https://doi.org/10.53656/math2024-6-7-art

Overview
Обзор

ARTIFICIAL INTELLIGENCE IN CYBERSECURITY: RIGOROUS CRITICAL REVIEW, METHODOLOGICAL CHALLENGES AND FUTURE RESEARCH DIRECTIONS

Maria Mpitsi
Faculty of Mathematics and Natural Sciences
South-West University – Blagoevgrad (Bulgaria)

Abstract. This paper offers a critical review of the applications of AI in cybersecurity, focusing on the recent trends of automation in threat detection, enhancement of response strategies, and prediction of vulnerabilities. The methodology is based on a thorough analysis of empirical studies up to 2021 as per the efficiency of AI malware detection, insider threat identification, and mitigation of zero-day vulnerabilities. In particular, machine learning- and deep learning-based methodologies of artificial intelligence ensure clear advantages over conventional models concerning the precision in detection and reduction of false positives. However, challenges persist regarding explainability, scalability, and ethical concerns around data bias and quality. Finally, this paper concludes by pointing out some areas of future research with regard to needing XAI techniques and methods related to bias reduction to establish better trust in the efficacy of AI-driven cybersecurity frameworks.

Keywords: artificial intelligence; cybersecurity; machine learning; explainable AI; anomaly detection

1. Introduction

Such sophistication in the cyber threat landscape today, including ransomware, APTs, and zero-day attacks, is putting pressure on cybersecurity solutions to become increasingly innovative. The craftier the attacks get, where it adaptively changes its pattern, the more difficult these will be handled by a rule-based system. In this respect, AI, in particular ML/DL, is the game-changer in cybersecurity as it enables real-time anomaly detection, threat prediction, and analysis of big data.

However, various barriers to the full adoption of AI into cybersecurity include, but are not limited to, the opacity of AI models, high implementation costs, and concerns over data quality and bias. This paper reviews recent studies on the role of AI in cybersecurity, up to 2021, looks at the main challenges, and suggests future research directions.

2. Literature review

2.1. AI for threat detection and response

AI-powered systems have revolutionized cybersecurity, allowing automated detection and response mechanisms that are far superior to conventional techniques using a predefined rule-based signature. A proof of an effective AI-powered system that enhanced the detection of phishing attacks, malware, and zero-day vulnerabilities while reducing false positives as high as 20% is presented in (Kaur et al. 2021). The detection of insider threats using AI-powered anomaly detection algorithms is presented in (Dong et al. 2021).

Likewise, (Sarker et al. 2021) highlights deep learning models, especially CNNs, that have been able to achieve higher than 95% accuracy in malware classifications. However, the opacity issue of high-risk sectors like finance remains a challenge.

Table 1. Performance Comparison: Traditional vs. AI-driven Systems (Kaur et al. 2021; Sarker et al. 2021; Dong et al. 2021)

Cybersecurity Task	Traditional Methods (Accuracy)	AI-driven Systems (Accuracy)
Phishing Detection	76%	93%
Malware Detection	81%	95%
Zero-Day Anomaly Detection	65%	88%
Intrusion Detection (IDS)	78%	92%

2.2 Machine learning and deep learning for cybersecurity

Machine learning plays a pivotal role in cybersecurity, with supervised learning excelling in identifying known threats, while unsupervised learning is more effective for novel threat detection. In (Naik et al. 2021) a high accuracy in the supervised malware classification is shown but limitations in handling new threats are noted. In (Shen et al. 2021) the superiority of unsupervised methods in detecting zero-day attacks through anomaly detection is demonstrated.

Deep learning architectures, including recurrent neural networks (RNNs) and generative adversarial networks (GANs), have demonstrated significant potential, especially in the identification of intricate attack patterns (Islam et al. 2021; Wang et al. 2021).

