust dimensions, specifically they are fairness, transparency and accountability—and showed that, if someone works on improving one of them, it eventually affecting other one in the pipeline [Nastoska25]. Overall, trustworthiness needs evidence, not promises [Li23]. XAI helps, but there are trade-offs with accuracy and complexity [BA20]. Teams must balance fairness, privacy, robustness, and other goals together [Nastoska25].

2.4 New security risks: Shadow AI and generative AI

Two fast-moving risks stand out. First, Shadow AI. This is when staff use AI tools without official approval or control. Silic et al. show that this is becoming common and can weaken governance and privacy if not managed [Silic25]. Puthal et al. call it a new "unseen frontier" in security because unofficial tools often skip normal reviews and monitoring [Deepak25]. The fix not only bans, but it also means tool registries, training, audits, safe alternatives, and better reporting paths [Silic25][Deepak25]. Second, Generative AI in the threat landscape. Ibrar et al. show that attackers can use generative AI to write better phishing, mutate malware, and create deepfakes at scale [WI25]. At the same time, defenders can also use AI to simulate attacks and improve detection [WI25]. Narula et al. note that "Generative AI security" is now a major research area, on both offense and defense [Narula25]. Overall, shadow AI expands risk inside organizations [Silic25][Deepak25]. Generative AI raises the stakes for both attackers and defenders [WI25][Narula25]. TRISM can help by adding controls, monitoring, and clear policies for these new fronts.

So, in a nutshell, it is clear that across the literature, one message repeats: responsible AI is a process, not a one-time fix. TRISM gives a practical frame to connect trust, risk, and security [Litan24][Habbal24][Mathew24]. Governance frameworks (EU AI Act, NIST, ISO) help put that into daily work [EC21][NIST23][ISO23]. Trustworthiness needs explainability, fairness, privacy, and real accountability, balanced case by case [BA20][Li23][Nastoska25]. New risks like Shadow AI and generative attacks make ongoing monitoring and human oversight even more important [Silic25][Deepak25][WI25][Narula25]. This review sets the stage for the next sections, where I discuss why TRISM matters in practice and how it can reduce risks while keeping AI useful and safe. The Picture as Figure 2 below illustrates the four main principles of AI TRISM.

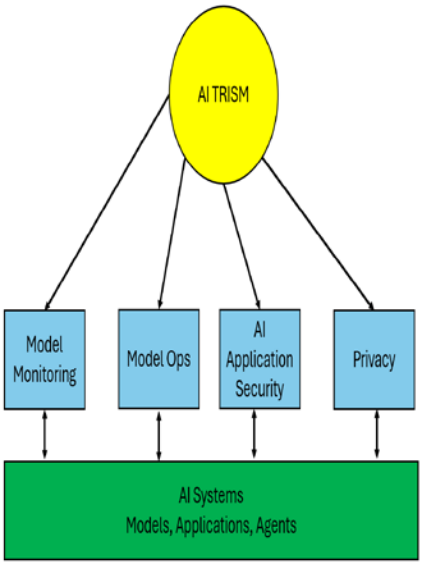


Figure 2: Four Principles of AI TRISM.

3. Importance of AI TRISM

3.1 Trust in AI Systems

If people cannot trust AI, they will eventually be going to avoid it. For example, a hospital may not want to use an AI system to suggest treatments if doctors are unsure how it works (and, that's where the explainable AI comes into the picture). A conclusive study by [NM25], and which is also a clinical chatbot trial, also reflects this need for close observation of this problem . Another example of that would be to see if customers would really like to let the AI decide what their credit scores are, although at present they are being calculated using mathematical modeling, but here trust issues can come from the customer end. Sensitive cases like this, where trust makes users feel confident that AI will behave in a transparent way are important.

3.2 Risk Management

The second part of TRISM is risk management. Every AI system has certain risks like biased data, prediction errors, or wrong use of private information. So, managing risk, here, means that, find the problems as early and as efficiently possible and then also fix them before they cause damage. An example would be about the organizations - who can test their models regularly to make sure they stay accurate. Risk management also can enforce companies to follow certain government rules like data protection laws.

