Ethical Governance of AI-Driven Information Security: Balancing Data Privacy, Cyber Resilience, and Digital Trust in Organizations

Muhammad Faris bin Nordin[1], Muhammad Din bin Khalid[2], Normal binti Mat Jusoh[3]

[123]Azman Hashim International Business School, Universiti Teknologi Malaysia, Kuala Lumpur, Malaysia

DOI: https://doi.org/10.51244/IJRSI.2026.1307000182

Received: 20 July 2026; Accepted: 25 July 2026; Published: 06 August 2026

ABSTRACT

Cyber threats represent a critical risk to global organisations, with cybercrime damages projected to exceed USD 10.5 trillion annually by 2025. Consequently, enterprises are rapidly adopting Artificial Intelligence (AI) to augment their security architectures, as traditional, rule based Security Information Systems (SIS) struggle to mitigate sophisticated, multi stage attacks. However, the application of AI in security raises significant ethical questions around data privacy, algorithmic transparency, and the balance between automated surveillance and civil liberties. As a conceptual paper, this study investigates how businesses can govern AI-Driven security solutions by bridging the theoretical divide between technical efficacy and ethical responsibility. To build this theoretical foundation, a systematic review of 32 peer reviewed articles was conducted using the PRISMA paradigm, synthesizing existing evidence across four core governance dimensions: technical performance, stakeholder accountability, regulatory compliance, and organisational process. Our conceptual analysis reveals a persistent "principles to practices gap"; while AI based SIS significantly outperform traditional systems in anomaly detection and incident response, these technological advancements have outpaced the operationalization of ethical norms within organisations. To address this gap, the paper proposes a novel, unified governance framework centred on digital trust. This model distinctly integrates the AI Trust Framework and Maturity Model (AI TMM), the Tiered Ethical Cybersecurity Model (TECM), and privacy preserving technologies such as federated learning to operationalize ethics by design. The article concludes with actionable policy pathways for legislators, organisational leaders, and researchers to increase cyber resilience while strictly respecting individual privacy rights.

Keywords: AI-Driven Information Security, Technological Change, Cyber Law, Corporate Social Responsibility/Ethics, IT Management.

INTRODUCTION

The cybersecurity landscape has undergone a radical transformation, shifting from static perimeter defenses to complex architectures necessitated by highly automated and targeted threats. Cybercrime caused approximately USD 6 trillion in damage globally in 2021 (Muhammad et al., 2024) and is anticipated to expand to USD 10.5 trillion each year in the near future, with a 72% rise in coordinated attack activity (Muhammad et al., 2024). In turn, companies are increasingly under pressure to move away from the reactive, rule based security paradigms, and embrace proactive, AI powered solutions. Traditional Security Information Systems (SIS) that rely on signature databases and human analyst evaluation are fundamentally ill equipped to mitigate the speed and complexity of today's threats. Advanced Persistent Threats (APTs) are a particular challenge. They are stealthy multi stage patient attacks within a network for weeks or months, mapping systems and escalating access before doing their most damaging acts (Ali et al., 2025).

Artificial Intelligence (AI) presents a paradigm shifting solution to these sophisticated vulnerabilities. AI incorporated in the SIS constantly monitors behaviour, abnormalities are detected early and incidents are automatically responded to in near real time. Gauhar Ali et al. (2025) showed that XGBoost based classification models could achieve detection accuracy of up to 99.6% with 0% false positive rate which would be difficult to be accomplished in practice using the usual rule based approaches. AI allows organisations to monitor security

activities at a scale that human teams cannot achieve alone, to simultaneously cover thousands of endpoints and users, in addition to pure performance (Shamnad Mohamed Shaffi et al., 2025). However, integrating AI into cybersecurity infrastructures introduces profound ethical challenges; because these systems are inherently data hungry, identifying anomalous conduct necessitates the continuous collection and analysis of massive datasets encompassing user, device, and network activities. However, there is an inherent paradox. The data collecting that enables AI to be effective for security, also has severe privacy difficulties, consent issues and the potential of monitoring going beyond the scope for which it was intended (Mbah and Evelyn, 2024). One of the first to openly emphasise this duality was Zheng Yan et al. (2020), who stressed that AI strengthens security defences while at the same time creates new vulnerabilities and ethical concerns in the systems it protects.

This ethical tension is particularly pronounced within rapidly developing digital economies, with Malaysia serving as a salient illustrative case. The country has spent roughly USD 185 billion on digital infrastructure since 2021 and will continue to do so until 2024, making it one of the top technology hubs in the area. However, only 2% of Malaysian firms are regarded to have a mature cybersecurity posture, according to the Cisco 2024 Cybersecurity Readiness Index (Muhammad et al., 2024). There's also a huge perception gap of 85% of Malaysian firms are confident in their security measures as ransomware and data breaches continue to grow. This mismatch shows a dependence on outdated, reactive strategies ill suited to tackle the current threat environment.

This study posits that the critical missing element is not merely technological advancement, but robust ethical governance. Without ethical guardrails, organisations that use AI powered security technologies risk creating systems that are powerful but unaccountable, effective but invasive, and speedy but opaque. This is the principles to practices gap, as Birkstedt et al. (2023) show: there are many policy level commitments to AI ethics, but these do not translate into day to day operational realities. The main topic of this research is how to bridge this gap. We undertake a Literature Review of 32 peer reviewed sources to synthesise the facts on how AI might boost cybersecurity while maintaining ethical accountability in this study. This paper proposes an integrated governance framework based on the AI Trust structure and Maturity Model (AI TMM), the Tiered Ethical Cybersecurity Model (TECM) and mature privacy preserving technology. The goal is to offer a practical, evidence based roadmap for companies interested in creating robust, trustworthy, and safe AI-Driven environments.

