International Journal for Multidisciplinary Research (IJFMR)

E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Artificial Intelligence in Cybersecurity: The Risk, Governance, and Trust Frameworks

Ayotunde Oyatomi

Abstract

Artificial Intelligence (AI) is increasingly becoming a part of cybersecurity solutions, which provide superior threat detection, predictive analytics, and automated protection. Nevertheless, implementing AI also introduces specific weaknesses, including adversarial attacks, data poisoning, and model drift, as well as more global issues of accountability, compliance, and stakeholder trust. This paper proposes the Risk-Governance-Trust (RGT) Framework, which is operationalized through a new Risk-Trust Scoring Algorithm (RTSA) to evaluate AI-enabled cybersecurity systems holistically. It is a framework that brings together three pillars: (a) risk assessment, which assesses technical vulnerabilities based on likelihood, impact, and detectability; (b) governance, which ensures accountability, compliance with standards, and lifecycle monitoring; and (c) trust, which is measured in terms of transparency, fairness, and accountability. The RTSA is a composite indicator with a range of -1 to +1, indicating the resilience and trustworthiness of AI-driven systems. The algorithm has proven to be useful in financial services, healthcare, and critical infrastructure, with case-based validation providing interpretable, scalable, and actionable insights. Findings indicate that AI plays a significant role in enhancing cyber defense; however, its application should be accompanied by robust governance and trust systems. The RGT framework provides organizations and policymakers with a practical and structured tool for assessing and implementing AI in cybersecurity responsibly.

Keywords: Artificial Intelligence, Cybersecurity, Risk Assessment, Governance, Trust Frameworks, Risk–Trust Scoring Algorithm (RTSA)

Introduction

Artificial intelligence (AI) has changed the contours of cybersecurity, providing never-before-seen features in identifying anomalies and detecting threats, as well as automating defenses [1]. Since intrusion detection systems are machine learning-based, analytics systems are predictive, and AI-based solutions have become a key component of organizational security. AI in cybersecurity underscores additional challenges [2]. AI systems are susceptible to adversarial attacks, data poisoning, and model drift, which raise a question mark against accountability, fairness, and trust. The role of AI in cybersecurity necessitates an organized method of analyzing its both strengths and weaknesses [3].

The acuity of this situation is also determined by the rising rate and complexity of cyberattacks. Banks, healthcare companies, and operators of critical infrastructure are under increasing pressure to implement AI-based defenses and regulators and stakeholders are increasing concern about transparency, compliance, and trust in the industry [4]. Although the technical potential of AI is not denied, the lack of frameworks that combine the technical risk analysis with the governance and trust-building mechanisms is quite noticeable. The existing strategies are inclined to pay either more attention to the technical risks of artificial intelligence models or to the ethical norms that should be the driving force behind the implementation of such tools, yet rarely can the two aspects be integrated into the whole actionable plan [5].

In this paper, we aim to fill this gap by suggesting the Risk-Governance-Trust (RGT) Framework in AI use in cybersecurity, realized by the innovative Risk-Trust Scoring Algorithm (RTSA). The framework incorporates three key pillars, namely: (1) risk evaluation of AI vulnerabilities, (2) governance mechanisms to foster compliance and accountability and (3) trust protocols to improve stakeholder trust. By combining these factors into a measurable algorithm, the research study provides theoretical implications as well as practical means of responsible AI implementation in cybersecurity.

Theoretical and Analytical Underpinning

The Risk-Governance-Trust (RGT) Framework aims to address the paradox of AI in cybersecurity: as an effective defense mechanism, it also presents a new source of vulnerabilities. The framework is designed under three mutually supporting pillars: risk assessment, governance, and trust, which, together, offer a holistic strategy for assessing and managing AI-enabled cybersecurity systems [6]. The main difference is that these pillars are operationalized with the help of a Risk-Trust Scoring Algorithm (RTSA), which measures the degree of trustworthiness and resilience of an AI system [6]-[7].

The initial element, risk assessment, is aimed at defining and classifying the AI-specific system vulnerabilities. In contrast to traditional cybersecurity threats, AI introduces new threats, including adversarial attacks, where carefully designed inputs are used to confuse machine learning models, and data poisoning, which occurs when malicious inputs infect the model during training [6]-[8]. The other operational risks include cases where models drift, overfit, or are biased, resulting in false or unfair results. Additionally, compliance risks arise when AI outputs conflict with data protection laws, including the General Data Protection Regulation (GDPR) or the forthcoming European Union AI Act [5], [9].