3.3 Security Management

The third part is security management. AI uses large volumes of data, and that data, which primarily were being used to generate contents, can be attacked or stolen. Some hackers can even mislead AI models by giving them incorrect input so that they make wrong decisions. TRISM

Here comes into the picture and helps prevent these attacks by checking security gaps and protecting models and data from unauthorized access. This is important because even a small security failure can cause big harm to users and organizations, thereby losing billions of dollars.

3.4 Growing Adoption and Standards

Together, these three ideas make TRISM important. It supports both technical safety and ethical responsibility. Many governments and organizations have already started to adopt TRISM principles. There already are different established frameworks, like AI RMF 1.0, [NIST23] and the EU AI Act [EC21], which denotes that the world is moving toward stronger AI governance and it is progressing that way very fast. Overall, these frameworks share the same goals as TRISM: to make AI systems more transparent than they are currently.

4. Benefits of AI TRISM

4.1 Benefits for Organizations and Individuals

AI Trust, Risk, and Security Management have many real benefits for both companies and everyday people. When AI is built and used with care, it becomes easier to trust and much safer to use. TRISM helps make that possible. It does not only focus on technology but also connects to ethics, rules, and how people feel about AI. One big reason TRISM is important is that it helps people trust AI. Most people don’t like systems they can’t understand.

4.2 Transparency and Security

In contrast, TRISM really encourages the AI to be very open and wide – perhaps more transparent so that, when people can see these things clearly, they can build the confidence that, the system is quite fair and build the trust eventually. Also, since AI uses a lot of parameters and information, hackers are always trying to change the data or sometimes trick the specific model to get wrong answers however, TRISM can help solve these issues by building trust from the user side. Also, TRISM puts security at the center. It reminds teams to protect data, control who has access, and keep an eye on unusual activity. So, when we see that, systems are being built this way, they are harder to attack and less likely to leak private information indeed. TRISM also helps make AI fair and ethical. It encourages designers to think about how their models affect people. For example, an AI that helps choose job candidates, basically doing the job of an HR, should not favor one candidate over another, although artificial intelligence has a higher chance of doing so since all these models contain data that are biased in the sense that how companies are feeding them. TRISM, the framework, pushes for fairness here, so that AI decisions can be explained. Also, many more countries at this point have rules about how AI should be utilized. The EU AI Act [EC21] and the ISO/IEC 23894:2023 standard [ISO23] are good examples. Many people still worry about AI — that it might take jobs, make unfair choices, or collect data without permission. TRISM helps reduce these fears. When companies show that they follow clear steps for safety, fairness, and responsibility, people start to feel more comfortable. Over time, this builds confidence. It shows that AI can be used in a good way — one that respects people and their privacy.

4.3 Overall Impact

Overall, TRISM makes AI stronger and safer in every way. It builds trust, lowers risks, improves security, and supports fairness. It also helps organizations follow laws and rebuild public confidence. In short, TRISM helps turn AI from something people worry about into something they can trust and depend on. Figure 3 denotes the Model monitoring of AI TRISM.

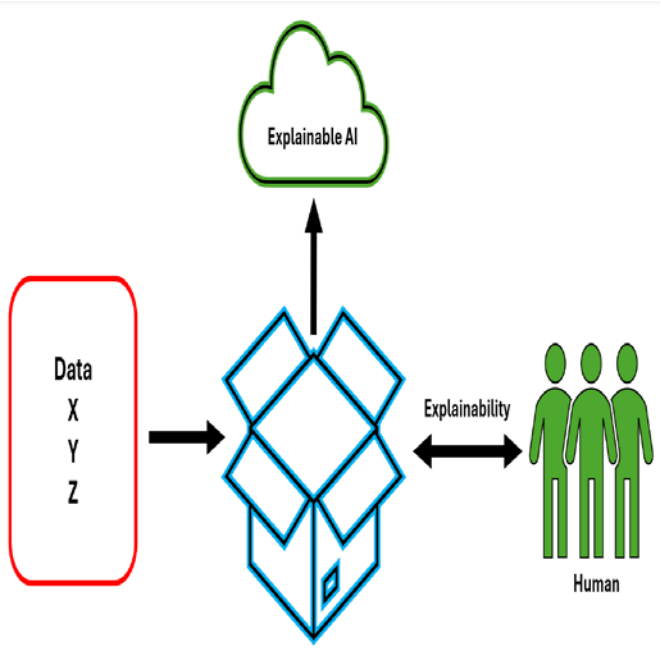


Figure 3: TRISM models monitoring functionality.

