International Journal of Innovation and Economic Development
ISSN 1849-7020 (Print)
ISSN 1849-7551 (Online)

URL:
DOI: 10.18775/ijied.1849-7551-7020.2015.121.2001

Volume 12
Issue 1
April, 2026
Pages 7-14

AI in Cybersecurity (2026): A Qualitative Inquiry into Adversarial Intelligence, Defensive Automation, and Governance in Emerging Digital Economies

[1]Stanley Mwangi Chege, [2]Mary Wainaina

[1]Department of Information and Computer Science, School of Science, Catholic University of Eastern Africa, Nairobi, Kenya

Abstract: Artificial Intelligence (AI) has become a foundational capability within modern cybersecurity operations, simultaneously amplifying adversarial capacity and enhancing defensive effectiveness. In 2026, organizations face a structurally altered threat landscape characterized by automated, adaptive, and scalable attacks, alongside growing reliance on AI-enabled security controls and decision-making systems. This study examines AI as a double-edged sword in cybersecurity through a qualitative inquiry grounded in emerging digital economies, with a specific focus on Kenya's mobile-first fintech and financial-services ecosystem. Using document analysis and thematic synthesis of peer-reviewed literature, industry threat reports, and regulatory instruments, the research explores (i) adversarial uses of AI, (ii) defensive AI applications in Security Operations Centres (SOCs), and (iii) governance and compliance challenges arising from automated decision-making. The study positions ISO/IEC 42001 as a unifying AI governance framework capable of operationalizing regulatory requirements under the Kenya Data Protection Act (KDPA). Findings indicate that without structured governance and human oversight, defensive AI may introduce legal, ethical, and operational risks comparable to those posed by adversarial AI. The paper contributes a governance-oriented lens for cybersecurity leaders in emerging digital economies and provides actionable recommendations for integrating AI innovation with accountability, transparency, and resilience.

Keywords: Artificial Intelligence, Cybersecurity, Adversarial AI, Defensive AI, ISO/IEC 42001, Kenya Data Protection Act, Qualitative Inquiry, Fintech Security, AI Governance

1. Introduction

Cybersecurity has entered an algorithmic era in which artificial intelligence increasingly shapes both offensive and defensive capabilities. Traditional threat models assumed human adversaries constrained by time, cost, and scale. By contrast, AI-enabled threat actors can operate autonomously, continuously, and at machine speed, fundamentally altering assumptions underpinning cybersecurity risk management (ENISA, 2024; Microsoft, 2024). At the same time, defenders are integrating AI into detection, response, and resilience functions to cope with escalating threat volumes and operational complexity. In the context of AI, innovation is precursor for resilience and protection from emerging threats (Maxamadumarovich et al., 2012). Beyond technical disruption, the rise of AI in cybersecurity has triggered heightened regulatory and policy scrutiny at national, continental, and global levels. In Kenya, the Office of the Data Protection Commissioner (ODPC) has emphasized accountability, impact assessments, and safeguards around automated decision-making under the Kenya Data Protection Act

(Office of the Data Protection Commissioner [ODPC], 2024; Republic of Kenya, 2019/2022). At the continental level, the African Union Continental Artificial Intelligence Strategy frames trustworthy AI as a prerequisite for digital transformation, financial inclusion, and protection of fundamental rights in Africa's mobile-first economies (African Union Commission, 2024). Globally, the OECD AI Principles establish widely adopted norms emphasizing transparency, robustness, accountability, and human-centered values for AI systems with societal impact (Organisation for Economic Co-operation and Development [OECD], 2019/2024).

For emerging digital economies such as Kenya, where financial services are predominantly mobile-first, these policy expectations intersect directly with cybersecurity practice. Mobile-money platforms, fintech APIs, and dense third-party integrations expand attack surfaces while compressing response windows. Concurrently, regulatory expectations concerning explainability, proportionality, and human oversight increasingly apply to AI-enabled security controls themselves. This convergence underscores the need for integrated governance models that balance adversarial pressure, defensive automation, and regulatory accountability, as conceptualized in the AI Offense–Defense–Governance framework proposed in this study (see Figure 1).

2. Problem Statement

Despite rapid adoption of AI-driven security tools, many organizations lack coherent governance structures to manage AI-related risks. Cybersecurity strategies frequently emphasize technological capability while underestimating the implications of automated decision-making, data integrity, and explainability. In Kenya and comparable emerging digital economies, this gap is exacerbated by mobile-first architectures and evolving regulatory frameworks.