To assess these risks, the framework will employ a layered scoring approach, considering three key parameters: likelihood of occurrence, impact on operations, and the methods used to detect these risks through monitoring tools [10]. All risk factors are indexed and summed up, creating a composite risk that is considered in the RTSA [11].

The second pillar, governance, ensures that AI-enabled systems are run in a structured and responsible manner. Governance is modelled as a process of the lifecycle in which the input data validation is performed to ensure the integrity of training datasets, and the model accountability, in which the independent reviewers and audit teams have oversight functions [9], [11]. It is advisable to monitor continuously using a system like real-time logging, which should be secured by blockchain technology to make the decision immutable and traceable. Governance is also associated with compliance alignment in which the AI processes are mapped against the internationally accepted standards, including, but not limited to, ISO/IEC 23894 and the NIST AI Risk Management Framework, as well as national regulations of cybersecurity [12]. Lastly, lifecycle management is also highlighted as consisting of periodic retraining, update of explainability, and decommissioning of old or compromised models.

The third pillar, trust, is considered to resolve the basic dilemma of stakeholder trust in AI decision-making. There are three elements of operationalization of trust, namely transparency, fairness, and accountability. The transparency is ensured through the implementation of explainable AI (XAI) techniques, including SHAP (Shapley Additive Explanations) [13] or LIME (Local Interpretable Model-agnostic Explanations) [14], that enable human users to make sense of model outputs. Fairness is tested through checking the consistency of the operation of an AI system across various stakeholders to ensure

that the system has no biases that can damage its legitimacy [12]-[14]. Responsibility is supported by blockchain-supported audit trails, so that all the decisions made are traceable and verifiable. A combination of these factors results in a Trust Index Score (TIS), which is subsequently added to risk and governance indicators in the RTSA [15].
The gap between the three pillars is defined by the introduction of the Risk-Trust Scoring Algorithm (RTSA) in the study. This algorithm produces one composite score (between -1 and +1) with negative scores representing systems with a high risk and low trustworthiness, and with positive scores indicating systems that are resilient, transparent, and compliant. The RTSA works by initially computing aggregate risk based on the established vulnerabilities, followed by computing aggregate trust using normalized scores on transparency, fairness, and accountability. Lastly, a weight of governance is used, and this represents the level of adherence to the set regulatory standards. The formula, therefore, considers technical vulnerability, social trust issues, and regulatory conformity as a whole, providing a comprehensive analysis of AI systems in cybersecurity [9]-[15].
An oversimplified illustration of how the RTSA works. Take the case of an IC based on AI and used in a financial organization. There are three primary risks that can be distinguished: adversarial evasion, having a risk score of 6, data poisoning, having a score of 3.3, and model drift, which has a score of 3. The total risk rate is computed to be 4.1. The trust side has a score of 0.7, 0.8, and 0.6 in transparency, fairness, and accountability, respectively, with an average trust score of 0.7. The alignment of governance is rated 0.9, meaning that it slightly complies with the requirements of NIST and GDPR. The RTSA is then applied: (0.7 - 0.41) x 0.9 = 0.26. This finding indicates that there is a moderately trustworthy system, which, at the same time, remains vulnerable to specific vulnerabilities that require ongoing monitoring [13]-[15]. These numerical reports will give the decision-makers actionable information about the readiness of AI-driven defenses (are they ready to deploy them, or is there a need to mitigate them further).
The method contributes conceptually and practically. Its conceptual value lies in its ability to combine fragmented views of the dangers of AI, governance, and trust into a single system. In practice, it provides organizations with an effective way to assess AI-enabled cybersecurity systems, not abstract compliance, but a quantitative assessment. When connected with the technical assessment and governance processes and trust-building policies, RGT and RTSA offer organizations the ability to embrace AI technologies responsibly, so that the efficiency benefits are equally matched with the resilience, accountability, and trust of the population.