LITERATURE REVIEW

The Shift from Traditional to AI-Driven Security

The limitations of standard SIS have been well documented in recent studies. Legacy systems are built on static rule sets and signature based detection. Such systems are capable of defending against known attacks but are by their nature reactive to new or mutated attack signatures. A primary limitation lies in the reactive, manual nature of traditional SIS, which contrasts starkly with the proactive, adaptive capabilities of AI-Driven architectures (Muhammad et al., 2024). Consequently, the transition toward AI-Driven architectures is not merely an operational upgrade, but a necessary paradigm shift to address highly sophisticated threat vectors.. In their comparative study, they demonstrated that AI-Driven SIS increases anomaly detection rates from 83% to 95% and reduces threat response times from 15 mins to 3 mins, while scaling from 5,000 to 15,000 users and lowering false positive rates from 14% to 6%.

The speed gains are not restricted to a particular system or circumstance. Gauhar Ali et al. (2025) mainly focused on the identification of Advanced Persistent Threats, demonstrating that advanced machine learning classifiers, such as XGBoost, could achieve an accuracy of 99.6% with a 0% false positive rate. Shamnad Mohamed Shaffi et al. (2025) extended this reasoning to cloud settings and discovered that AI in the Zero Trust Architecture helped to decrease the unauthorised access instances by 85%, while deep learning based fraud detection had 97% accuracy. These empirical performance gains across diverse environments ranging from cloud architectures to highly vulnerable healthcare IoT networks underscore the technological imperative of AI adoption. Muhammad Sajid Nawaz et al. (2025) tried three machine learning models on IoT healthcare networks and found out that the Convolutional Neural Network (CNN) beat both Recurrent Neural Networks and Random Forest with 95% accuracy, 94% precision and 93% recall.

The Principles to Practices Gap in AI Governance

One of the most constant themes in the literature is the mismatch between proclaimed ethical values and the reality of governance. Despite a proliferation of high level ethical frameworks, empirical operationalization remains severely underdeveloped, reflecting a structural disconnect between policy formulation and technical deployment. Birkstedt et al. (2023) discovered that whilst there is increased scholarly interest in AI ethics, the vast majority of existing research is conceptual rather than empirical, the literature has a Western bias, and that there are very few real methods for operationalising principles.

McCormack et al. (2026) further explore this topic with structured interviews with 15 industry executives. They found many organisations struggled with unclear ownership of compliance, separate departments and a tendency to prioritise profitability over eliminating bias. The outcome is what the authors refer to as ethical washing, whereby businesses make a formal commitment to responsible AI but do not change their practices in any meaningful way. This finding aligns with the one by Ababneh et al. (2025), which suggests that the cybersecurity infrastructure in developing nations develops quicker than the legal and ethical frameworks designed to regulate it.

From a technical standpoint, Korobenko et al. (2024) addressed this issue by completing an extensive literature assessment of 73 works on AI ethics and discovered that just 7 of these studies gave formal definitions of key ethical concepts. They established a four pillar structure of Data, Technology, People and Process and mapped governance practices against a risk continuum from high to low. Their wider conclusion is that good governance requires not merely ideals but structured and risk aware implementation methods.

Chaudhary (2024) also addressed the issue from a legal perspective, looking at the EU's regulation of algorithmic transparency in the GDPR and the 2024 EU AI Act. This paper offers the idea of qualified transparency, according to which the duty to disclose information should be tailored to different audiences and should not be a "one size fits all" approach. Furthermore, Chaudhary asserts that technical limitations, especially 'decontextualisation,' render absolute transparency practically unattainable.

Privacy, Surveillance, and the Ethical Tension at the Heart of AI Security

The use of AI in security contexts creates what many experts refer to as an inherent ethical tension; comprehensive data collection for threat detection inherently risks evolving into unwarranted surveillance, thereby undermining the privacy rights it aims to defend. Mbah and Evelyn (2024) well understood this contradiction, stating that roughly 88% of data breaches are caused by human error, thus offering a strong organisational incentive for the extensive monitoring. But they also warned that AI surveillance should be governed by the principles of privacy by design and legislative frameworks such as the GDPR and CCPA.

This conflict has led to extensive discussion in the literature of the idea of 'privacy by design' as a possible remedy. Rather of retrofitting privacy as a compliance requirement on systems after they are built, privacy by design incorporates data minimisation, anonymisation and access controls into the architecture of AI systems from the outset. This approach was highly recommended by Emehin et al. (2024) which have demonstrated through case studies in banking and healthcare that AI based systems with differentiated privacy and federated learning may significantly reduce the threat surface without compromising detection efficacy.

Federated learning is one of the most promising technology solutions to the surveillance vs privacy problem. Federated architectures do not centralise sensitive data to train a model but store the data on local devices and only exchange changes in model parameters. Xu et al. (2024) reviewed distributed and decentralised learning processes extensively and concluded that all three approaches, namely federated learning, split learning and swarm learning, solve the problem of GDPR compliance by not centralising data. Saeed and Alsharidah (2024) added to this, noting that differential privacy is a strategy that provides mathematical guarantees of anonymity through the addition of calibrated noise to data sets or queries, so that it is computationally infeasible to reverse engineer individual records.

Barthwal et al. (2025) contribute to this issue with research of 482 participants from three groups: AI professionals, parents, and youth. They found considerable differences in the experience of privacy by different groups in AI environments. Trust in AI systems was substantially greater among AI specialists (4.18 out of 5) than among teenagers (3.09). This conclusion is relevant for governance design in that it suggests trust is not homogenous, and that the design of governance frameworks has to include the diverse expectations and vulnerabilities of different user populations.

Digital Trust as a Governance Outcome

Trust has emerged as a key concept in the AI governance literature, as a social benefit and as a quantifiable governance outcome with tangible organisational implications. Meenalochini (2025) introduced a concept of Digital Trust Engineering that combines AI, cybersecurity and cloud governance. This methodology has been used and has resulted in a 60% reduction in privacy breaches, a 60% increase in decision reliability and a 71% increase in organisational reputation. Consequently, trust transitions from an abstract aspirational ideal to a quantifiable governance metric.

Livia Levine (2019) provided a basic theoretical theory of digital trust, by extending the Integrative Social Contract Theory (ISCT) to online organisations. Her research demonstrated that digital business communities qualify as a true community according to ISCT where the core social norms are trust and cooperation. Although this work was established before many of the AI specific advancements in the field, it provides an important ethical grounding to understanding why trust matters in digital governance circumstances, not only as a management tool but also as a normative commitment.