Specifically, organizations face three interrelated problems: (i) adversarial AI is reducing the effectiveness of traditional controls; (ii) defensive AI is being deployed without sufficient transparency, oversight, or accountability; and (iii) regulatory obligations under data protection laws such as the Kenya Data Protection Act are insufficiently integrated into cybersecurity governance. This misalignment exposes organizations to legal, ethical, and operational failure even as they invest heavily in AI-enabled security.

3. Literature Review

3.1 Adversarial AI and the Evolving Threat Landscape

Recent literature documents the rise of adversarial AI, including polymorphic malware, automated reconnaissance, and generative social engineering (ENISA, 2024; Lin, 2025). These capabilities reduce attacker costs and increase scale, shifting the balance of power toward automation. Industry reports further indicate that a significant proportion of contemporary cyber incidents involve some form of AI augmentation (Microsoft, 2024; Verizon, 2025).

3.2 Defensive AI and SOC Transformation

Scholarly and practitioner studies highlight the role of AI in transforming Security Operations Centres through automation, orchestration, and predictive analytics (Bono et al., 2024; Wang et al., 2024). Defensive AI has been shown to reduce alert fatigue, improve detection accuracy, and enhance analyst productivity. AI systems necessarily pose new governance, accountability and security challenges (Obrenovic et al., 2026). However, research also cautions against over-reliance on opaque models that may introduce new failure modes (Giarimpampa et al., 2026).

3.3 AI-Specific Vulnerabilities

Emerging research identifies AI-specific attack vectors, including prompt injection, data poisoning, and model manipulation, which bypass traditional network-based defenses (Lee et al., 2025; Srivastava et al., 2024). These vulnerabilities challenge conventional cybersecurity assumptions and necessitate new governance approaches.

3.4 Governance, Regulation, and AI Standards

The governance literature emphasizes the growing importance of accountability and explainability in AI systems, particularly where automated decisions produce significant effects (Gartner, 2024). As evidenced across broader digital environments, in accordance to credibility theory system trustworthiness is essential for important decision-making (Khudaykulova et al., 2026). ISO/IEC 42001 has been introduced as the first certifiable AI management system standard, aiming to institutionalize responsible AI practices across organizational lifecycles (ISO, 2023). Regulatory frameworks such as the Kenya Data Protection Act further reinforce requirements for impact assessments and safeguards around automated decision-making (Republic of Kenya, 2019/2022).

Figures and Tables

This infographic visualizes the "Double-Edged Sword of AI" concept central to the research. The graphic is anchored by a vertical sword split down the center, dividing the landscape into two opposing forces:

- **Left Side (Adversarial AI):** Represented in red, this side illustrates the offensive capabilities of AI, featuring icons of robot swarms to denote "automated & scalable attacks" and a speedometer symbolizing the amplification of adversarial capacity.
- **Right Side (Defensive AI):** Represented in blue, this side depicts the protective role of AI, using a shield/eye icon for "AI-enabled security controls" and gears to symbolize "defensive automation" in SOCs.
- **Context & Foundation:** The background features a digital map of Kenya, grounding the study in the "mobile-first fintech ecosystem". The base of the graphic shows an interlocking foundation where **ISO/IEC 42001** connects with the **Kenya Data Protection Act (KDPA)**, labeled "Operationalizing Compliance," highlighting the governance solution proposed in the study.
- **Key Finding:** A summary box at the bottom warns that without governance, defensive AI risks are comparable to adversarial risks.