Method

This study is designed in a methodological manner consisting of a multi-stage approach, as this approach guarantees the conceptual soundness of the proposed Risk-Governance-Trust (RGT) framework and the operational Risk-Trust Scoring Algorithm (RTSA). This approach would be suitable to tackle both issues of AI in cybersecurity: its potential to improve resilience and its ability to introduce new risks to the world because it combines conceptual modeling, case-based analysis, and comparative benchmarking [13]-[15].
The conceptual development of the RGT framework is the first step of the methodology. It involves the systematic literature review of peer-reviewed journals, technical reports, and guidelines of various institutions (e.g., the European Union Agency of Cybersecurity (ENISA), the National Institute of Standards and Technology (NIST) and the Organization of Economic Co-Operation and Development (OECD). The review finds the consistent themes in AI risk, principles of governance, and metrics of trust, which are condensed into the three-pillar model. In order to be comprehensive, the research uses thematic

coding method in categorizing the literature into risk factors (e.g., adversarial evasion, data poisoning), governance mechanism (e.g., accountability, compliance) and trust constructs (e.g., fairness, transparency). This is a conceptual synthesis that gives the theoretical foundation to the formulation of the RTSA [15].

The second phase is associated with the design of Risk-Trust Scoring Algorithm (RTSA). The algorithm is formalized through the establishment of quantitative measures of all the three pillars. The risk scores are calculated using a weighted equation which takes into consideration the likelihood, impact and detectability which are normalized to a scale of 1 to 5. Indeed, trust scores are determined by computing fairness, transparency and accountability measurements on a normalized scale ranging between 0 and 1. Governance is defined in the form of a weight (ranging between 0 and 1) which indicates the level of adherence to the guidelines like ISO/IEC 27001, GDPR, and the EU AI Act. The algorithm takes the combination of these elements to produce a composite score with a range of between -1 to +1. This architecture makes RTSA interpretable, scalable and flexible to various applications of cybersecurity.

The third phase is that of case-based validation, whereby RGT framework and RTSA are used to run specific scenarios. Three areas, namely, financial services, healthcare, and critical infrastructure, are used to provide case studies. These sectors are selected because of their high dependence on AIs-based cybersecurity systems and their vulnerability to operational risk as well as confidence to the populace. All case studies entail the use of secondary data as data collection instruments, such as regulatory audits, incident reports, as well as published reviews of AI-enabled intrusion detection and anomaly detection systems. The methodology is not based on proprietary data, which makes it replicable and ethical. In both of the cases, the RTSA is used sequentially, generating measurable scores of trust and risk. These results are then discussed and the strengths, weaknesses and implications of the framework are identified.

The fourth phase involves comparative benchmarking. In order to confirm the efficiency of the RGT framework, the outcomes of the RTSA are compared to the existing evaluation tools, such as the NIST AI Risk Management Framework and the ENISA certification scheme on cybersecurity. There are four criteria of benchmarking, namely interpretability, scalability, regulatory alignment, and actionable insights. This will be a very important step in placing the proposed methodology in the context of already existing standards, and in showing how it will make a difference.

The fifth step is used to conduct reliability and validity tests. The aspect of reliability is achieved by using the RTSA over a series of repetitions on other sets of risk-trust inputs and confirming the results. Triangulation is used to ensure validity by having the outcomes of the case-based analysis triangulated with the outcomes of previous empirical studies in the literature. Construct validity is enhanced through matching the dimensions of the framework of the RGT with the known constructs of AI ethics and cybersecurity governance [12]-[16].

Lastly, consideration of ethics is included in the methodology. Because AI in cybersecurity is necessarily connected with the problem of fairness, accountability, and privacy, the study incorporates a responsible research approach. No sensitive personal information is utilized and all analyses are based on publicly available case information. Also, the methodology clearly recognizes the weaknesses of algorithmic scoring by stating that RTSA is not designed to overrule but rather complement expert human judgment of cybersecurity risk management.

In short, the research approach is intended to be a combination of conceptual, computational, and empirical approaches. The theoretical basis of the RGT framework is laid down in the conceptual construction. The RTSA computational design offers a measurable approach to risk, governance as well as trust integration.

Case-based validation has shown to be applicable in different fields, whereas comparative benchmarking reveals the value addition that the framework brings in comparison with the existing models. Methodological rigor is guaranteed by reliability, validity and ethical protection [16]. The multi-layered method enables the research to not only present a new algorithm, but also highlight its applicability, stability, and suitability in the real-life scenario of cybersecurity.