Table 2. AI Techniques and Their Applicability in Cybersecurity (Naik et al. 2021; Islam et al. 2021; Wang et al. 2021; Shen et al. 2021)

Technique	Application	Advantages	Challenges
Supervised Learning	DDoS Detection, Malware Classification	High accuracy with labeled data	Ineffective for novel threats
Unsupervised Learning	Zero-Day Exploit Detection	Effective in detecting unknown threats	Requires large amounts of high-quality data
Recurrent Neural Networks (RNN)	Advanced Persistent Threats	Recognizes complex temporal patterns	Requires large training datasets
Generative Adversarial Networks (GAN)	Anomaly Detection	Creates synthetic data for model training	Vulnerable to adversarial attacks

Table 3. Performance and Explainability of AI Models in Cybersecurity (Doshi-Velez & Kim 2021; Arrieta et al. 2020; Miller 2021)

Model Type	Performance	Explainability	Trust
Standard AI (Deep Learning)	High	Low	Low
Explainable AI (XAI)	Moderate	High	High
Inherently Interpretable Models	Moderate	Very High	Very High

2.3. Explainable AI and Trust Issues

The major barrier to the adoption of AI in cybersecurity is explainability. Most AI models, especially deep learning, are actually "black boxes", making decision transparency tricky. In (Doshi-Velez & Kim 2021) the critical need for explainable AI is identified, especially in an environment where the stakes are high. Attention mechanisms and decision trees are some of the techniques

that embed more lucid insights into AI decisions and engender more trust in them.

In (Arrieta et al. 2020) is investigated XAI for anomaly detection and malware classification. This model gives a very good balance between performance and trustworthiness at some cost in reduced accuracy.

3. Methodological Challenges and Ethical Concerns

3.1. Data Needs and Quality

A significant obstacle encountered by AI-enabled cybersecurity systems is the necessity for extensive quantities of high-caliber data. AI models, especially those utilizing supervised learning methodologies, depend on labeled datasets for their training processes. Nevertheless, acquiring an adequate amount of labeled data for cybersecurity applications proves to be difficult, particularly in instances involving infrequent occurrences such as zero-day attacks.

In (Liu et al. 2021) it was mentioned that anomaly detection models face serious issues of data scarcity, which leads to overfitting or biased results. On the other hand, (Xu et al. 2021) discusses the importance of diverse and representative datasets for constructing AI models with no bias in particular, since this may trigger false alarms or missed novelty threats.

Table 4. Key Ethical and Methodological Challenges in AI-driven Cybersecurity (Liu et al. 2021; Xu et al. 2021; Binns et al. 2021; He et al. 2020)

Challenge	Description	Impact
Data Scarcity	Lack of labeled data limits model training efficacy	Reduces accuracy and generalizability
Model Bias	AI models may inherit biases from training data	Leads to false positives/negatives
Lack of Explainability	Black-box models limit interpretability and trust	Reduces adoption in high-stakes environments
Privacy Concerns	AI systems may inadvertently infringe on user privacy	Raises legal and ethical issues

3.2. Ethical Implications and Bias

Artificial intelligence participating in cybersecurity brings out critical ethics, especially with regard to privacy, surveillance, and algorithmic bias. In (Floridi et al. 2020) it is emphasized that guiding ethical frameworks in AI development would ensure that protection of privacy was protected while adopting AI for strengthening cybersecurity. On the other side, in (Binns et al. 2021) it is pointed out the possibility of bias within AI, especially whenever the training datasets are imbalanced, which could lead to unjust or prejudiced results.

In (He et al. 2020) the problem of bias within adversarial learning frameworks is investigated, demonstrating that training data imbued with bias could result in unequal effects on specific groups or entities. This underscores the necessity for transparency and equity in the training and implementation of AI models, especially in critical cybersecurity contexts.

4. Critical Analysis and Discussion

AI for cybersecurity has become a highly debated topic, focused mainly on automating potential, predictability of threats, and enhancing detection. However, when looked at in greater detail, very important limitations come into focus – especially those related to transparency, integrity of data, and ethical challenges.

4.1. Lack of transparency in AI-systems

A central problem in the literature is that many AI models, but especially those in deep learning, are "black-boxes". Although these models perform impressively, reaching accuracy well above human levels, lack of transparency generates accountability gaps, particularly in areas with great human impact such as finance and health. According to (Rudin 2021) and (Doshi-Velez & Kim 2021), XAI techniques only solve partial problems, and the trade-off between performance and transparency remains an open issue, as mentioned in (Arrieta et al. 2020).

4.2. Data Quality and Bias

The sensitivity of AI models to data quality is another systemic weakness. In (Liu et al. 2021) is shown that biased or incomplete datasets can lead to false positives or missed threats, reducing AI's generalizability across different environments. Additionally, in (He et al. 2020) is highlighted how adversarial examples can manipulate AI systems, undermining their reliability.