5. Systematic Risks

5.1 Bias in Data and Models

It is true that AI has bias. Figuratively, all it does is learn from data, and if the data contains just one- sided examples, the model may also produce biased results. For example, if an AI model for hiring is trained mostly on data from one group of people, it may favor that group and ignore others. This is, therefore, something which can lead to decisions that are basically unfair. By doing or promoting regular testing, TRISM greatly helps reduce bias. The framework also encourages AI to use transparent datasets so that designers can see where biases are. Certainly, afterwards, the reason for that was to rectify it in a way that it solves the problem.

5.2 Security Risks and Data Attacks

Another important risk that is to be discussed here is security. AI can be manipulated in many ways. Hackers use methods such as poisoning of data. These attacks can cause serious harm, especially in areas that are quite sensitive to handle - for example banking, or healthcare - which

is also protected by several different federal laws. TRISM here takes that very seriously and makes sure to have additional and strong security checks at every stage of this generative development.

5.3 Data Misuse and Privacy Concerns

This, however, creates the risk of data misuse because if sensitive data is not stored carefully, it can be stolen. For example, an AI system that analyzes patient data could expose private health information if not protected well. TRISM supports privacy by design. It recommends data encryption to protect users' personal information. It also enforces the AI systems to follow data protection laws, for example, GDPR.

5.4 Lack of Transparency (Black-Box Models)

A common problem in AI that many people encounter these days, is that many of the currently available AI models, which are also quite very popular among its users act like a black box – meaning they make decisions, but it is not explainable and so overall it's hard to see how it made this decision. This lack in transparency makes it difficult for people to trust AI systems for the decisions which are critical. For example, sensitive issues like defense security cannot just trust AI without a clear explanation. TRISM here encourages AI to explain each theory and thus, this problem can be solved. Overall, by making models easier to interpret, organizations can show how decisions are made.

5.5 Compliance and Legal Risks

As AI becomes part of business and government, compliance is another risk that everyone should take note of. There are many companies who may face legal and other regulatory problems if their AI systems break regulatory rules. For instance, if the AI fails to reveal how it made a certain decision – then it could violate transparency laws. There are also ethical risks that are being attached with it because it can harm people in different ways. TRISM helps organizations manage all the above-mentioned risks by supporting AI methods with appropriate legal frameworks. It makes sure that all stages—from design to deployment—follow accountable standards.

5.6 Lack of Human Oversight

Finally, a growing risk is the lack of human control. There are some AI systems, which are pretty advanced that make decisions automatically without enough needing a human supervision. This can be very dangerous if the system behaves in unexpected ways, like we have seen in Si-Fi movies. TRISM stresses human oversight and accountability, and thus to be making sure that people remain part of the decision-making process, especially in critical applications indeed..

6. Comparison

In order to understand how AI TRISM works, a visually structured tech stack of AI PRISM is being illustrated since it helps to look at its structure visually and compare it with systems that do not follow TRISM principles. A visual model can show how different parts of AI TRISM connect

and theoretically depend on each other. Figure 4 below how it works in layers, from the basic technology up to advanced AI governance.