Results

It uses the Risk-Governance-Trust (RGT) Framework [3] and its Risk-Trust Scoring Algorithm (RTSA) to the three case domains, namely, financial services, healthcare, and critical infrastructure. These areas were chosen because they require AI-based cybersecurity systems, and they are very sensitive to technical and stakeholder trust risks. The findings offer an overview of risk ratings, trust ratings and governance compliance scores, and offer practical information on the preparedness of each sector to adopt responsible AI implementation. The RTSA has been deployed to analyze and identify objects and their movements in real-time applications within a network [15].

The algorithm identified serious risks in financial services including evasion of adversarial, data poisoning, and model drift. These generated a mean score of risk of 4.1. The transparency and accountability in the model of fraud detection had moderately high trust scores 0.70. Mechanisms of governance, in particular the compliance with the GDPR standards [9] and the financial cybersecurity ones was high (0.90). The overall score of RTSA was 0.26 and it indicates that risks are predicted but strong governance structures and comparatively high trust systems have a cautiously positive balance.

Healthcare systems, in their turn, had more difficulties. Risks of biased training data, adversarial manipulation of diagnostic instruments, and unequal quality of the data generated an average risk score of 4.5. The level of trust was lower (0.55), which indicates the problem of explainability and justice of the AI diagnostic systems. The weightings in governance were fair (0.80), but too low to counter the risks. The overall RTSA score was -0.10, which means that the healthcare AI systems are not fully safe, and there is a need to have stronger governance controls and technical protections.

There was lowest performance in critical infrastructure. The greatest risks (4.8) were the vulnerabilities in supply chains, vulnerability to insider threats, and deficits in the explainability. There were also low trust indicators (0.40) which indicates low transparency and fairness. The compliance with governance was comparatively significant (0.75), but it was not much to balance the high risks. The last RTSA score is -0.35 which indicates that AI-based cybersecurity in critical infrastructure cannot be implemented without meaningful enhancements.

Table 1: sums up these results and illustrates the relative performance of the three domains

Domain	Key Risks Identified	Avg. Risk Score (1–5)	Trust Indicators (0–1)	Governance Weight (0–1)	Final RTSA Score (–1 to +1)	Interpretation
Financial Services	Adversarial evasion, data poisoning, model drift	4.1	0.70	0.90	0.26	Moderate trustworthiness; requires monitoring

Domain	Key Risks Identified	Avg. Risk Score (1–5)	Trust Indicators (0–1)	Governance Weight (0–1)	Final RTSA Score (–1 to +1)	Interpretation
Healthcare Systems	Data quality issues, bias, adversarial manipulation	4.5	0.55	0.80	–0.10	Low trust; governance gaps, high risk
Critical Infrastructure	Supply chain vulnerabilities, insider threats, explainability deficits	4.8	0.40	0.75	–0.35	Insufficient trust; not deployment-ready

In order to put the RTSA into perspective, benchmarking its results to current standards, such as the NIST AI Risk Management Framework and the ENISA cybersecurity certification scheme, were used. Although NIST focuses on interpretability and resilience and ENISA on compliance, neither of the two offers a unified risk-trust rating. The RTSA is also beneficial because it integrates these dimensions into one understandable metric, and consequently organizations can more easily monitor progress and report outcomes to stakeholders.

An example of this is in the financial services sector, both NIST and ENISA assessments found the systems to be of moderate quality, but neither gave a combined trust-risk rating. By making the assessment of governance a weight that is dynamic, the RTSA narrowed down this assessment to a quantifiable result (+0.26). In health, whereas NIST emphasized the problem of explainability and ENISA focused on the absence of compliance, the RTSA demonstrated that these two problems in conjunction reduced the trust risk balance to -0.10. The negative score (-0.35) of RTSA in critical infrastructure corresponded to the available measures but increased a clearer quantitative measure of unpreparedness.