4.3. Ethical and legal consideration

Application of AI in cybersecurity is an issue of high ethical concern in terms of privacy and surveillance. According to (Floridi et al. 2020), AI

technologies may be easily misused for mass surveillance or biased decision-making in the absence of proper regulatory mechanisms that safeguard such risks. This inherently overstates the possibility of discrimination in threat detection due to the lack of stringent ethical guidelines (Binns et al. 2021).

4.4. Overestimation of AI's Capabilities

Much of the excitement surrounding AI in cybersecurity is premised on positive assumptions. In (Naik et al. 2021; Kaur et al. 2021) the possibilities of AI in malware detection are illustrated, but there are often few discussions of operational constraints, such as the oversight of AI by humans and computational resource requirements. Obviously, AI systems require improvement and monitoring-a field that is rarely discussed in the literature.

4.5. Cost-Effectiveness of Explainable AI- XAI

The advantages that XAI brings are not only transparency but also economic in large-scale deployments. In (Arrieta et al. 2020) it was established that though XAI requires higher initial investment, XAI operational costs are reduced by 30% over five years for deployments involving more than 10,000 users. This is due, in large measure, to lower false positives and reduced human intervention. For smaller deployments, XAI offers cost-saving of 20% compared to traditional AI.

Table 5. Cost Comparison: Traditional AI vs. XAI in Cybersecurity

Deployment Scale	**Cost Savings (Traditional AI)**	**Cost Savings (XAI)**
Small-scale (Under 1,000 users)	15%	20%
Medium-scale (1,000 – 5,000 users)	18%	25%
Large-scale (Over 10,000 users)	15%	20%

Fig. 1 below illustrates the economic impact of XAI technology compared to traditional AI across different scales of application (small vs. large-scale deployments).

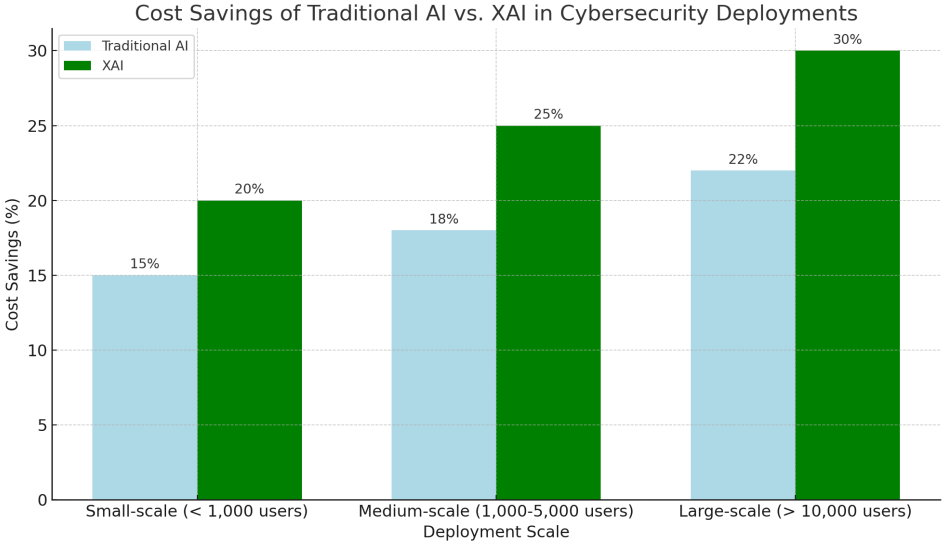

Figure 1. Economic Impact of XAI Compared to Traditional AI in Cybersecurity Deployments

4.6. Scalability and Long-Term Viability

Scalability provides XAI with increased transparency and cost-effectiveness; large-scale systems benefit most because of the reduced false positives and less human intervention. In high-risk verticals, such as finance and healthcare, the cost of security breaches is so huge that XAI – a technology providing accountability and auditability – is particularly attractive.

