Guardians or threats? The double-edged role of artificial intelligence in cyber security

Ina VIRTOSU,
PhD in EU Law, University of Macau, SAR Macau, China
yb67199@connect.um.edu.mo, ivirtosu3@gmail.com

Abstract

Objectives: Artificial Intelligence (AI) has emerged as a transformative force within the domain of cyber security, offering significant advancements in threat detection, risk assessment, and automated response. **Prior work:** Leveraging machine learning, natural language processing, and anomaly detection algorithms, AI-driven systems enable organizations to process vast volumes of data, identify patterns of malicious behavior, and mitigate potential cyber threats with increased speed and accuracy. These innovations have positioned AI as a cornerstone in modern cyber defense strategies. However, the same technologies that enhance defensive capabilities are increasingly exploited by adversaries to conduct more sophisticated, targeted, and evasive cyberattacks. Malicious actors utilize AI to develop adaptive malware, automate vulnerability exploitation, and execute advanced phishing and social engineering schemes - including the deployment of deepfakes and synthetic identities. This dual-use nature of AI introduces a complex landscape in which the boundaries between protective tools and offensive weapons are increasingly blurred. **Approach:** This article offers a comprehensive examination of AI's dual role in cyber security, analyzing both its utility in enhancing defensive frameworks and its potential as a tool for cyber offense. **Implications**: It further explores the ethical, legal, and governance challenges posed by AI integration, including issues of algorithmic transparency, accountability, and regulatory compliance. **Value:** Through a review of contemporary applications, emerging threats, and current policy responses, the article provides a critical assessment of the opportunities and risks associated with AI in cyber security, and proposes recommendations for building resilient, ethically grounded, and forward-looking security infrastructures.

Keywords: algorithmic accountability, adversarial attacks, algorithmic ethics, machine learning, identity manipulation.

1. Introduction to AI in cyber security

Artificial Intelligence (AI) is increasingly becoming a cornerstone in the evolving field of cyber security. Broadly defined, AI refers to the development of computer systems capable of performing tasks that normally require human intelligence, such as reasoning, learning, pattern recognition, and problem-solving [1]. In cyber security, AI enhances the ability to detect, prevent, and respond to threats in real time, especially in high-volume environments that are beyond human analytical capacity [2], [3]. AI's utility lies in its ability to process large datasets, recognize complex patterns, and make predictive decisions - capabilities that are crucial given the volume, velocity, and sophistication of modern cyber threats [4], [5]As digital infrastructures grow and the number of endpoints and threat vectors increases, AI's role in bolstering cyber resilience has shifted from complementary to indispensable [6].

The historical development of AI in cyber security can be traced back to early expert systems in the 1980s, which used predefined rules to flag suspicious behaviour [7]. However, these systems lacked adaptability and were often overwhelmed by new forms of attack. The advent of machine learning in the late 1990s and early 2000s marked a significant leap, enabling systems to learn from data and improve over time without being explicitly programmed [8]. Recent advancements in deep learning, neural networks, and natural language processing have further enhanced the effectiveness of AI in cyber defence

[9], [10]. Today, AI is embedded in threat detection systems, behavioural analysis tools, fraud prevention platforms, and autonomous response mechanisms [11], [12]. Moreover, the use of reinforcement learning and federated learning has opened up new possibilities in real-time and decentralized security solutions [13], [14].