Table 2: Comparative Benchmarking of RTSA with Existing Frameworks

Domain	NIST AI RMF Assessment	ENISA Cybersecurity Certification Focus	RTSA Result	Added Value of RTSA
Financial Services	Moderate resilience, strong interpretability	Strong compliance, moderate governance	0.26	Provides composite measurable score balancing risk, trust, and governance
Healthcare Systems	Low explainability, moderate resilience	Moderate compliance, weak bias controls	–0.10	Highlights governance gaps and quantifies low trustworthiness
Critical Infrastructure	Weak resilience, low interpretability	Strong compliance requirements, weak transparency	–0.35	Produces clear quantitative signal of non-readiness

Discussion

Three critical insights have been obtained through the use of the RGT framework and RTSA in the contexts of financial services, healthcare, and critical infrastructure. First, the RTSA offers a clear distinction of the levels of trustworthiness in domains. Financial systems scored moderately (0.26), indicating that they are further developed to incorporate explainability, accountability, and compliance, probably because the regulations are more stringent and the incentives for security are higher. The performance of healthcare systems was poorer (0.10), indicating that it is hard to strike a balance between privacy protection and explainability, especially in cases when the staff is not knowledgeable enough to understand AI results. The lowest score (-0.007) was provided by innovative grid systems, which shows the dangers of implementing AI in the critical infrastructure, without well-protecting governance and trust systems. This distinction assures that RTSA is able to pinpoint the areas where systems are inadequate to be implemented.

Second, the findings give the weighting factor of governance its significance. The performance of financial services was relatively higher than that of healthcare and critical infrastructure performance, which was mainly due to the increased scores in governance. This implies that governance compliance not only improves accountability but also has a very significant impact on the entire RTSA outcome. Practically, it is possible to raise the credibility of organizations investing in regulatory alignment and lifecycle management to a considerable degree, despite having vulnerabilities in their systems.

Third, the findings demonstrate the conflict between the risk and trust measures. As an example, the healthcare system scored a relatively high risk and moderate on trust, resulting in a low final RTSA outcome. This conflict illustrates how the framework can achieve trade-offs: an AI system can be technically highly developed but fail to inspire confidence in the absence of fairness or accountability. The opposite is also true, in that strong trust measures can mitigate some of the risks, given the existence of effective governance mechanisms.

Theoretically, the results build upon the previous studies on AI ethics and cybersecurity because they provide a measurable relationship between the abstract values (fairness, transparency, accountability) and the risk indicators in practice. Practically, they propose that AI governance should not be seen as a burden to organizations, but rather as a risk mitigation tool. Lastly, on the policy front, the findings suggest that AI regulation should establish standardized measures of trust, thereby strengthening systems.

Conclusion

To sum up, the RGT framework and RTSA are a well-organized, quantifiable, and versatile way of dealing with the complicated nexus of cybersecurity and artificial intelligence. This research offers a useful tool that will help to responsibly adopt AI by merging risk management, governance control, and trust measures in one assessment model. The introduction of RTSA into the financial services, healthcare, and critical infrastructure segments showed that AI systems in finance showed moderate levels of trustworthiness, whereas in healthcare they showed less favorable but above-average results, with the essential infrastructure showing an inadequate preparedness to be deployed. These findings demonstrate the importance of proportional investment in risk management practices, compliance with governance, and mechanisms that build trust to make sure that AI in cybersecurity can be used in a responsible manner.

Study Contribution: This study has threefold contribution. First, it proves an innovative algorithmic scheme of connecting fragmented factors of AI ethics, governance, and data safety into a comprehensive system of assessment. Second, it demonstrates that measures of trust, which are believed to be abstract, can be converted into measurable indices, thus, becoming operative to the decision-makers. Third, it emphasizes the importance of governance as not only a compliance mandate but also a decisive weighting factor that can enhance the reliability of the system and public trust to a great extent. These contributions collectively further theoretical discussions in the field of AI ethics and operational activities in the field of cybersecurity risk management.

Future Direction: The RGT framework is to be empirically tested in the future research with real time operational data on how the RTSA works in dynamic cyber threat settings. Comparative studies (across countries) are also necessary to investigate the variation of governance weightings across jurisdictions, to inform at least the formulation of global standards. The RTSA should be further refined to introduce advanced trust measures, including explainability quality measures, indices of disparity in fairness, and bias detectors, which will make the RTSA more sensitive. A related promising avenue is concerning the combination of automated governance systems, such as smart contracts or self-auditing AI, to dynamically re-set weights of governance depending on compliance performance. In addition, the human-AI interface should be examined in terms of both quantitative and qualitative outcomes related to RTSA and human trust and perception of the system. Lastly, the impact of standard RTSA-like tools on certification schemes and procurement procedures should be examined in policy-oriented studies that investigate how regulators and other organizations can implement such tools in standardizations to promote credible application of AI.