Xia et al. (2026) empirically demonstrate that responsible AI governance is positively correlated with company performance. The authors utilised a difference in differences regression model on 342 securities businesses from 2016 to 2023 and found that firms that embraced responsible AI governance witnessed large increases in Tobin's Q, a standard measure of market value relative to asset replacement cost. The most influential factors were compliance certification (33.15%), transparency (16.93%) and AI patent applications (17.32%). The conclusion has a vital implication: Ethical AI governance is not just a legal burden but a source of economic advantage and market trust.

Alyileili and Opoku (2026) extended this to the public sector, looking at the impact of governance design on consumer satisfaction with AI enabled municipal services in the UAE. In a mixed methods study of 272 staff and consumers, they showed that perceived helpfulness was a greater predictor of happiness than system speed and that transparency and accountability were crucial for user confidence. In a number of situations, high system responsiveness was negatively related to satisfaction, suggesting that consumers are more concerned with quality of resolution than speed of response. This is an important finding for the design of security governance since the urge to favour automated speed can come at a cost of real human engagement.

Theoretical Frameworks Guiding AI Security Governance

Different theoretical frameworks have been used in the literature to comprehend and guide the management of AI enabled security systems. Most closely comparable maybe is the AI Trust Framework and Maturity Model (AI TMM) by Mylrea and Robinson (2023). The AI TMM accounts for the unpredictability of AI systems in terms of entropy. It gives a maturity rating from 0 to 3 for seven ethical pillars: explainability, data privacy, robustness, transparency, data use and design, societal well being and responsibility. The framework puts ethical governance into practice so that organisations can assess their present practices and pinpoint particular areas for advancement. Most crucially, the authors observed that low explainability maturity leads to both security and productivity losses, indicating that ethics and effectiveness are not separate issues.

Ababneh et al. (2025) suggested the Tiered Ethical Cybersecurity Model (TECM) to solve the same problem with a different view. The TECM is structured on five levels: Technical, Data, Operational, Organisational and Policy. Ethics are incorporated in each level of security practice. The validation survey carried out in Jordan, Saudi Arabia and Malaysia with 500 participants scored well on clarity and utility but lower on adaption to new technologies which is an area to improve in future.

In this study many scholars have utilised the Technology Acceptance Model (TAM) to explain why AI systems that are theoretically superior sometimes may not work in practice. Similarly, perceived transparency and trustworthiness were found to be key for effective adoption of AI technologies (McCormack et al., 2026; Utami & Aimin, 2026). If a system functions effectively but cannot explain why it makes the decisions it does, the human analysts who must use the system will resist it. This finding underscores the significance of explainable AI (XAI) not only as an ethical imperative, but also as a practical necessity for effective deployment.

Conceptual Framework

The Integrated Governance Architecture

This study proposes a conceptual framework that consists of three complementary components: 1) AI Trust Framework and Maturity Model (AI TMM) as the organisational backbone; 2) privacy preserving technologies as the technical infrastructure; and 3) principles of digital trust engineering as the underlying governance philosophy. A mature AI TMM implementation enables the appropriate use of privacy preserving technologies and together they create the digital trust that companies need to sustain stakeholder confidence over time.

The AI TMM operationalizes ethical AI governance by offering a practical, measurable framework rather than abstract philosophical ideals (Mylrea and Robinson, 2023). AI solutions at the lowest maturity stages are used in silos, with few integrations into larger security operations and no formal governance monitoring. As they mature, companies begin to formalise processes, use machine learning for threat assessments and build formal policies for AI risk management. At the highest maturity level, AI systems are not only technically capable, but also aligned with corporate values, regulatory needs and stakeholder expectations (Mylrea and Robinson, 2023). This evolution parallels the organisational learning cycles articulated by George (2025) who advocated a strategic move away from reactive tool acquisition to integrated risk based prioritisation.

Privacy preservation technologies are a convergence of technical and ethical governance. Federated learning helps businesses to construct powerful AI models without centralising sensitive data, tackling the surveillance problem that has made AI adoption controversial in privacy sensitive sectors. Differential privacy delivers demonstrable guarantees that no individual record can be learned from aggregate analysis, enabling robust security analytics at scale, while yet giving strong anonymity guarantees. Saeed & Alsharidah (2024) also identified homomorphic encryption as a supporting technology. Homomorphic encryption is a technique that allows calculations to be performed on encrypted data without decrypting it at any stage. Collectively these solutions operationalise privacy by design without needing companies to sacrifice on security performance.

Meenalochini (2025) conceptualizes digital trust engineering as a rigorous discipline yielding measurable outcomes, transcending its traditional perception as a soft organizational ideal. Organizations that proactively embed trust mechanisms into system architectures consistently outperform those that relegate governance to reactive compliance exercises. In practice, this means that governance frameworks need to be built into the architecture of AI security systems, not added on after deployment.

The Four Governance Dimensions

The proposed approach is grounded in the taxonomy of AI governance research by Birkstedt et al. (2023) and arranged along four governance dimensions: technical, stakeholder, regulatory and process. The levels test different facets of responsible AI deployments in security contexts.

The technological dimension refers to system performance, explainability and robustness. Masud et al. (2024) demonstrate that XAI tools, including LIME, SHAP, and LRP, significantly mitigate false positive rates in anomaly detection by isolating the specific features triggering alarms. This is important not only for accuracy, but also for accountability—if analysts understand why a system picked up an event, they may make better decisions on how to respond. However, the authors also note that intrinsically explainable models such as decision trees are often less accurate than complex black box models, a genuine trade off that governance systems must grapple with.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VI June 2026

Birkstedt et al. (2023) advised the formation of AI Oversight Units (AOU), which are multi disciplinary teams that integrate technological, legal and ethical knowledge to monitor the conduct of AI systems to ensure that they continue to be aligned with the principles of the organization. This approach is consistent with the Human in the Loop (HITL) principles proposed by McCormack et al. (2026) and Pahune et al. (2025) who argued that high stake assessments should involve significant human validation rather than be entirely delegated to automated systems.