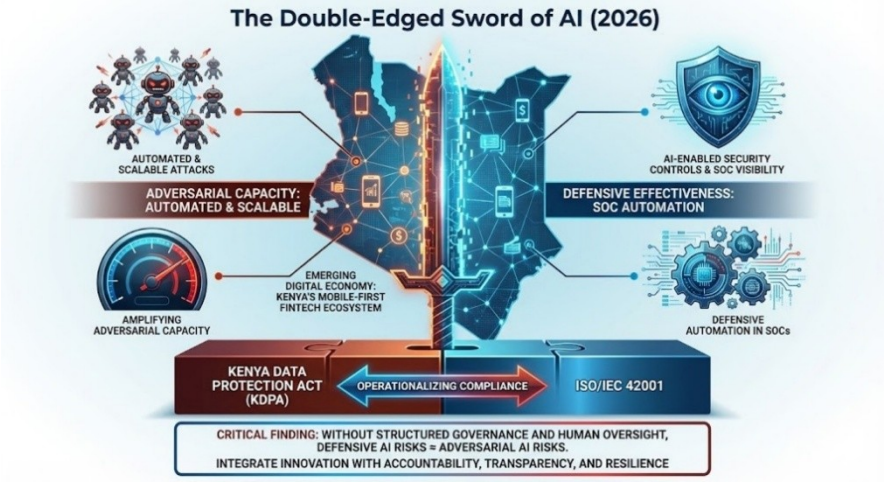

Figure 1:AI as the Double-Edged Sword

4.ResearchMethodology

This study adopts a qualitative research design grounded in interpretive inquiry. A document analysis approach was employed to synthesize insights from peer-reviewed academic literature, authoritative industry reports, and regulatory and standards documentation.

4.1 Data Sources

Primary sources included academic journal articles on AI and cybersecurity, industry threat intelligence reports (e.g., ENISA, Microsoft, Verizon), international standards documentation (ISO/IEC 42001), and Kenyan regulatory instruments and guidance issued by the Office of the Data Protection Commissioner.

4.2 Analytical Approach

Thematic analysis was conducted to identify recurring patterns related to adversarial AI capabilities, defensive AI practices, and governance challenges. Themes were iteratively refined through comparison across sources to enhance credibility and analytical rigor.

4.3 Rigor and Trustworthiness

Rigor was enhanced through triangulation across academic, industry, and regulatory sources, transparent documentation of analytical steps, and alignment with established qualitative research principles. The focus on a clearly defined contextual setting—emerging digital economies with mobile-first infrastructures—supports analytical transferability rather than statistical generalization.

5.Findings

The analysis yielded four key findings. First, adversarial AI has materially altered threat dynamics by compressing attack timelines and reducing reliance on human expertise, increasing asymmetry between attackers and defenders. Second, defensive AI offers demonstrable operational benefits but introduces opacity and governance risks when deployed without oversight. Third, AI-specific vulnerabilities represent a distinct and under-addressed category of cybersecurity risk. Fourth, regulatory and standards-based governance mechanisms remain poorly integrated into operational cybersecurity practices, particularly in emerging economies.

Taken together, these findings support the central proposition of the AI Offense–Defense–Governance model: that effective cybersecurity in the AI era depends not on technological capability alone, but on the alignment between offensive pressures, defensive automation, and governance controls (see Figure 1).

6.Findings

The analysis yielded four key findings. First, adversarial AI has materially altered threat dynamics by compressing attack timelines and reducing reliance on human expertise. Second, defensive AI offers demonstrable operational benefits but introduces opacity and governance risks when deployed without oversight. Third, AI-specific vulnerabilities represent a distinct and under-addressed category of cybersecurity risk. Fourth, regulatory and standards-based governance mechanisms remain poorly integrated into operational cybersecurity practices, particularly in emerging economies.

7.Theoretical and Practical Contributions

7.1 Theoretical Contributions

This study contributes to the cybersecurity and information systems literature in three ways. First, it conceptualizes AI in cybersecurity as a **dual-force phenomenon**, simultaneously enabling offensive capability and defensive resilience. This framing extends existing threat-centric models by explicitly incorporating governance as a co-equal analytical dimension. Second, the research advances theory on **AI governance in emerging digital economies** by contextualizing regulatory compliance, infrastructure characteristics, and organizational maturity within mobile-first environments. Third, by integrating ISO/IEC 42001 into cybersecurity discourse, the study bridges the gap between abstract ethical AI principles and operational management system theory.

7.2 Practical Contributions

From a practitioner perspective, the study provides cybersecurity leaders, regulators, and policymakers with actionable guidance. It demonstrates how ISO/IEC 42001 can be operationalized as a governance backbone for AI-enabled security, reducing legal exposure under the Kenya Data Protection Act while enhancing trust and accountability. The findings further inform CISOs on the necessity of Human-in-the-Loop controls and structured AI risk assessments, supporting defensible decision-making in high-impact scenarios. Insufficiently regulated high-stake systems lead to adverse outcomes (Abueva et al., 2025).

7.3 Implications for Policy and Regulation

The findings have direct implications for policymakers and regulators overseeing AI adoption in cybersecurity and critical digital infrastructure. First, regulators should recognize AI-enabled cybersecurity systems as **high-risk automated processing**, warranting explicit guidance on impact assessments, transparency, and human oversight. Second, alignment between data protection authorities and cybersecurity regulators is essential to avoid fragmented compliance obligations that undermine both security effectiveness and individual rights. Third, the study suggests that management-system standards such as ISO/IEC 42001 can function as regulatory complements, providing auditable evidence of due diligence, accountability, and proportional risk management. For emerging digital economies, embedding such standards into regulatory guidance can accelerate responsible AI adoption while strengthening institutional trust and cross-sector consistency.