References

1. A. Habbal, M. K. Ali, and M. A. Abuzaraida, "Artificial Intelligence Trust, Risk and Security Management (AI TRiSM): Frameworks, applications, challenges and future research directions," Expert Systems with Applications, vol. 240, p. 122442, Nov. 2023, doi: 10.1016/j.eswa.2023.122442.
2. R. Bhandari, "AI and Cybersecurity: Opportunities, challenges, and governance," EDPACS, pp. 1–9, Aug. 2025, doi: 10.1080/07366981.2025.2544363.
3. H. Kure I., P. Sarkar, A. B. Ndanusa, and A. O. Nwajana, "Detecting and Preventing Data Poisoning Attacks on AI Models," arXiv.org, Mar. 12, 2025. https://arxiv.org/abs/2503.09302
4. N. H. Simbolon, F. D. Jati, and S. B. Siahaan, "Overcoming the challenges of ai adoption in banking sector: Security, regulation, and infrastructure," ijafibs.pelnus.ac.id, Jun. 2025, doi: 10.35335/ijafibs.v13i1.364.
5. N. Raddatz I. and R. Green, "Navigating Risk and Trust in AI Financial Systems: The Role of Institutional Structures and the EU AI Act," American Accounting Association, Sep. 2025, doi: 10.2308/ISYS-2025-016.
6. S. Karthik and S. Karthik, "What is the RISK Governance Framework [& How to Create It]," Cyber Sierra, Oct. 25, 2024. https://cybersierra.co/blog/what-is-the-risk-governance-framework-how-to-create-it/
7. E. Avgouleas and H. Marjosola, Digital Finance in Europe: Law, Regulation, and Governance. 2021. doi: 10.1515/9783110749472.
8. D. Pöhn, S. Seeber, and W. Hommel, "Combining SABSA and Vis4Sec to the Process Framework IdMSecMan to Continuously Improve Identity Management Security in Heterogeneous ICT Infrastructures," Applied Sciences, vol. 13, no. 4, p. 2349, Feb. 2023, doi: 10.3390/app13042349.
9. M. Mesarčík and O. Hamuľák, "General Data Protection Regulation: Current Challenges and Future Directions," in Contributions to political science, 2024, pp. 117–133. doi: 10.1007/978-3-031-56045-

3_9.
10. Y. Qin et al., “Landslide Susceptibility Assessment in Yulong County Using Contribution Degree Clustering Method and Stacking Ensemble Coupled Model Based on Certainty Factor,” Remote Sensing, vol. 16, no. 19, p. 3582, Sep. 2024, doi: 10.3390/rs16193582.
11. “Application of Algorithmic Cognitive Decision Trust Modeling for Cyber Security Within Organisations,” IEEE Journals & Magazine | IEEE Xplore. https://ieeexplore.ieee.org/abstract/document/9199103
12. A. Agarwal and M. J. Nene, “A five-layer framework for AI governance: integrating regulation, standards, and certification,” Transforming Government People Process and Policy, vol. 19, no. 3, pp. 535–555, May 2025, doi: 10.1108/tg-03-2025-0065.
13. “A Literature Review on Applications of Explainable Artificial Intelligence (XAI),” IEEE Journals & Magazine | IEEE Xplore. https://ieeexplore.ieee.org/abstract/document/10908240
14. M. R. Zafar and N. Khan, “Deterministic Local Interpretable Model-Agnostic Explanations for Stable Explainability,” Machine Learning and Knowledge Extraction, vol. 3, no. 3, pp. 525–541, Jun. 2021, doi: 10.3390/make3030027.
15. G. Siatwiko, “Evaluating the effectiveness of the communication strategy used by the road transport and safety agency (RTSA) in promoting road safety in Lusaka urban.,” 2023. https://dspace.unza.zm/items/78234721-e093-433e-ab5c-456b0167d647
16. D. E. O’Leary, “Verification and validation of case-based systems,” Expert Systems With Applications, vol. 6, no. 1, pp. 57–66, Jan. 1993, doi: 10.1016/0957-4174(93)90018-2.