The regulatory dimension positions AI governance in the context of existing and evolving legal systems. Chaudhary (2024) noted that the EU AI Act and GDPR put specific obligations on algorithmic transparency and the right to explain. Formal compliance measures with these requirements are missing in most organisations, which poses a legal risk and a practical governance vacuum (Korobenko et al., 2024). For organisations operating in Malaysia, the national Cybersecurity Act adds another layer of regulatory requirement however as pointed out by Muhammad et al. (2024) the gap between legislative purpose and operational implementation remains considerable.

The process dimension addresses incorporating governance in the organisational processes of the day to day work. George (2025) proposed six strategic priorities for constructing cyber resilience, including (i) architecture integration, (ii) software supply chain security, (iii) SIEM modernisation, (iv) data centric protection, (v) attack surface reduction, and (vi) post quantum cryptography readiness. Furthermore, Khawar et al. (2026) mentioned that the arrival of quantum computing will need companies to start transitioning to post quantum cryptography standards immediately, before the spread of quantum enabled attacks.

Table 1. Traditional SIS vs. AI-Driven SIS: Performance Comparison

Performance Metric	Traditional SIS	AI-Driven SIS	Source
Anomaly Detection Rate	83%	99.6%	Muhammad et al. (2024); Ali et al. (2025)
Incident Response Time	15 minutes	Under 3 minutes	Muhammad et al. (2024)
False Positive Rate	14%	6%	Muhammad et al. (2024)
System Scalability	~5,000 users	15,000+ users	Muhammad et al. (2024)
Fraud Detection Accuracy	N/A	97%	Shaffi et al. (2025)
IoT Intrusion Detection	N/A	95% (CNN)	Nawaz et al. (2025)
Unauthorized Access Reduction	N/A	85%	Shaffi et al. (2025)

Table 2. Governance Framework Dimensions and Tools

Dimension	Key Challenge	Governance Tool	Source
Technical	Black box opacity and adversarial vulnerability	XAI (LIME, SHAP), AI TMM	Masud et al. (2024); Mylrea & Robinson (2023)

Stakeholder	Ethics washing and unclear accountability	AI Oversight Units (AOU), HITL protocols	Birkstedt et al. (2023); McCormack et al. (2026)
Regulatory	Principles to practices gap	GDPR, EU AI Act, TECM	Chaudhary (2024); Ababneh et al. (2025)
Process	Reactive posture and tool fragmentation	SIEM modernization, Zero Trust, CTEM	George (2025); Khawar et al. (2026)

METHODOLOGY

Research Design

The research strategy of this study is Literature Review. The SLR method was selected to provide a systematic, transparent, and reproducible synthesis of multidisciplinary evidence, ensuring the proposed governance framework is grounded in empirical data rather than theoretical assumptions. This study was performed according to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta Analyses) approach to guarantee methodological rigour through a transparent process of identification, screening, eligibility assessment and inclusion.

This SLR approach aligns with recent methodological practice in AI governance research. Birkstedt et al. (2023) conducted a systematic literature review (SLR) based on PRISMA to synthesise 68 papers on organisational AI governance. Utami and Aimin (2026) employed similar methods to synthesise 59 publications on AI-Driven customer experience management. The SLR in this work allows for a systematic analysis of the intersection of evidence on AI security performance with the governance and ethical literature to produce insights that are not generated by each strand of research alone.

Search Strategy and Source Selection

Academic literature was searched in three main databases: Scopus, IEEE Xplore and Web of Science. The reason for choosing these databases is that they encompass both technical computer science and social science literature, reflecting the interdisciplinary nature of AI governance research. The search terms included combinations of the following: AI-Driven cybersecurity, Security Information Systems, ethical AI governance, data privacy, digital trust, cyber resilience, federated learning, explainable AI and organisational AI governance.

Sources should be peer reviewed and published in indexed journals or conference proceedings and directly connected to at least one of the four governance elements of the study and available in full form. We exclude sources if they are not peer reviewed, if they are technical without governance consequences, or if they were published before 2019 (with the exception of foundational theoretical work such as Livia Levine's (2019) work on digital trust). We concentrated on Q1 and Q2 indexed literature to ensure the evidence base was of established criteria of academic excellence.

Analytical Approach

The literature was evaluated using a thematic analysis framework proposed by Braun and Clarke which provides a structured, yet flexible, approach to the identification of patterns across qualitative and mixed methods research. Thematic coding was conducted iteratively, undergoing multiple review cycles to maximize internal coherence and maintain strict fidelity to the source material.

This method yielded four key themes that were consistent with the four governance elements included in the conceptual framework: technical performance and explainability, stakeholder responsibility and oversight, regulatory compliance and transparency, and organisational process and resilience. To ensure that the framework

reflected the true complexity of the evidence, and did not provide an imbalanced impression, for each theme supporting and contrasting evidence was documented.

Variable Framework

For the operationalisation of the research issues the structured variable framework is introduced in the study. Independent variables comprise AI embedded technologies within the SIS, including machine learning algorithms, deep learning architectures, and explainable AI tools. The dependent variables are the governance and security outcomes: accuracy of detection, time for incident reaction, false positive rates, system scalability, and measures of digital trust and organisational resilience. A mediating variable that describes how governance structures (such AI Oversight Units, HITL protocols, privacy preserving architectures) transform AI capacity into ethical and effective security practice. The three layer paradigm is based on the understanding that the application of AI technology does not automatically lead to beneficial governance outcomes, but that the impact of AI depends on the organisational and regulatory framework in which it is utilised.

RESULTS AND ANALYSIS

AI Performance: Compelling Evidence with Important Caveats

The proof of AI performance in security contexts is remarkable and consistent across several studies and sectors. The context of governance was benchmarked by Muhammad et al. (2024) who show that AI-Driven SIS improve anomaly detection from 83% to 95% and response time from 15 to 3 minutes. This finding was subsequently strengthened and enlarged by Ali et al. (2025) whose XGBoost model achieved 99.6% detection accuracy against APTs, the most sophisticated and hazardous kind of threat. Similarly, Shamnad Mohamed Shaffi et al. (2025) showed that Zero Trust Architecture with AI resulted in 85% reduction in the number of unauthorised access in cloud settings.

