IBIMA Publishing
Journal of Information Assurance & Cyber security
https://ibimapublishing.com/articles/JIACS/2026/274939/
Vol. 2026 (2026), Article ID 274939, 12 pages, ISSN: 2165-9923
https://doi.org/10.5171/2026.274939

Research Article

From Detection to Trust: Managing Cybersecurity Risks of Generative AI Systems in Organizations

Mariusz ŁAZARSKI

WSB-NLU University in Nowy Sącz**,** Poland
mlazarski@wsb-nlu.edu.pl

Received date:22 December 2025; Accepted date:16 January 2026; Published date: 23 February 2026

Abstract

The rapid adoption of generative artificial intelligence (AI) systems in organizational environments introduces new and complex challenges in the field of information assurance and cybersecurity. While generative AI technologies, including large language models, provide significant benefits in automation, decision support, and knowledge management, they also create emerging risk vectors related to data confidentiality, integrity, system misuse, and regulatory compliance. This paper aims to examine cybersecurity risks associated with generative AI systems from an organizational perspective, with a particular focus on trust, governance, and cybersecurity risk management. The study is based on a structured review of recent academic literature, industry reports, and regulatory frameworks addressing AI security and information assurance. Key threat categories are identified, including prompt injection attacks, unauthorized data disclosure, model manipulation, lack of transparency, and limited auditability of AI-driven systems. The analysis highlights how these risks affect organizational decision-making, accountability, and compliance with data protection regulations. Based on the findings, the paper proposes a conceptual framework for managing cybersecurity risks of generative AI systems that integrates technical safeguards, organizational policies, and compliance mechanisms. The framework emphasizes the transition from reactive detection toward trust-oriented governance and continuous risk monitoring. The paper concludes by outlining future research directions focused on empirical validation of generative AI security controls and their effectiveness in strengthening organizational information assurance.

Keywords: generative artificial intelligence, cybersecurity risk management, information assurance, organizational trust

Cite this Article as: Mariusz ŁAZARSKI (2026), "From Detection to Trust: Managing Cybersecurity Risks of Generative AI Systems in Organizations", Journal of Information Assurance & Cyber security, Vol. 2026 (2026), Article ID 274939, https://doi.org/10.5171/2026.274939

Introduction

The rapid advancement and widespread adoption of generative artificial intelligence (AI) systems have significantly transformed organizational information systems and decision-making processes. Technologies such as large language models and foundation models are increasingly deployed in areas including customer service automation, software development support, data analysis, and knowledge management (Bommasani et al., 2021). While these systems offer substantial operational benefits, their integration into organizational environments introduces new and complex challenges in the domain of information assurance and cybersecurity (Shneiderman, 2020).

Unlike traditional information systems, generative AI systems rely on probabilistic models trained on extensive datasets, often characterized by limited transparency and explainability. Prior research highlights that such characteristics amplify risks related to data confidentiality, integrity, and system misuse, particularly when generative AI is applied to sensitive or regulated organizational data (Bender et al., 2021; Floridi et al., 2018). Emerging studies further indicate that generative AI systems may unintentionally disclose confidential information, generate misleading or biased outputs, or be exploited through adversarial techniques, thereby undermining trust in AI-supported processes (Weidinger et al., 2021).

Existing cybersecurity research has largely focused on technical threat detection and mitigation mechanisms, including intrusion detection systems, malware analysis, and network security controls. However, generative AI systems introduce novel threat vectors that extend beyond conventional detection-oriented approaches. These include prompt injection attacks, unauthorized inference of sensitive data, limited auditability of model behavior, and challenges related to accountability and governance (Zhou et al., 2023; Kroll et al., 2017). As a result, organizations face increasing difficulties in ensuring reliable decision-making, regulatory compliance, and sustained trust in AI-driven systems (Raji et al., 2020).

From an information assurance perspective, the secure adoption of generative AI requires a shift from purely reactive cybersecurity measures toward integrated governance and risk management approaches. International guidelines and standards emphasize the importance of aligning technical safeguards with organizational policies, legal requirements, and continuous monitoring mechanisms throughout the AI system lifecycle (NIST, 2023; ISO/IEC 23894, 2023). In parallel, regulatory frameworks such as the General Data Protection Regulation and the emerging Artificial Intelligence Act underscore the necessity of accountability, transparency, and risk-based governance for AI systems deployed in organizational contexts (European Commission, 2016; European Commission, 2024).

The objective of this paper is to examine cybersecurity risks related to generative AI systems from an organizational perspective, with a particular focus on trust, governance, and cybersecurity risk management. Based on a structured review of academic literature, industry reports, and regulatory frameworks, the study identifies key threat categories and analyzes their implications for organizational information assurance (OECD, 2022; Gillespie, 2020). Furthermore, the paper proposes a conceptual framework that supports the transition from detection-centric security models toward trust-oriented governance and continuous risk monitoring, providing a foundation for future empirical research in this emerging field.