Figure 4: This figure shows AI TRISM Technical Stack

Table 1. AI systems with and without TRISM

Element	With TRISM	Without TRISM
Trust	Users recognize decisions	Users are not sure
Risk Management	Risks recognized and reduced subsequently	Bias
Security	Data leak protection	Exposed to cyber threats
Fairness	Ethical results	Possible discrimination
Compliance	Follows policy standards	Higher chance of violations
User Confidence	High adoption	Low adoption due to safety concerns

7. Conclusion

In the history of civilization, there were many disruptive technologies the world has encountered, but as many experts believe, AI is going to be disruptive that basically will change how people

and organizations work. Although, it is true that AI can make decisions fast, it also brings serious challenges related to fairness. So, to make sure that the AI systems are reliable, the organizations, who are involved in this process or at least utilize AI into a great portion of their work, need to know how these systems are built, and this is where TRISM comes into the picture. This work discussed what TRISM means and why it matters. It showed that TRISM has three key ideas: trust, risk, and security. Trust helps users believe in the system and its results. Risk management identifies possible problems and controls them before they cause harm. There are also many studies which dig deeper on TRISM, but studies such as those by Habbal et al. [Habbal24] and Litan [Litan24] describe TRISM as a modern framework that connects technical safety with ethical responsibility. Standards from organizations like NIST [NIST23] and ISO [ISO23] give practical guidance on how to manage AI risks. These studies have made that clear that TRISM is not only a theory but also a growing practice supported by real policies and tools. Although there are many benefits with TRISM, there are still many challenges that need to be taken care of. Many organizations do not yet have the tools or skills to apply TRISM fully. There are couple of future directions that can be worked on, however, focusing on creating common standards and training programs will be important so that TRISM can be used across all industries.

8. References

[Silic25] M. Silic, K. Kind-Trüller, “From Shadow It to Shadow AI–Threats, Risks and Opportunities for Organizations,” Strategic Change, 2025, DOI 10.1002/jsc.2682, https://onlinelibrary.wiley.com/doi/10.1002/jsc.2682

[Wang24] Y. Wang, “Human Subjectivity in Information Practice and AI Governance,” Proceedings of the Association for Information Science and Technology, 2024, DOI 10.1002/pra2.1111, https://asistdl.onlinelibrary.wiley.com/doi/10.1002/pra2.1111

[Habbal24] A. Habbal, M. K. Ali, M. A. Abuzaraida, “Artificial Intelligence Trust, Risk and Security Management (AI TRISM): Frameworks, applications, challenges and future research directions,” Expert Systems with Applications, 2024, pp. 122442, https://www.sciencedirect.com/science/article/pii/S0957417423029445

[Mathew24] A. Mathew, “AI TRISM: Trust, Risk, and Security Management in Cybersecurity,” Multidisciplinary International Journal of Research and Development, 2024, pp. 84-90, https://www.mijrd.com/papers/v4/i3/MIJRDV4I30004.pdf

[T24] P. Tasatanattakool, “Digital Asset Management Process using AI TRISM,” 2024 IEEE 6th International Conference on Power, Intelligent Computing and Systems (ICPICS), 2024, pp. 234-239, https://ieeexplore.ieee.org/document/10532376

[Narula25] S. Narula, M. Ghasemigol, J. Carnerero-Cano, A. Minnich, E. Lupu, D. Takabi, “Exploring Research and Tools in AI Security: A Systematic Mapping Study,” IEEE Access, 2025, pp. 84057-84080, https://ieeexplore.ieee.org/document/10988535

[Rupani25] V. S. K. Rupani, C. M. Rao, “AI TRISM in Cloud Security: Leveraging AI-Powered DevSecOps for Robust Security Management,” Integrating Advanced Technologies for

Enhanced Security and Efficiency, Springer, 2025, pp. 383-398, https://link.springer.com/chapter/10.1007/978-3-031-91798-1_25

[Anand25] S. Anand, W. M. Wan Mohamed, "AI as an Autonomous Decision-Maker in Governance: A New Era of Policy Formulation," The Power of Agentic AI: Redefining Human Life and Decision-Making, Springer, 2025, pp. 81-96, https://link.springer.com/chapter/10.1007/978-3-031-89424-4_6

[Deepak25] Deepak Puthal, A. K. Mishra, S. P. Mohanty, A. Longo, and C. Y. Yeun, "Shadow AI: Cyber Security Implications, Opportunities and Challenges in the Unseen Frontier," SN Computer Science, vol. 6, no. 5, Apr. 2025, doi: https://doi.org/10.1007/s42979-025-03962-x