Mbah and Evelyn (2024) presented case studies of genuine AI implementations, resulting in a 70% decrease in fraud instances and a 90% decrease in ransomware cases in hospitals. Meenalochini's (2025) Digital Trust Engineering strategy achieved a measurable improvement in organisational reputation and a 60% reduction in privacy violations. Alsulami (2024) showed that the AI based detection model that utilised Artificial Fish Swarm driven Weight normalized Adaboost (AF WAdaBoost) attained an accuracy of 98.69% on benchmark datasets, surpassing traditional machine learning techniques like Multi Layer Perceptron and Random Forest.

However, the literature cautions against an uncritical interpretation of these performance metrics. XAI visualisation techniques are sensitive to slight perturbations in the input which indicates that the interpretations produced may not be resistant to small changes" (Masud et al., 2024). Even though behavior based AI models are exceedingly accurate, they are not immune to adversarial attacks that might poison training data or evade detection algorithms (Andreea Dinu et al., 2024). The results of this study suggest that excellent governance does not decrease in relevance with high performance measures, but rather boosts it.

Perhaps the most realistic assessment of the hurdles of AI adoption comes from Alnaabi and Al Mahruqi (2026) using survey data from 138 practitioners in Oman. The AI's institutional knowledge was high (4.22 out of 5), but the actual advantages were somewhat lower (3.85 out of 5). The biggest barriers were skills shortages (87%), legacy infrastructure (73%) and regulatory uncertainties (67%). The results strongly contradict the idea that adoption of AI will automatically result in higher security performance, and indicate the organisational and structural characteristics that mediate performance.

The Governance Gap: Where Principles and Practice Diverge

In addition to the performance evidence, the literature also shows a recurrent and disturbing pattern: organisations that make formal pledges to ethical AI governance typically fail to live up to those claims in practice. Birkstedt et al. (2023) found in their SLR that 61 of 68 papers were conceptual and not empirical, meaning that the majority of the AI governance research is addressing what should happen, rather than what

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VI June 2026

happens. This discrepancy permits organizations to publicly endorse AI ethics while internally deploying systems that fail to meet these normative standards.

McCormack et al. (2026) gave the topic a particular organisational texture in industry interview research. Governance gaps noted by respondents included the lack of clarity about who owns AI compliance duties, the inclination to consider AI ethics as a marketing instead of an operational issue, and financial imperatives that always trump bias reduction efforts. These are not one time events. Ababneh et al. (2025) examined cybersecurity practitioners in Jordan, Saudi Arabia and Malaysia and found similar patterns, with the lowest scoring attribute they measured being the adaptation of ethical frameworks to new technology.

There is also a similar problem for the governance of large language models that Pahune et al. (2025) call data opacity, when companies do not have enough insight into how their AI systems are using data, even if they nominally claim to be governing it. The result is a contradiction: the governing structures exist on paper, but cannot work well in practice, because the systems below are too complicated and opaque to be supervised in any meaningful way. This is partially a technological hurdle. Chaudhary (2024) highlights decontextualisation as a barrier to meaningful transparency, but also an organisational problem inherent in how AI development processes are structured.

Privacy Preserving Technologies: Promising but Unevenly Adopted

Although privacy preserving technologies offer practical resolutions to the surveillance privacy dichotomy, their organizational adoption remains notably inconsistent. Federated learning has been emphasised greatly. Xu et al. (2024) presented the most complete technical evaluation of federated learning, split learning, swarm learning, and hybrid techniques such as split federated learning (SFL), which combines the resource efficiency of split learning and the parallelism of federated learning. Their examination of more than 170 publications confirmed that such approaches work in solving the problems of GDPR compliance by not collecting data centrally.

Saeed & Alsharidah (2024) also included a review of partner technologies, namely differential privacy and homomorphic encryption. Differential privacy gives mathematical guarantees that are especially valuable in areas with a high risk of de anonymization, such as healthcare and financial services. Homomorphic encryption takes this a step further, enabling computation on encrypted data without decryption, however the authors noted significant scale challenges currently limiting its practical implementation.

Taddese et al. (2025) looked at data governance in AI powered genomics research and found that although federated learning and blockchain based data integrity mechanisms existed in theory, their implementation was limited by infrastructure gaps and absence of clear organisational mandates for privacy preserving practice. A similar finding was reported in the case studies of Emehin et al. (2024): organisations tend to deploy privacy preserving technologies under regulatory duress, rather than as a proactive governance decision, constraining the extent such tools are embedded in security architectures.

Organizational and Strategic Dimensions of Cyber Resilience

George (2025) gave the most comprehensive strategic blueprint for resilience. He used Gartner research to identify six architectural priorities, including moving from point solutions to integrated security architectures, securing software supply chains with Software Bills of Materials (SBOMs), modernising SIEM platforms, shifting to data centric protection models, reducing attack surfaces with Continuous Threat Exposure Management (CTEM) and beginning the transition to post quantum cryptography.

George also discovered that 60% of CISOs believe macroeconomic volatility to be a big security challenge, and 48% said they don't trust their abilities to analyse AI specific risks. This result is valid for a maturity gap in not only technology deployment but also in organisational risk management competency. Organizations that isolate cybersecurity as a purely technical function exhibit lower resilience compared to those that integrate security considerations into executive strategic planning.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VI June 2026

Khawar et al. (2026) further generalised this research to explore the future of quantum computing. Meta analyses have shown that Quantum Support Vector Machines achieved 92% accuracy on benchmark datasets, compared to 87% for standard SVMs. Quantum Neural Networks have the potential to reduce training time by 40%. The authors advised organisations to start moving to post quantum cryptographic standards such as ML KEM and ML DSA soon, as the lead times for enterprise cryptographic migrations are large and quantum enabled cybersecurity solutions are not yet extensively deployed. In addition, Jain et al. (2025) note that decentralised AI and blockchain technology provide further resilience benefits at the level of organisational governance by providing tamper proof record keeping and reducing the risk of majority abuse in corporate decision making.