Research Questions and Hypotheses

In order to systematically investigate the cybersecurity and information assurance challenges associated with generative AI systems in organizational contexts, this study is guided by the following research questions and hypotheses.

Research Questions:

What are the primary cybersecurity and information assurance risks associated with the use of generative AI systems in organizational environments?

How do these risks affect organizational trust, decision-making reliability, and compliance with data protection and cybersecurity regulations?

To what extent can integrated governance and risk management approaches mitigate cybersecurity risks related to generative AI systems?

How does the transition from detection-oriented security measures to trust-oriented governance influence organizational information assurance in AI-driven systems?

Based on the reviewed literature and the conceptual foundations of information assurance, the following hypotheses are proposed:

Hypotheses:

The adoption of generative AI systems in organizations significantly increases exposure to novel cybersecurity risks, including data leakage, system manipulation, and reduced auditability.

Higher levels of governance and organizational oversight are positively associated with increased trust in generative AI-driven systems.

Organizations that integrate technical safeguards with organizational policies and compliance mechanisms achieve lower cybersecurity risk levels in generative AI deployments.

Trust-oriented governance approaches are more effective than purely detection-based security strategies in enhancing information assurance for generative AI systems.

Background and Related Work on Generative AI Security

Generative artificial intelligence systems, particularly large language models (LLMs) and foundation models, have rapidly transitioned from experimental technologies to widely deployed organizational tools. These systems are increasingly used to support decision-making, automate content generation, enhance customer interaction, and improve knowledge management processes. Their growing integration into critical organizational workflows has significantly expanded the cybersecurity and information assurance attack surface.

From a security perspective, generative AI systems differ fundamentally from traditional information systems. Unlike rule-based or deterministic software, generative AI models operate on probabilistic mechanisms and are trained on large-scale datasets that may include sensitive, proprietary, or biased information. As a result, concerns related to data confidentiality, integrity, availability, and accountability are amplified. Prior research highlights that the opacity of model behavior, limited explainability, and restricted auditability complicate the identification and mitigation of security incidents involving AI-driven systems (Bender et al., 2021; Bommasani et al., 2021).