[WI25] W. Ibrar, D. Mahmood, A. S. Al-Shamayleh, G. Ahmed, S. Z. Alharthi, and A. Akhunzada, "Generative AI: a double-edged sword in the cyber threat landscape," Artificial Intelligence Review, vol. 58, no. 9, Jul. 2025, doi: https://doi.org/10.1007/s10462-025-11285-9

[BA20] A. Barredo Arrieta et al., "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," Information Fusion, vol. 58, pp. 82–115, Jun. 2020, doi: https://doi.org/10.1016/j.inffus.2019.12.012

[AB25] A. Batool, D. Zowghi, and M. Bano, "AI governance: a systematic literature review," AI and Ethics, vol. 5, no. 3, pp. 3265–3279, Jan. 2025, doi: https://doi.org/10.1007/s43681-024-00653-w

[EC21] European Commission, "Proposal for a Regulation laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act)," COM(2021) 206 final, Brussels, 21 April 2021. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206

[ISO23] ISO/IEC 23894:2023. Information technology — Artificial intelligence — Guidance on risk management. International Organization for Standardization, Feb. 2023. URL: https://www.iso.org/standard/77304.html

[Li23] C. Lahusen, M. Maggetti, and M. Slavkovik, "Trust, trustworthiness and AI governance," Scientific Reports, vol. 14, no. 1, Sep. 2024, doi: https://doi.org/10.1038/s41598-024-71761-0

[Litan24] A. Litan, "Tackling Trust, Risk and Security in AI Models," Gartner, 24 Dec. 2024. URL: https://www.gartner.com/en/articles/ai-trust-and-ai-risk

[Nastoska25] A. Nastoska, B. Jancheska, M. Rizinski, and D. Trajanov, "Evaluating trustworthiness in AI: Risks, metrics, and applications across industries," Electronics, 14(13) (July 2025), Art. 2717. DOI: https://doi.org/10.3390/electronics14132717

[NIST23] NIST, "Artificial Intelligence Risk Management Framework 1.0 (AI RMF 1.0)," National Institute of Standards and Technology, Jan. 2023. DOI: https://doi.org/10.6028/NIST.AI.100-1

[Papagiannidis25] E. Papagiannidis, P. Mikalef, and K. Conboy, "Responsible artificial intelligence governance: A review and research framework," Journal of Strategic Information Systems, 34(2) (2025), Article 101885. DOI: https://doi.org/10.1016/j.jsis.2024.101885

[Polemi24] N. Polemi, I. Praça, K. Kioskli, and A. Bécue, "Challenges and efforts in managing AI trustworthiness risks: a state of knowledge," Frontiers in Big Data, 7 (May 2024). DOI: https://doi.org/10.3389/fdata.2024.1381163

[AS25] A. S. Al-Ghamdi and S. Z. Alharthi, "Holism of Thermal Energy Storage: A Data-Driven Strategy for Industrial Decarbonization," Sustainability, vol. 17, no. 19, pp. 8745–8745, Sep. 2025, doi: https://doi.org/10.3390/su17198745

[NM25] N. M. Azmi et al., "A prospective single-masked, non-inferiority, parallel-group randomized controlled trial of the efficacy of a ChatGPT-based AI chatbot to improve Boston bowel preparation scores for colonoscopy preparation: A trial protocol," PLoS ONE, vol. 20, no. 10, pp. e0334349–e0334349, Oct. 2025, doi: https://doi.org/10.1371/journal.pone.0334349

9. Acronyms

Acronym	Full Form
AI	Artificial Intelligence
TRISM	Trust, Risk, and Security Management
XAI	Explainable Artificial Intelligence
NIST	National Institute of Standards and Technology
ISO	International Organization for Standardization
GDPR	General Data Protection Regulation
EU	European Union
RMF	Risk Management Framework
AI RMF 1.0	Artificial Intelligence Risk Management Framework

Last modified on November 17, 2025
This and other papers on recent advances in networking are available online at
http://www.cse.wustl.edu/~jain/cse5700-25/index.html
Back to Raj Jain's Home Page