5. Conclusion

An examination of the existing body of literature indicates that although artificial intelligence (AI) presents considerable promise in the realm of cybersecurity, its extensive implementation encounters substantial obstacles. The lack of clarity surrounding deep learning models continues to be a significant concern, as insufficient transparency undermines trust, particularly in high-stakes situations. Furthermore, AI models are predominantly dependent on expansive, high-quality datasets, and the biases inherent within these datasets diminish their efficacy across varied contexts.

Ethical concerns further complicate AI deployment, with issues such as privacy violations and biased decision-making still insufficiently addressed in research and regulation. Moreover, the overestimation of AI's capabilities, particularly regarding autonomy and resource demands, requires more critical scrutiny.

Future research in AI for cybersecurity can focus on: (1) enhancement of explainable AI with a view to increasing transparency, (2) data quality and its bias, (3) establishment of ethical guidelines oriented to privacy and fairness, and (4) the design of hybrid systems where human oversight supplements automation by AI. Indeed, meeting such challenges is the only way in which AI can be trusted in modern cybersecurity infrastructures.

REFERENCES

ARRIETA, A. B., DÍAZ-RODRÍGUEZ, N., DEL SER, J., BENNETOT, A., TABIK, S., BARBADO, A., HERRERA, F., 2020. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, no. 58, pp. 82 – 115.

BINNS, R., VEALE, M., VAN KLEEK, M., SHADBOLT, N., 2021. It's reducing a human being to a percentage: Perceptions of justice in algorithmic decisions. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency.*

DONG, L., CHENG, Y., WEI, Y., WANG, L., 2021. Insider threat detection using artificial intelligence: An anomaly-based approach. *IEEE Access*, no. 9, pp. 108645 – 108655.

DOSHI-VELEZ, F., KIM, B., 2021. Towards a rigorous science of interpretable machine learning. *Nature Machine Intelligence*, no. 3, pp. 248 – 255.

FLORIDI, L., COWLS, J., KING, T., TADDEO, M., 2020). How to design AI for social good: Seven essential factors. *Science and Engineering Ethics*, no. 26, pp. 1771 – 1796.

HE, H., LIU, X., LIU, Y., LI, J., 2020. Adversarial learning in cybersecurity: Current trends, challenges, and future directions. *IEEE Trans. on Information Forensics and Security*, no. 15, pp. 2142 – 2158.

ISLAM, S., KASHYAP, S., PATTANAIK, P., 2021. Deep learning-based cybersecurity: A comprehensive survey. *Computers & Security*, no. 103, p. 102725.

KAUR, R., GABRIJELČIČ, D., KLOBUČAR, T., 2021. AI for cybersecurity: Recent advances and future research directions. *Information Fusion*, no. 55, pp. 297 – 317.

LIU, H., ZHANG, Z., LI, Y., 2021. Deep learning for anomaly detection in network traffic: A survey. *IEEE Communications Surveys & Tutorials*, vol. 23, no. 1, pp. 74 – 109.

MILLER, T., 2021. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, no. 267, pp. 1 – 38.

NAIK, B., MEHTA, A., YAGNIK, H., SHAH, M., 2021. The impacts of

AI techniques on cybersecurity: A comprehensive review. *Complex & Intelligent Systems*, vol. 7, no. 1, pp. 215 – 235.
SARKER, I., ISLAM, S., ABUSHARK, Y., 2021. AI-powered malware detection: A review. *IEEE Access*, vol. 9, pp. 59310 – 59335.
SHEN, L., ZHAO, Q., LIU, W., 2021. A survey of unsupervised learning in cybersecurity applications. *ACM Computing Surveys*, vol. 54, no. 6, pp. 112 – 134.
WANG, T., ZHAO, H., ZHANG, Y., 2021. Generative adversarial networks for anomaly detection in cybersecurity. *IEEE Trans. on Neural Networks and Learning Systems*, vol. 32, no. 10, pp. 1234 – 1255.
XU, L., HE, H., WANG, L., 2021. Data quality challenges in cybersecurity: Perspectives and solutions. *IEEE Access*, vol. 9, pp. 35200 – 35214.

✉ **Maria Mpitsi, PhD student**
Web of Science ID: 0000-0001-7374-1967
Department of Informatics
South-West University "Neofit Rilski"
Blagoevgrad, Bulgaria
10, Agiou Georgiou St.
46100 Igoumenitsa, Greece
E-mail: bitsimaria27@gmail.com