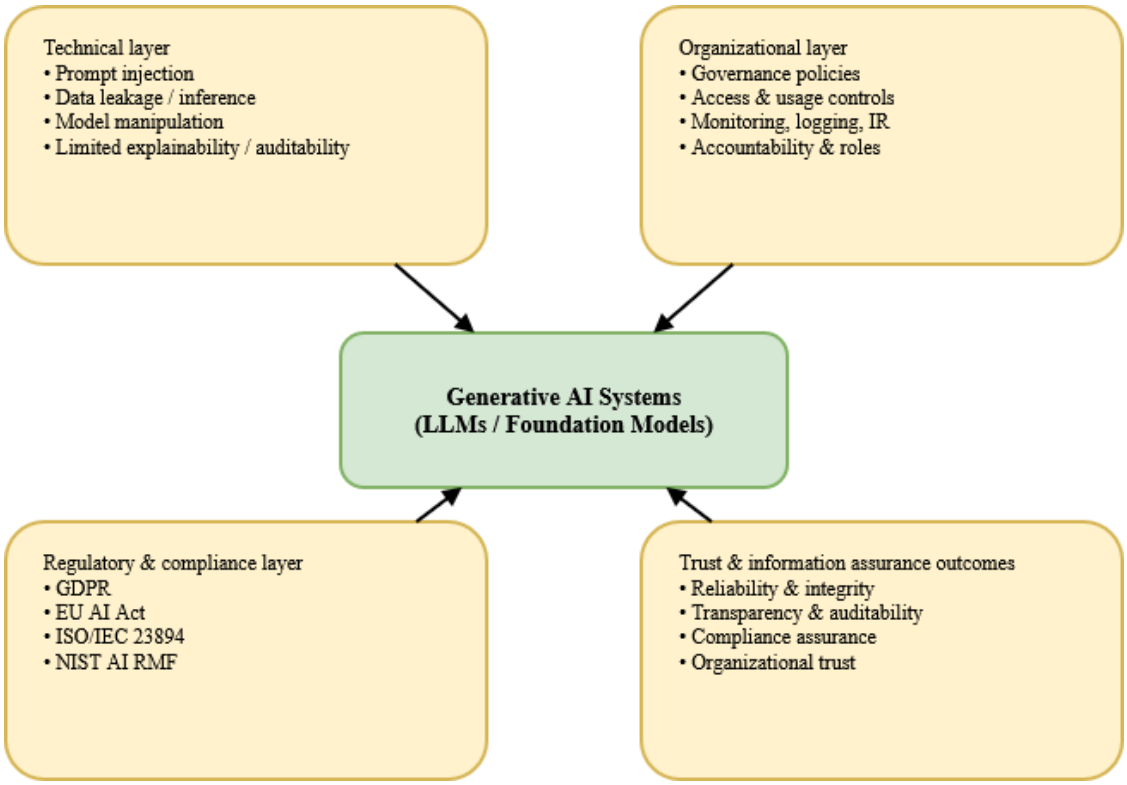

Fig. 1. Conceptual overview of the generative AI security landscape

Source: Author's own conceptual work.

Figure 1 illustrates a multi-layered view of the generative AI security landscape, highlighting how technical vulnerabilities, organizational controls, and regulatory frameworks collectively influence trust, accountability, and information assurance in organizational deployments of generative AI systems.

Existing literature on generative AI security identifies several emerging threat categories. One of the most widely discussed risks is unauthorized data disclosure, where models may unintentionally reproduce sensitive information present in training data or supplied through user prompts. Closely related are prompt injection and prompt manipulation attacks, which exploit the model's instruction-following behavior to bypass safeguards, extract restricted information, or generate malicious outputs (Zhou et al., 2023). These threats challenge traditional access control and input validation mechanisms commonly used in organizational IT systems.

Another critical area addressed in prior studies concerns the lack of transparency and explainability in generative AI models. Researchers emphasize that limited insight into model decision processes undermines trust and complicates accountability, particularly in regulated environments subject to data protection and compliance requirements (Floridi et al., 2018; Kroll et al., 2017). This issue is further compounded by difficulties in logging, monitoring, and auditing AI-generated outputs, which weakens incident response and forensic investigation capabilities.

Governance-oriented research highlights the necessity of integrating security controls with organizational policies and regulatory frameworks. International standards and guidelines, such as the NIST AI Risk Management Framework and ISO/IEC standards on AI risk management, emphasize a lifecycle-based approach that incorporates risk identification, assessment, mitigation, and continuous monitoring (NIST, 2023; ISO/IEC 23894, 2023). Similarly, regulatory initiatives, including the European Union's General Data Protection Regulation and the emerging Artificial Intelligence Act, stress the importance of accountability, transparency, and risk-based governance for AI systems deployed in organizational contexts.

Despite the growing body of literature, existing studies often focus either on technical vulnerabilities or on ethical and regulatory considerations in isolation. Fewer works address generative AI security as a socio-technical challenge that requires the alignment of technical safeguards, organizational procedures, and trust-oriented governance mechanisms. This gap indicates the need for integrated frameworks that move beyond reactive detection and toward proactive risk management and trust assurance, particularly in environments where generative AI systems influence strategic and operational decisions.

Cybersecurity Risk Landscape of Generative AI Systems

The deployment of generative artificial intelligence systems in organizational environments introduces a distinct cybersecurity risk landscape that differs substantially from traditional information systems. These risks arise from the probabilistic nature of generative models, their reliance on large-scale training data, and their increasing autonomy in generating content, recommendations, and decisions. As a result, generative AI systems expand both the technical and organizational attack surface, requiring a reassessment of existing cybersecurity risk management practices.

One of the most prominent threat categories is prompt injection and manipulation attacks, which exploit the instruction-following behavior of large language models. By crafting malicious or deceptive prompts, adversaries may bypass system safeguards, override intended constraints, or induce the model to generate unauthorized or harmful outputs. Such attacks undermine conventional access control mechanisms and challenge assumptions about trusted user input, particularly in systems where generative AI is integrated into automated workflows or decision-support tools (Zhou et al., 2023). The key cybersecurity risk categories associated with generative AI systems in organizational contexts are summarized in Table 1.

Another significant risk involves unauthorized data disclosure and inference. Generative AI models may inadvertently reproduce sensitive information contained in training datasets or reveal confidential organizational data provided during interactions. Even when direct data leakage does not occur, attackers may infer sensitive attributes through systematic querying and output analysis. These risks are particularly critical in sectors handling personal, financial, or proprietary data, where breaches may lead to regulatory non-compliance, financial penalties, and reputational damage (Bender et al., 2021; Bommasani et al., 2021).

A further challenge concerns limited transparency, explainability, and auditability of generative AI systems. The complexity of model architectures and training processes often prevents organizations from fully understanding how specific outputs are produced. This opacity hinders effective monitoring, incident detection, and forensic analysis following security events, and weakens accountability mechanisms in regulated environments. Limited transparency and reporting mechanisms further complicate accountability (Mitchell et al., 2019).