DISCUSSION

Rethinking the Ethics Performance Trade Off

Across both practice and academic debate, there is a shared perception of a trade off between ethical governance and AI performance, that increasing the transparency, accountability and privacy respect of AI systems will inevitably come at a cost to speed, accuracy or scalability. The evidence reviewed in this study contradicts that assumption. Organisations that have implemented privacy protecting technologies such as federated learning have not seen a significant drop in performance. Indeed, hybrid strategies such as split federated learning enhance efficiency and parallelism (Xu et al., 2024). Meenalochini (2025) showed that applying a governance first strategy to digital trust engineering brings measurable improvements in terms of security outcomes and reputation.

In a quasi-experiment, Xia et al. (2026) found that responsible AI governance improves Tobin's Q, which is probably the best proof that ethics and performance are complementing rather than competing. Good AI governance promotes stakeholder confidence, reduces regulatory risk, and nurtures the institutional reliability that markets value at higher valuations. This re framing has substantial implications for the way in which organisations think about governance investments. They need to consider ethical compliance not as a cost center but as a source of competitive advantage and long term resilience.

That said, there are actual trade offs in the literature that governance models have to face up to honestly. Masud et al. (2024) found a true accuracy explainability trade off: innately interpretable models such as decision trees are easier to explain, but they are less accurate than sophisticated deep learning models. Saeed and Alsharidah (2024) found the scalability restrictions of homomorphic encryption that now hinder its practical use. These trade offs are not an argument against ethical government. They are an argument for a nuanced context sensitive governance design that consciously decides where to accept trade offs and why.

Building Human Centered AI Governance in Security Contexts

One of the most obvious indications from the literature is that effective governance of AI in security contexts requires substantial human participation at all levels, from the design of systems to operational decision making. As noted by McCormack et al. (2026) and Pahune et al. (2025), the Human in the Loop (HITL) idea is not just a risk management measure. It's a design philosophy that acknowledges the irreplaceable importance of human judgement in situations where automated judgements have real world consequences for real people.

Birkstedt et al. (2023) propose AI Oversight Units (AOU) as an institutional vehicle to translate this theory into organisational practice. An AOU is not merely a compliance position. It is a cross functional team with the ability and power to monitor the behaviour of AI systems, evaluate anomalous outputs and intervene when systems make conclusions that are inconsistent with the organisation's values or regulatory requirements. This importance of such institutional monitoring was further supported by a study on the public sector by Alyileili and Opoku (2026) which found that governance antecedents significantly influenced users' trust in AI enabled services.

The training dimension of human centred governance deserves special consideration. Alnaabi and Al Mahruqi (2026) identified a dearth of skilled workers as the main challenge for the successful application of AI in cybersecurity, with 87% of the practitioners they surveyed reporting this as a primary challenge. Slesinger et al.

(2024) show how co creation approaches, in which marginalised groups are directly involved in the development of AI systems, can improve justice results and promote the wider societal legitimacy necessary for AI systems to work well. They discovered that co creation training gives non expert participants the language to raise concerns about AI bias, which is directly relevant to governance design: educated stakeholders are better positioned to hold AI systems to account.

The Malaysia Context: Lessons for Developing Digital Economies

Malaysia's ambitious digital investment mixed with dangerously low cybersecurity maturity is a pattern that is expected to be repeated in other rising digital economies. The 2% mature readiness figure provided by Muhammad et al. (2024) is not only a technological gap, it is a governance gap, it is a skills gap, it is a structural gap in how cybersecurity is defined and prioritised at the corporate and national levels.

The literature identifies a variety of concrete pathways to address issue. In the similar context of Oman, Alnaabi and Al Mahruqi (2026) proposed the phased adoption framework as part of their Oman Cybersecurity AI Framework (OCAIF) that prioritises the AI adoption investments based on the organisational readiness rather than assuming the entire sector can adopt AI at the same time. Regulatory sandboxes, public private collaborations and targeted skills development programmes are all cited as facilitators to accelerate the introduction of safe AI without creating governance shortcuts.

For Malaysia in particular, a national Cybersecurity Act is a necessary foundation. But it is interesting to look at a similar situation in South Africa discussed by Mahomed (2025): a robust regulatory basis on paper does not necessarily translate into reality. Research Ethics Committees need upskilling Compliance ownership must be clearly allocated and practical instructions for implementation must accompany legal duties. This paper sets out a governance structure that is flexible enough to apply to this kind of regulatory and organisational situation and provides a systematic route from policy commitment to operational practice.

Practical Implementation Challenges, Case Illustrations, And Alignment with Established Standards

A key limitation identified across the literature is that the implementation of AI-Driven security governance frameworks is often constrained by organisational readiness, budget availability, technical complexity, and workforce capability. Although AI enabled Security Information Systems deliver significant improvements in threat detection and response, the deployment of governance mechanisms such as AI Oversight Units, Explainable AI controls, federated learning infrastructures, and continuous auditing programmes requires substantial investment. Organisations frequently encounter barriers including legacy system integration challenges, shortages of cybersecurity and AI specialists, uncertainty regarding regulatory compliance obligations, and resistance to organisational change. These implementation constraints partly explain the persistent gap between ethical principles and operational practice discussed throughout the literature.

Practical evidence from healthcare, cloud computing, and financial service environments illustrates both the value and challenges of AI-Driven cybersecurity governance. In healthcare IoT environments, AI based intrusion detection systems have demonstrated improved detection accuracy while requiring strict privacy safeguards because of the sensitivity of patient information. Similarly, cloud service providers applying AI enabled Zero Trust Architecture have reported substantial reductions in unauthorised access incidents, although successful deployment depended on continuous monitoring, staff training, and governance oversight. These examples suggest that technological performance alone is insufficient; successful adoption depends on aligning security objectives with privacy protection, accountability mechanisms, and organisational capabilities.

The proposed framework may also be evaluated against established cybersecurity governance standards. The stakeholder and process dimensions align closely with the governance and risk management principles contained in the NIST Cybersecurity Framework 2.0, particularly the Govern, Identify, Protect, Detect, Respond, and Recover functions. Likewise, the framework complements ISO/IEC 27001 requirements through its emphasis on risk based controls, continual improvement, accountability, and information security management processes. However, unlike conventional cybersecurity standards, the proposed model explicitly incorporates AI specific concerns, including algorithmic transparency, explainability, bias mitigation, privacy preserving computation,

and digital trust measurement. This integration represents the framework's principal contribution by extending traditional cybersecurity governance toward responsible AI enabled security operations.