8. Recommendations

Based on the findings, the study recommends that organizations: 1. Integrate AI risk assessments into enterprise cybersecurity and risk management processes. 2. Adopt ISO/IEC 42001 as a unifying governance framework linking cybersecurity, legal compliance, and ethical AI. 3. Maintain Human-in-the-Loop controls for high-impact automated decisions. 4. Invest in workforce upskilling to manage AI governance alongside technical security controls. 5. Strengthen collaboration with national cyber response and regulatory bodies.

8.1 Implications for Policy and Regulation (Concise)

This study's findings align closely with existing national, continental, and international policy instruments governing responsible AI adoption.

First, under the **Office of the Data Protection Commissioner (ODPC) of Kenya**, AI-enabled cybersecurity systems should be explicitly classified as *high-risk processing activities* under the Kenya Data Protection Act, thereby triggering mandatory Data Protection Impact Assessments (DPIAs), enhanced transparency obligations, and safeguards against solely automated decision-making. Regulators should issue sector-specific guidance clarifying how AI-driven security controls—such as automated fraud detection or account freezing—must incorporate meaningful human oversight.

Second, the **African Union AI Strategy** emphasizes trustworthy, inclusive, and development-oriented AI. In this context, cybersecurity AI governance should be framed as a digital trust enabler for Africa's mobile-first economies. Policymakers are encouraged to adopt harmonized governance approaches that balance innovation with protection of fundamental rights, particularly in critical sectors such as finance, healthcare, and digital public infrastructure.

Third, the findings operationalize the **OECD AI Principles**, particularly accountability, transparency, robustness, and human-centered values (OECD, 2019/2024). Embedding management-system standards such as **ISO/IEC 42001** into regulatory guidance provides a practical mechanism for translating these principles into auditable organizational practices. For emerging digital economies, such alignment supports regulatory coherence, cross-border interoperability, and institutional trust while avoiding overly prescriptive or innovation-stifling regulation.

9. Conceptual Framework: AI Offense–Defense–Governance Model

Figure 1 here

The conceptual framework presented in this section (see Figure 1) synthesizes the study's theoretical and empirical insights into a unified analytical model.

Figure 1. AI Offense–Defense–Governance Conceptual Framework

Figure 1 illustrates the proposed AI Offense–Defense–Governance framework, which conceptualizes cybersecurity in the AI era as an interaction between three interdependent dimensions: adversarial AI (offense), defensive AI (defense), and AI governance.

The **AI Offense** dimension captures adversarial uses of artificial intelligence, including automated reconnaissance, polymorphic malware, and generative social engineering techniques such as deepfakes and hyper-personalized phishing. These capabilities accelerate attack velocity and scale while reducing reliance on human expertise.

The **AI Defense** dimension represents organizational deployment of AI-enabled security controls, including predictive analytics, behavioral monitoring, security orchestration and automation (SOAR), and self-healing response mechanisms. Defensive AI aims to counter machine-speed threats by improving detection accuracy, response speed, and operational resilience.

The **AI Governance** dimension forms the stabilizing and moderating layer of the framework. It encompasses regulatory compliance (e.g., Kenya Data Protection Act), policy oversight (ODPC guidance), international norms (OECD AI Principles), continental strategy (African Union AI Strategy), and management-system controls operationalized through ISO/IEC 42001. Governance mechanisms ensure transparency, accountability, explainability, and meaningful human oversight over AI-enabled cybersecurity decisions.

The framework posits that sustainable cybersecurity resilience emerges only when defensive AI capabilities are balanced by robust governance structures. Overemphasis on offensive or defensive AI without corresponding governance increases legal exposure, ethical risk, and loss of institutional trust. The model provides a transferable analytical lens for researchers and practitioners evaluating AI-enabled cybersecurity strategies across emerging and developed digital economies.

10. Conclusion

AI has irreversibly transformed cybersecurity into a contest between competing algorithms. While defensive AI provides essential capabilities for managing scale and complexity, it also introduces new categories of risk that extend beyond technical failure to legal and ethical domains. For emerging digital economies, sustainable cybersecurity resilience depends on balancing AI-enabled innovation with structured governance, regulatory alignment, and human judgment. This study demonstrates that ISO/IEC 42001, when aligned with data protection regulation, offers a viable pathway for achieving this balance.