Model misuse and abuse represent an additional risk dimension. Generative AI systems can be repurposed to automate social engineering attacks, generate convincing phishing messages, produce malicious code, or disseminate disinformation at scale. In such scenarios, the AI system itself becomes an enabler of cyber threats rather than a direct target. This dual-use characteristic complicates risk assessment and requires organizations to consider not only how models are protected, but also how their outputs may be exploited (Weidinger et al., 2021).

Table 1. Key cybersecurity risks of generative AI systems in organizational contexts

Risk category	Description	Potential organizational impact
Prompt injection and manipulation	Exploitation of instruction-following behavior to bypass safeguards or induce unauthorized outputs	Loss of control over AI behavior, generation of harmful or misleading content
Unauthorized data disclosure and inference	Unintentional leakage or inference of sensitive training or interaction data	Data breaches, regulatory non-compliance, reputational damage
Model misuse and abuse	Use of generative AI to automate phishing, social engineering, malware creation, or disinformation	Increased scale and effectiveness of cyberattacks
Limited transparency and explainability	Opacity of model decision-making processes and outputs	Reduced trust, weakened accountability, challenges in compliance
Limited auditability and logging	Insufficient monitoring and traceability of AI interactions	Ineffective incident response and forensic analysis
Integration and dependency risks	Weak integration with existing systems or reliance on external AI providers	Expanded attack surface, supply chain and third-party risks

Source: Author's own work.

As reflected in Table 1, risks related to transparency, explainability, and auditability pose significant challenges for organizations deploying generative AI systems. Limited visibility into model behavior complicates monitoring, incident response, and accountability, particularly in regulated environments (Floridi et al., 2018; Kroll et al., 2017).

From an operational standpoint, integration risks emerge when generative AI systems are embedded into existing information infrastructures. Inadequate segregation of privileges, insufficient logging of AI interactions, or weak integration with security monitoring tools may amplify the impact of AI-related incidents. Furthermore, dependencies on external AI service providers introduce supply chain risks, including loss of control over data handling practices and limited visibility into third-party security controls (NIST, 2023; ISO/IEC 23894, 2023).

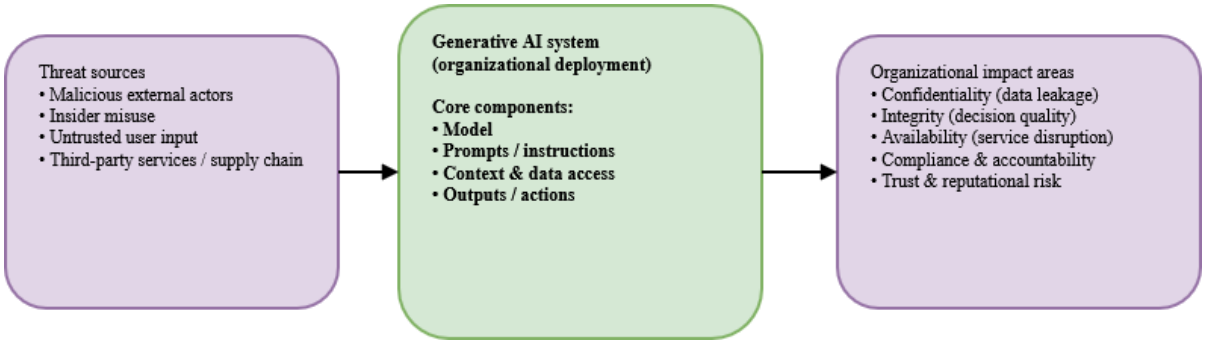

Fig. 2. Conceptual representation of the cybersecurity risk landscape of generative AI systems.

Source: Author's own conceptual work.

Figure 2 illustrates the multi-dimensional nature of cybersecurity risks associated with generative AI systems and their potential organizational impact.

Taken together, these risk categories demonstrate that cybersecurity threats associated with generative AI systems are multi-dimensional and interconnected. They span technical vulnerabilities, misuse scenarios, organizational processes, and governance structures. Addressing this complex risk landscape requires moving beyond isolated technical controls toward holistic risk management approaches that integrate security engineering, organizational oversight, and continuous monitoring. This perspective provides the foundation for examining organizational and regulatory implications and for developing trust-oriented governance frameworks for generative AI systems in subsequent sections.

Organizational and Regulatory Implications

The cybersecurity risks associated with generative artificial intelligence systems have significant organizational and regulatory implications that extend beyond traditional information security management. As generative AI technologies are increasingly embedded in operational workflows, decision-support systems, and customer-facing applications, organizations must adapt their governance structures and compliance mechanisms to address AI-specific threat characteristics.