Recent analyses emphasize the necessity of integrating AI-driven tools with strategic policy-making and resilient organizational governance structures to protect digital environments effectively [15]. Furthermore, the socio-technical challenges of embedding algorithmic decision-making into security infrastructures reveal the need for ethically robust frameworks that account for both technological complexity and human agency [16]**.** Studies also underline the importance of cultivating responsible innovation in digital ecosystems, particularly where AI may increase vulnerabilities instead of mitigating them [17]**.**

AI's potential is not solely defensive - it is also dual-use in nature. The same algorithms that detect anomalies in network traffic can be reverse engineered by malicious actors to identify system vulnerabilities or launch adaptive attacks [14], [18]. Deepfakes, AI-generated phishing emails, and intelligent malware are growing threats that exploit AI's capabilities for malicious ends [19], [20], [21]**.** This dual-use dilemma highlights the ethical and regulatory challenges associated with AI deployment in cyber security contexts [22], [23] It underscores the importance of designing secure, transparent, and controllable AI systems that enhance trustworthiness without creating new risks [24], [25]**.**

The purpose of this article is to explore the multidimensional role of AI in cyber security, focusing on both its defensive and offensive applications. It aims to provide a critical overview of the state-of-the-art AI tools used to safeguard digital systems, while also examining how AI is weaponized by cybercriminals and state actors. By analysing key case studies and recent developments, the article seeks to reflect the complex landscape of AI-driven security, the emerging regulatory frameworks, and the ethical considerations at play.

2. AI as a defence mechanism

2.1. AI-powered intrusion detection systems

AI has significantly transformed Intrusion Detection Systems (IDS), providing more accurate and adaptive capabilities compared to traditional rule-based methods. Conventional IDS often rely on static signatures, which makes them less effective against unknown or zero-day attacks. AI, particularly machine learning (ML) and deep learning (DL) algorithms, can learn from historical attack patterns and detect anomalies in real time, even in encrypted traffic [2], [4].

Machine learning-based IDS, such as anomaly detection models, analyse baseline network behaviour and flag any deviations that may signify malicious activities. These systems are trained on vast datasets and continuously adapt, improving their precision with each iteration [10]. For instance, decision trees, random forests, support vector machines, and neural networks have all been successfully applied in modern IDS solutions [26]**.**

A real-world example of AI in IDS can be found at Amazon, where AI-driven tools like MadPot (a global honeypot network) track malicious actors' behaviours and tactics. These tools detect approximately 750 million cyber threats daily, showcasing AI's role in scaling up cybersecurity operations [27].

2.2. Behavioural analysis and anomaly detection

AI plays a pivotal role in behavioural analytics by learning what constitutes "normal" user behaviour and identifying any deviations that suggest unauthorized access or insider threats. This is especially important in combating advanced persistent threats (APTs) and low-and-slow attacks that evade traditional detection systems [28].

Unsupervised ML techniques, such as clustering and autoencoders, are particularly valuable for detecting behavioural anomalies without requiring labelled datasets [29]. For example, AI can analyse login patterns, geolocation, device use, and time-based behaviours to build adaptive risk profiles and flag suspicious activity [3].

According to a Deloitte and Wall Street Journal report (2023), Chief Information Security Officers (CISOs) are increasingly relying on AI-driven behavioural analytics to augment human analysis. These systems enable faster incident triage, reduce false positives, and enhance threat prioritization [4].

2.3. Threat intelligence automation

AI significantly accelerates and improves the accuracy of threat intelligence gathering. It does so by analysing structured and unstructured data sources, including threat feeds, cybersecurity blogs, darknet forums, and even social media platforms [30]. Natural Language Processing (NLP) allows AI systems to extract relevant indicators of compromise (IOCs) from vast volumes of text and correlate them with known attack vectors [8].

Automated threat intelligence systems like IBM's QRadar or Palo Alto's Cortex XSOAR incorporate AI to reduce response time, prioritize alerts, and offer context-aware remediation [5]. These systems are often used in conjunction with Security Orchestration, Automation, and Response (SOAR) platforms to fully automate routine decision-making processes [6].

According to the Deloitte CISO Guide (2023), organizations using AI for threat intelligence experience 30 - 40% faster detection rates and improved decision-making under pressure [4].

2.4. Real-world case studies

Amazon is one of the leading examples of successful AI implementation in cybersecurity. By using AI tools that monitor attack patterns via honeypots and machine learning-based detection systems, Amazon has drastically improved its cyber threat identification capability. The company now detects nearly 750 million threats daily - a fourfold increase from 2020 levels [27].