To address the study's conceptual nature, future research should empirically validate the framework through multi case studies, surveys, and interviews involving cybersecurity professionals, organisational leaders, auditors, regulators, and policymakers. Such empirical investigations could assess implementation maturity, adoption barriers, governance effectiveness, stakeholder trust, and cybersecurity outcomes across different industries. Longitudinal studies would further help determine whether organisations adopting higher levels of AI governance maturity achieve sustained improvements in cyber resilience, privacy protection, and digital trust over time.

CONCLUSION AND RECOMMENDATIONS

Summary of Findings

A systematic review of 32 peer reviewed publications yields four principal conclusions. First, AI powered SIS demonstrate substantial operational superiority over traditional security architectures. There is clear evidence of increased detection, faster reaction and greater scalability across numerous sectors and contexts. Second, there is an ongoing disconnection between professed ethical objectives and the practice of governance, motivated by organisational incentives, unclear accountability metrics and the opacity of the AI systems themselves. Third, privacy preserving technologies such as federated learning and differential privacy offer practical answers to the issue of surveillance vs privacy without compromising security performance. Fourth, digital trust is not a soft value but a quantitative governance outcome with strong implications for organisational resilience and competitive position.

The Proposed Governance Framework

In this study we propose an integrated governance framework which comprises the AI TMM, the TECM and the principles of digital trust engineering within a four dimensional framework of technological, stakeholder, regulatory and process governance. This paradigm departs from existing techniques in two important ways. First, technological performance and ethical accountability are viewed as complementing rather than conflicting objectives, providing opportunities to negotiate real trade offs rather than denying their existence. Second, we aim to make it scalable across different levels of organisational maturity so that organisations in the earliest stages of AI adoption are able to engage with governance principles in a manner that is appropriate for their current capabilities and resources.

Recommendations

Organisational leaders must prioritize establishing cross functional AI Oversight Units with definitive authority over AI behavior, underpinned by HITL protocols that explicitly delineate the boundaries between automated actions and human mandated security decisions. Privacy protecting technologies should be built into the design, not tacked on as compliance afterthoughts, with federated learning and differential privacy prioritised for deployment in privacy sensitive industries.

For policymakers, the evidence strongly necessitates regulatory frameworks that transcend abstract principles, offering actionable implementation guidance, independent audit mechanisms, and targeted compliance support for small and medium enterprises. The Malaysian Cybersecurity Act needs to be supported with sector specific recommendations for implementation, targeted skills development programmes and public private collaborations that reduce the cost of responsible AI adoption for resource constrained organisations.

The gap in the existing academic literature is the absence of longitudinal empirical studies that follow the implementation of AI governance frameworks over time in real organisational contexts. Longitudinal studies that track organisations over many years of AI governance adoption, measuring the co evolution of stakeholder trust, regulatory compliance and security outcomes evolve together, would provide the evidence base needed to refine and validate frameworks like the one proposed here. Future research should also engage with the emerging

challenges of quantum computing, agentic AI, and autonomous decision making systems, which are likely to introduce governance challenges that existing frameworks were not designed to address.

REFERENCES

1. Ababneh, J., Alnemari, M. M., Attar, H., Alghamdi, H. A., Al Shawaheen, M., Saqarat, B., Abu Romman, L., & Iqtait, M. (2025). Cybersecurity Ethical Aspects (CEA). *IEEE Access*. https://doi.org/10.1109/ACCESS.2025.3641180
2. Ali, G., Shah, S., & ElAffendi, M. (2025). Enhancing cybersecurity incident response: AI-Driven optimization for strengthened advanced persistent threat detection. *Results in Engineering*, 25, 104078. https://doi.org/10.1016/j.rineng.2025.104078
3. Alnaabi, A. H, & Al Mahruqi, A.H. (2026). Artificial intelligence methods for enhancing cybersecurity in Oman: a comprehensive review. *Journal of Cloud Computing*, 15:48. https://doi.org/10.1186/s13677 026 00860 2
4. Anil, V. K. S., & Babatope, A. B. (2024). The role of data governance in enhancing cybersecurity resilience for global enterprises. *World Journal of Advanced Research and Reviews*, 24(01), 1420–1432. https://doi.org/10.30574/wjarr
5. Barthwal, A., Campbell, M., & Shrestha, A. K. (2025). Privacy Ethics Alignment in AI: A Stakeholder Centric Framework for Ethical AI. *Systems*, 13, 455. https://doi.org/10.3390/systems13060455
6. Birkstedt, T., Minkkinen, M., Tandon, A., & Mäntymäki, M. (2023). AI governance: themes, knowledge gaps and future agendas. *Internet Research*, 33(7), 133 167. https://doi.org/10.1108/INTR 01 2022 0042
7. Chaudhary, G. (2024). Unveiling The black box Bringing Algorithmic transparency to AI. *Masaryk University Journal of Law and Technology*, 18(1), 95 122. https://scholar.google.com/scholar?q=Unveiling+The+black+box+Chaudhary
8. Dinu, A., et al. (2024). AI-Driven solutions for cybersecurity comparitive analysis and etical aspects. *Revista Română de Informatică și Automatică (Romanian Journal of Information Technology and Automatic Control)*, 34(3), 35 48. https://doi.org/10.33436/v33i4y202304
9. Emehin, O., Emeteveke, I., & Akanbi, I. (2024). Enhancing Cybersecurity with Safe and Reliable AI: Mitigating Threats While Ensuring Privacy Protection. *International Journal of Computer Applications Technology and Research*, 13(11), 1 12. https://doi.org/10.7753/IJCATR1311.1001
10. George, A. S. (2025). Cyber Resilience in an AI-Driven World: A Strategic Framework. *Partners Universal Innovative Research Publication (PUIRP)*, 3(6), 86 118. https://doi.org/10.5281/zenodo.18002783
11. Iqbal, A. (2024). Digital Trust and Governance: Navigating Privacy and Innovation in the Data Driven Era. *Journal of Management and Social Sciences Review*, 2(2), 40 48. https://scholar.google.com/scholar?q=Ayesha+Iqbal+Digital+Trust+Governance
12. Jain, G., Singh, H., & Hashimy, L. (2025). Enhancing corporate governance with decentralized AI: a structuration theory perspective. *Journal of Science and Technology Policy Management*. https://doi.org/10.1108/JSTPM 10 2024 0425
13. Jeon, K. H. (2026). Limitations and mitigation strategies for using generative artificial intelligence in medical writing: a narrative review. *Journal of the Korean Medical Association*, 69(2), 118. https://doi.org/10.5124/jkma.25.0163
14. Kanchi, M., & Mittal, A. (2026). Trust, Privacy, and Risk Perception in AI Enabled Cyber Security Systems. *5th International Conference on Innovative Practices in Technology and Management (ICIPTM)*, 11465835. https://doi.org/10.1109/ICIPTM69057.2026.11465835
15. Khawar, M., Khalid, S., Rehman, M. U., Usman, A., Al Malwi, W., & Asiri, F. (2026). Shaping the future of cybersecurity: The convergence of AI, quantum computing, and ethical frameworks for a secure digital era. *Computer Science Review*, 60, 100882. https://doi.org/10.1016/j.cosrev.2025.100882
16. Korobenko, D., Nikiforova, A., & Sharma, R. (2024). Towards a Privacy and Security Aware Framework for Ethical AI: Guiding the Development and Assessment of AI Systems. *25th Annual International Conference on Digital Government Research (DGO 2024)*, 742 753. https://doi.org/10.1145/3657054.3657141