From an organizational perspective, a central challenge concerns the allocation of responsibility and accountability for AI-driven outcomes. Generative AI systems often operate with varying degrees of autonomy, combining human input with automated content generation and recommendations. This blurring of responsibilities complicates incident attribution and raises questions regarding accountability for security breaches, erroneous outputs, or compliance violations. Prior research highlights that insufficient accountability mechanisms in algorithmic systems may undermine effective governance and trust, particularly in complex organizational environments (Kroll et al., 2017; Raji et al., 2020).

Another important implication relates to the alignment of generative AI deployments with existing cybersecurity and risk management frameworks. Traditional security controls such as access management, logging, and incident response procedures are typically designed for deterministic systems and may not adequately capture the dynamic and context-dependent behavior of generative models. Limited visibility into AI interactions, prompts, and generated outputs can weaken monitoring capabilities, reduce auditability, and hinder post-incident forensic analysis. International guidelines emphasize the need for lifecycle-based risk management approaches that integrate continuous monitoring and governance into AI system operation (NIST, 2023; ISO/IEC 23894, 2023).

Regulatory compliance further intensifies these organizational challenges. Data protection regulations, including the General Data Protection Regulation (GDPR), impose strict requirements related to transparency, data minimization, and lawful processing of personal data. Generative AI systems that process sensitive or personal information must therefore be governed across their entire lifecycle, from data collection and model training to deployment and ongoing operation. Inadequate control over training data, prompts, or AI-generated outputs may expose organizations to legal liabilities, financial penalties, and reputational damage (European Commission, 2016).

In addition to existing data protection laws, emerging regulatory initiatives such as the European Union's Artificial Intelligence Act introduce new obligations for organizations deploying high-risk AI systems. These obligations include requirements for risk assessment, documentation, human oversight, and continuous monitoring of system performance and impact. Compliance with such regulations necessitates a shift from ad hoc or purely technical security measures toward structured governance models that integrate technical safeguards with organizational policies and regulatory controls (European Commission, 2024).

From an information assurance perspective, these organizational and regulatory pressures highlight the critical role of trust in the adoption and sustained use of generative AI systems. Trust is influenced not only by the technical robustness of AI solutions but also by the presence of transparent governance structures, clear accountability mechanisms, and demonstrable compliance

practices. Research on ethical and human-centered AI emphasizes that trustworthiness is closely linked to explainability, accountability, and responsible oversight in socio-technical systems (Floridi et al., 2018; Shneiderman, 2020).

Overall, the organizational and regulatory implications of generative AI cybersecurity risks underscore the need for holistic, trust-oriented approaches that combine security engineering, governance structures, and compliance mechanisms. Addressing these challenges requires moving beyond isolated technical controls toward integrated frameworks that support accountability, regulatory compliance, and continuous risk monitoring. These considerations provide the foundation for the development of trust-oriented governance frameworks for managing cybersecurity risks of generative AI systems, which are presented in the following section.

Managing Cybersecurity Risks of Generative AI

Effectively managing cybersecurity risks associated with generative artificial intelligence systems requires a systematic and proactive approach that extends beyond traditional, detection-focused security controls. Due to the probabilistic nature of generative models, their evolving behavior, and their deep integration into organizational processes, cybersecurity risk management must address technical, organizational, and governance dimensions in a coordinated manner.

From a technical perspective, organizations should adopt security-by-design principles across the entire lifecycle of generative AI systems. This includes controlling training data sources, implementing safeguards for prompt handling, and applying output filtering mechanisms to reduce the risk of unauthorized data disclosure and model misuse. Access control mechanisms should be expanded to govern not only system users but also interactions with the AI model itself, including prompt submission, contextual data injection, and integration with downstream systems. Continuous monitoring of model behavior and generated outputs is essential for detecting anomalies, misuse patterns, and emerging threats that may not be captured by conventional intrusion detection systems (NIST, 2023; ISO/IEC 23894, 2023).

Organizational processes represent a second critical pillar of generative AI cybersecurity risk management. Clear internal policies governing acceptable use, data handling, and human oversight are necessary to reduce ambiguity and prevent misuse. Organizations should explicitly define roles and responsibilities related to AI system operation, security monitoring, and incident response, ensuring accountability for AI-related risks is clearly assigned. Prior research emphasizes that insufficient accountability mechanisms in algorithmic systems may weaken governance and erode trust, particularly in complex socio-technical environments (Kroll et al., 2017; Raji et al., 2020). In addition, targeted training and awareness programs for employees interacting with generative AI systems are essential, as human behavior remains a key factor influencing both risk exposure and mitigation effectiveness.