References

- Abueva, N., Buzelo, A., Wu, Y., Turniyazova, Z., Karakushev, D., & Obrenovic, B. (2025). Digital technologies and student mental health: Risks of social media and the promise of virtual reality and autonomous sensory meridian response interventions. *Psychology Research and Behavior Management*, 2179-2191.
- African Union Commission. (2024). *African Union continental artificial intelligence strategy*. African Union. https://au.int/en/documents/202404
- Almeida, G. (2023). Self-healing networks: Adaptive responses to ransomware attacks. *Preprints*. https://www.preprints.org/manuscript/202312.1538
- Bena, N. (2025). Protecting machine learning from poisoning attacks: A risk-based perspective. *Computers & Security*, 137, 103605. https://doi.org/10.1016/j.cose.2025.103605
- Bono, J., Smetters, K., & Williams, J. (2024). *Generative AI and security operations center productivity: Evidence from live operations*. Microsoft Security Research. https://www.microsoft.com/security/blog/
- Chege, S. M. (2024). The adoption of generative AI in Kenya: A critical analysis of opportunities, challenges, and strategic imperatives. *International Journal of Innovation and Economic Development, 10*(2), 46–57.
- ENISA. (2024). *ENISA threat landscape 2024*. European Union Agency for Cybersecurity. https://www.enisa.europa.eu/publications/enisa-threat-landscape-2024

- Gartner. (2024). *Tackling trust, risk and security in AI models (AI TRiSM)*. Gartner Research. https://www.gartner.com/en/articles/ai-trust-and-ai-risk
- Giarimpampa, D., Meier, R., Bissyandé, T. F., Lenders, V., & Klein, J. (2026). Exploring the role of artificial intelligence in enhancing cybersecurity operations: A systematic literature review. *arXiv*. https://arxiv.org/abs/2401.01234
- International Organization for Standardization. (2023). *ISO/IEC 42001:2023 — Artificial intelligence management systems — Requirements*. ISO. https://www.iso.org/standard/81230.html
- Khudaykulova, M., He, Y., Obrenovic, B., Khudaykulov, A., & Abueva, N. (2026). The role of streamer trustworthiness and attractiveness in enhancing viewer engagement and mindfulness in live streaming. *Acta Psychologica*, *264*, 106367.
- Lee, R. W., Kuo, A. M. H., & Chen, Y. (2025). Vulnerability of large language models to prompt injection attacks. *JAMA Network Open, 8*(1), e245678. https://doi.org/10.1001/jamanetworkopen.2025.5678
- Lin, L. S. F. (2025). Organisational challenges in law enforcement's response to deepfakes and related threats. *Laws, 14*(4), 46. https://doi.org/10.3390/laws14040046ž
- Maxamadumarovich, U. A., Obrenovic, B., & Amonboyev, M. (2012). Understanding the innovation concept. *Journal on Innovation and Sustainability RISUS*, *3*(3), 19-26.
- Microsoft. (2024). *Microsoft digital defense report 2024*. Microsoft. https://www.microsoft.com/security/security-insider/threat-landscape
- Obrenovic, B., Asa, A. R., & Oblakovic, G. (2026). The use of ChatGPT in the workplace: a bibliometric analysis of integration and influence trends. *AI & SOCIETY*, *41*(1), 655-668.
- Office of the Data Protection Commissioner (ODPC). (2024). *Guidance note on data protection impact assessment (DPIA)*. https://www.odpc.go.ke/
- Organisation for Economic Co-operation and Development. (2019/2024). *OECD principles on artificial intelligence*. OECD Publishing. https://www.oecd.org/ai/principles/
- Republic of Kenya. (2019/2022). *Data Protection Act, 2019 (Revised Edition)*. Kenya Law Reports. https://new.kenyalaw.org/
- Srivastava, M., Kaushik, A., Loughran, R., & McDaid, K. (2024). Data poisoning attacks in the training phase of machine learning models: A review. *CEUR Workshop Proceedings*, 3910, 1–12. https://ceur-ws.org/
- Verizon. (2025). *2025 data breach investigations report*. Verizon Business. https://www.verizon.com/business/resources/reports/
- Wang, X., Zhang, Y., & Liu, J. (2024). Combating alert fatigue with context-aware methods. *Computers & Security*, 129, 103116. https://doi.org/10.1016/j.cose.2023.103116