17. Koukaras, C., Hatzikraniotis, E., Mitsiaki, M., Koukaras, P., Tjortjis, C., & Stavrinides, S. G. (2025). Revolutionising Educational Management with AI and Wireless Networks: A Framework for Smart Resource Allocation and Decision Making. *Applied Sciences*, 15, 5293. https://doi.org/10.3390/app15105293
18. Levine, L. (2019). Digital Trust and Cooperation with an Integrative Digital Social Contract. *Journal of Business Ethics*, 160, 393–407. https://doi.org/10.1007/s10551 019 04201 z
19. Masud, M. T., Keshk, M., Moustafa, N., Linkov, I., & Emge, D. K. (2025). Explainable Artificial Intelligence for Resilient Security Applications in the Internet of Things. *IEEE Open Journal of the Communications Society*, 6. https://doi.org/10.1109/OJCOMS.2024.3413790
20. Mbah, G. O., & Evelyn, A. N. (2024). AI powered cybersecurity: Strategic approaches to mitigate risk and safeguard data privacy. *World Journal of Advanced Research and Reviews*, 24(03), 310 327. https://doi.org/10.30574/wjarr
21. Muhammad, M. H. B., Abas, Z. B., Ahmad, A. S. B., & Sulaiman, M. S. B. (2024). AI-Driven Security: Redefining Security Informations Systems within Digital Governance. *International Journal of Research and Innovation in Social Science (IJRISS)*, 8(9). https://doi.org/10.47772/IJRISS
22. Mylrea, M., & Robinson, N. (2023). Artificial Intelligence (AI) Trust Framework and Maturity Model: Applying an Entropy Lens to Improve Security, Privacy, and Ethical AI. *Entropy*, 25, 1429. https://doi.org/10.3390/e25101429
23. Muhammad Sajid Nawaz, Muhammad Ahsan Raza, Binish Raza, Manal Ahmad and Farial Syed. "AI-Driven Intrusion Detection Systems for Securing IoT Healthcare Networks". International Journal of Advanced Computer Science and Applications (IJACSA) 16.6 (2025). http://dx.doi.org/10.14569/IJACSA.2025.0160647
24. Pahune, S., Akhtar, Z., Mandapati, V., & Siddique, K. (2025). The Importance of AI Data Governance in Large Language Models. *Big Data and Cognitive Computing*, 9, 147. https://doi.org/10.3390/bdcc9060147
25. Saeed, M. M., & Alsharidah, M. (2024). Security, privacy, and robustness for trustworthy AI systems: A review. *Computers and Electrical Engineering*, 119, 109643. https://doi.org/10.1016/j.compeleceng.2024.109643
26. Shaffi, S.M., Vengathattil, S., Sidhick, J.N., & Vijayan, R. (2025). AI-Driven Security in Cloud Computing: Enhancing Threat Detection, Automated Response, and Cyber Resilience. ArXiv, abs/2505.03945.
27. Slesinger, I., Yalaz, E., Rizou, S., Gibin, M., Krasanakis, E., & Papadopoulos, S. (2024). Training in Co Creation as a Methodological Approach to Improve AI Fairness. *Societies*, 14, 259. https://doi.org/10.3390/soc14120259
28. Taddese, A. A., & Chekole, A. (2025). Data stewardship and curation practices in AI based genomics and automated microscopy image analysis for high throughput screening: a scoping review. *Human Genomics*, 19:16. https://doi.org/10.1186/s40246 025 00716 x
29. Utami, R. D., & Aimin, W. (2026). Balancing Personalization, Privacy, and Value: A Systematic Literature Review of AI Enabled Customer Experience Management. Information, 17(2), 115. https://doi.org/10.3390/info17020115
30. Xia, H., Chen, H., Zhang, J. Z., & Kamal, M. M. (2026). Exploring the impact of responsible AI governance on corporate performance: A quasi natural experiment. *Technological Forecasting & Social Change*, 223, 124425. https://doi.org/10.1016/j.techfore.2025.124425
31. Xu, H., Seng, K. P., Ang, L. M., & Smith, J. (2024). Decentralized and Distributed Learning for AIoT: A Comprehensive Review, Emerging Challenges, and Opportunities. *IEEE Access*, 12, 101016 101051. https://doi.org/10.1109/ACCESS.2024.3422211
32. Yan, Z., Susilo, W., Bertino, E., Zhang, J., & Yang, L. T. (2020). AI-Driven data security and privacy. *Journal of Network and Computer Applications*, 172, 102842. https://doi.org/10.1016/j.jnca.2020.102842