Governance and compliance considerations further shape the management of generative AI cybersecurity risks. Lifecycle-based risk management approaches stress the importance of continuous risk assessment, documentation, and review, particularly as AI systems are updated, reconfigured, or deployed in new operational contexts (NIST, 2023; ISO/IEC 23894, 2023). Maintaining comprehensive documentation of system design decisions, data sources, risk assessments, and mitigation measures supports transparency and enhances auditability. Such practices are increasingly important for demonstrating compliance with regulatory requirements related to data protection, accountability, and oversight (European Commission, 2016; European Commission, 2024).

A key challenge in managing generative AI cybersecurity risks lies in balancing innovation and control. Overly restrictive security measures may limit the potential benefits of AI adoption, while insufficient controls may expose organizations to unacceptable levels of risk. Effective risk management therefore requires adaptive strategies that combine preventive, detective, and corrective controls. These strategies may include periodic risk reassessment, red-teaming or adversarial testing of AI systems, and integration of AI-specific risks into enterprise-wide risk management and governance processes. Research on ethical and human-centered AI highlights that maintaining trust in AI-supported systems depends on transparency, accountability, and responsible oversight rather than purely technical robustness (Floridi et al., 2018; Shneiderman, 2020).

Overall, managing cybersecurity risks of generative AI systems necessitates a shift from reactive security practices toward proactive, trust-oriented risk management. By integrating technical

safeguards, organizational controls, and governance mechanisms, organizations can better anticipate emerging threats, respond effectively to incidents, and maintain confidence in AI-enabled processes. These management principles provide the foundation for the trust-oriented governance framework for generative AI systems proposed in the following section.

Proposed Trust-Oriented Governance Framework

Building on the identified cybersecurity risk landscape and the organizational and regulatory challenges associated with generative AI systems, this section proposes a trust-oriented governance framework for managing cybersecurity risks of generative artificial intelligence in organizational contexts. The framework supports a transition from detection-centric security models toward proactive, trust-based governance that integrates technical controls, organizational processes, and regulatory compliance mechanisms (NIST, 2023; ISO/IEC 23894, 2023).

The proposed framework adopts a socio-technical perspective, recognizing that trust in generative AI systems emerges from the interaction between technology, organizational governance, and external regulatory environments. Rather than treating cybersecurity as a purely technical function, the framework positions trust as an overarching governance objective that is continuously reinforced through accountability, transparency, and risk management across the AI system lifecycle (Floridi et al., 2018; Shneiderman, 2020).

At the technical layer, the framework emphasizes security-by-design and continuous monitoring of generative AI systems. This includes controls related to data governance, prompt handling, output validation, access management, and anomaly detection. These safeguards aim to reduce exposure to threats such as prompt injection, unauthorized data disclosure, and model misuse, while enabling early detection of abnormal model behavior. Continuous monitoring and logging of AI interactions support effective incident response and forensic analysis, which are central to AI risk management frameworks (NIST, 2023; ISO/IEC 23894, 2023).

The organizational layer focuses on policies, processes, and human oversight mechanisms that shape how generative AI systems are deployed and used. This layer addresses role definition, accountability structures, acceptable-use policies, and employee awareness. Prior research highlights that insufficient accountability mechanisms in algorithmic systems may undermine governance and trust, particularly in complex socio-technical environments (Kroll et al., 2017; Raji et al., 2020). Clearly defined organizational responsibilities are therefore essential for managing AI-related cybersecurity risks and ensuring consistent decision-making.

The regulatory and compliance layer ensures alignment with external legal and normative requirements. This includes obligations arising from data protection regulations and emerging AI-specific legislation, such as requirements for transparency, documentation, and risk assessment. Embedding regulatory considerations directly into governance processes supports accountability and reduces legal uncertainty, particularly in regulated sectors (European Commission, 2016; European Commission, 2024). Figure 3 illustrates the proposed trust-oriented governance framework and its technical, organizational, and regulatory layers, supported by a continuous monitoring and review feedback loop.

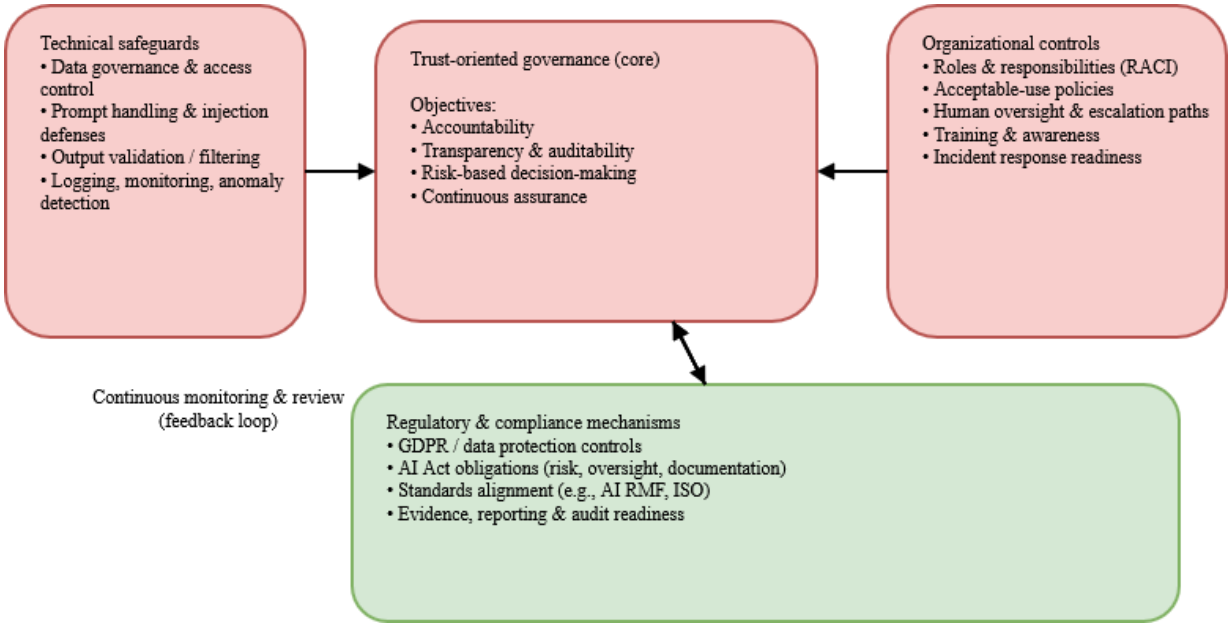

Fig. 3. Trust-oriented governance framework for managing cybersecurity risks of generative AI systems in organizations.

Source: Author's own conceptual work.

A central characteristic of the proposed framework is its lifecycle-oriented and adaptive nature. Trust is treated not as a static property of a generative AI system, but as a dynamic outcome that must be continuously assessed and reinforced as systems evolve. Feedback loops between technical monitoring, organizational oversight, and regulatory review enable ongoing risk reassessment and adaptation to new threats, system updates, or changes in operational context (OECD, 2022).

Overall, the proposed trust-oriented governance framework provides an integrated approach to managing cybersecurity risks of generative AI systems in organizations. By aligning technical safeguards with organizational governance and regulatory compliance, the framework supports accountability, transparency, and sustained trust in AI-enabled processes. This conceptual model offers a foundation for future empirical research aimed at validating the effectiveness of trust-oriented governance mechanisms and refining best practices for a secure and responsible deployment of generative AI systems.

Discussion

The findings of this study contribute to the growing body of research on generative AI security by framing cybersecurity risks not solely as technical vulnerabilities, but as a broader governance and trust challenge within organizational environments. While the existing literature extensively documents specific threats such as prompt injection, data leakage, and model misuse, fewer studies integrate these risks into a coherent organizational and regulatory perspective. This paper addresses this gap by synthesizing technical, organizational, and compliance dimensions into a trust-oriented governance framework.

The analysis confirms that traditional, detection-centric cybersecurity models are insufficient for managing risks associated with generative AI systems. Unlike conventional information systems, generative AI introduces probabilistic behavior, limited explainability, and dual-use capabilities that complicate both prevention and incident response. These characteristics reinforce prior observations that accountability and transparency are central to maintaining trust in AI-driven systems, particularly in regulated and high-risk domains.

From an organizational standpoint, the proposed framework highlights the importance of clearly defined roles, human oversight, and escalation mechanisms. These elements align with prior

research emphasizing that governance failures, rather than purely technical weaknesses, often undermine trust in algorithmic systems. Embedding cybersecurity responsibilities within organizational structures supports consistent decision-making and reduces ambiguity when AI-related incidents occur.

Regulatory and compliance considerations further shape the discussion. The integration of requirements derived from data protection regulations, AI-specific legislation, and international standards demonstrates that compliance should not be treated as an external constraint, but as an integral component of AI governance. Aligning security controls with regulatory expectations enhances accountability and provides organizations with defensible evidence of responsible AI deployment.

The trust-oriented governance framework proposed in this study extends existing AI risk management approaches by explicitly positioning trust as a dynamic and measurable outcome. Continuous monitoring, feedback loops, and lifecycle-based risk reassessment enable organizations to adapt governance mechanisms as generative AI systems evolve. This approach complements existing standards while offering a practical conceptual model that organizations can tailor to their specific operational contexts.

Despite its contributions, this study is subject to limitations. The framework is conceptual in nature and has not yet been empirically validated in real-world organizational settings. Additionally, the rapidly evolving landscape of generative AI technologies and regulations may introduce new risk categories not fully captured in the current analysis. These limitations highlight the need for further empirical research and longitudinal studies.

Conclusion and Future Research Directions

Generative artificial intelligence systems are increasingly embedded in organizational processes, expanding both their strategic value and their cybersecurity risk profile. This paper examined the cybersecurity risks associated with generative AI systems from an organizational perspective, emphasizing the interdependencies between technical vulnerabilities, organizational governance, and regulatory compliance. The analysis demonstrates that effective protection against generative AI–related threats requires moving beyond isolated technical controls toward integrated, trust-oriented governance models.

The proposed trust-oriented governance framework offers a structured approach for managing cybersecurity risks of generative AI systems by aligning technical safeguards, organizational controls, and regulatory mechanisms. By treating trust as a continuous governance objective rather than a static system property, the framework supports accountability, transparency, and informed decision-making across the AI system lifecycle. This perspective is particularly relevant for organizations operating in regulated environments or relying on generative AI for critical decision support.

From a practical standpoint, the framework provides organizations with guidance on how to operationalize AI security governance through continuous monitoring, clear responsibility allocation, and compliance-oriented documentation. It also highlights the importance of addressing dual-use risks, where generative AI systems may inadvertently enable malicious activities if governance mechanisms are insufficient.

Future research should focus on the empirical validation of the proposed framework through case studies, surveys, or experimental deployments in organizational settings. Investigating how trust-oriented governance mechanisms influence risk perception, compliance outcomes, and incident response effectiveness would provide valuable insights. Further studies could also explore sector-specific adaptations of the framework and examine their applicability across different regulatory regimes and organizational maturity levels.

As generative AI technologies continue to evolve, developing robust governance models that integrate cybersecurity, trust, and accountability will remain essential. The framework presented in this paper provides a foundation for advancing both academic research and practical approaches to securing generative AI systems in organizational environments.

References

- Bender, E. M., Gebru, T., McMillan-Major, A. and Shmitchell, S. (2021) ‘On the dangers of stochastic parrots: Can language models be too big?’, *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, pp. 610–623.

- Bommasani, R., Hudson, D. A., Adeli, E. et al. (2021) 'On the opportunities and risks of foundation models', *Stanford Center for Research on Foundation Models*, Stanford University.
- European Commission (2016) *General Data Protection Regulation (GDPR) - Regulation (EU) 2016/679*. [Online]. Available at: https://eur-lex.europa.eu/eli/reg/2016/679/oj [Retrieved 19 December 2025].
- European Commission (2024) Artificial Intelligence Act (AI Act). [Online]. Available at: https://artificialintelligenceact.eu [Retrieved 19 December 2025].
- Floridi, L., Cowls, J., Beltrametti, M. et al. (2018) 'AI4People - An ethical framework for a good AI society', *Minds and Machines*, 28(4), pp. 689–707.
- Gillespie, T. (2020) 'Content moderation, AI, and the question of trust', *Social Media + Society*, 6(2).
- ISO/IEC 23894 (2023) *Information technology - Artificial intelligence - Risk management*. International Organization for Standardization, Geneva.
- Kroll, J. A., Huey, J., Barocas, S. et al. (2017) 'Accountable algorithms', *University of Pennsylvania Law Review*, 165(3), pp. 633–705.
- Mitchell, M., Wu, S., Zaldivar, A. et al. (2019) 'Model cards for model reporting', *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 220–229.
- NIST (2023) *AI Risk Management Framework (AI RMF 1.0)*. National Institute of Standards and Technology. [Online]. Available at: https://www.nist.gov/itl/ai-risk-management-framework [Retrieved 19 December 2025].
- OECD (2022) *OECD Framework for the Classification of AI Systems*. [Online]. Available at: https://www.oecd.org/ai/classification [Retrieved 19 December 2025].
- Raji, I. D., Smart, A., White, R. N. et al. (2020) 'Closing the AI accountability gap', *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, pp. 33–44.
- Shneiderman, B. (2020) 'Human-centered artificial intelligence: Reliable, safe & trustworthy', *International Journal of Human-Computer Interaction*, 36(6), pp. 495–504.
- Weidinger, L., Mellor, J., Rauh, M. et al. (2021) 'Ethical and social risks of harm from language models', *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, pp. 658–668.
- Zhou, Y., Han, X., Xu, C. and Yu, S. (2023) 'Prompt injection attacks against large language models', arXiv preprint arXiv:2302.12173.