Another case is that of SplxAI, a cybersecurity startup using rapid red-teaming simulations to predict and neutralize potential attacks. SplxAI's platform simulates over 2,000 attacks per client in less than an hour and uses AI to refine threat assessments and remediation strategies dynamically [31]. These innovations demonstrate the growing reliance on AI as a core element in proactive defence architectures.

Furthermore, Darktrace, a well-known AI cybersecurity firm, leverages unsupervised learning to autonomously detect and respond to cyber threats in real time. Its AI Immune System monitors internal communications within a network, flagging unusual or threatening patterns based on learned baselines [32]. These solutions show how AI empowers organizations to act not only responsively but also preemptively.

3. AI in the hands of cybercriminals

3.1. Phishing 2.0: AI-generated content and language models in spear phishing

The advent of large language models (LLMs) has revolutionized phishing attacks, enabling cybercriminals to craft highly personalized and convincing messages. Traditional phishing often relied on generic templates riddled with grammatical errors. In contrast, AI-powered tools like WormGPT and FraudGPT can generate contextually relevant emails that mimic the tone and style of legitimate communications, making detection significantly more challenging [33].

For instance, in 2024, attackers utilized AI to analyse LinkedIn profiles, extracting specific details such as job titles, recent projects, and professional connections. This information was then used to generate spear-phishing emails that appeared to come from trusted colleagues or superiors, leading to successful credential theft and unauthorized access to corporate networks. Moreover, AI-driven phishing kits have emerged, capable of adapting their tactics in real-time based on user interactions. These kits can modify email content, subject lines, and even sender information dynamically to increase the likelihood of deceiving the target [34].

3.2. Deepfakes & social engineering: Identity manipulation and voice cloning

Deepfake technology, powered by AI, has introduced a new dimension to social engineering attacks. By creating realistic audio and video imitations of individuals, cybercriminals can deceive victims into believing they are interacting with trusted contacts.

A notable case occurred in 2019 when fraudsters used AI-generated voice cloning to impersonate a UK-based CEO, convincing an employee to transfer €220,000 to a fraudulent account [35]. Similarly, in 2025, scammers exploited TikTok videos to clone the voices of individuals' grandchildren, making distress calls to grandparents and requesting emergency funds. This tactic led to significant financial losses among the elderly population [36].

Furthermore, cybercriminals have employed deepfake videos in live conference calls, impersonating executives to authorize fraudulent transactions. In one instance, a finance employee was deceived into transferring $25 million after participating in a video call where all participants were AI-generated deepfakes of company executives [37].

3.3. Self-learning malware: AI-supported obfuscation and polymorphic code

AI has also been harnessed to develop self-learning malware capable of evading traditional security measures. These malicious programs can analyse their environment and modify their code to avoid detection, a technique known as polymorphism [38]. For example, the BlackMamba malware utilizes generative AI to create polymorphic code that changes its structure with each infection, rendering signature-based detection methods ineffective.

Additionally, AI enables malware to identify and exploit vulnerabilities in real-time, adapting its behaviour to maximize impact [39].

The use of AI in malware development lowers the barrier to entry for cybercriminals, allowing even those with limited technical expertise to launch sophisticated attacks. This democratization of cybercrime tools poses a significant threat to organizations worldwide.

3.4. Illustrative examples of recent AI-driven attacks

Several high-profile incidents illustrate the growing threat of AI-enhanced cyberattacks. In the Lumma Infostealer Takedown (2025), an international operation dismantled the Lumma malware network, which had infected over 394,000 computers. Lumma utilized AI to bypass security defences and was heavily used in phishing attacks [40].

In the case of North Korean IT Worker Infiltration, operatives from North Korea posed as IT professionals by using AI-generated deepfakes and stolen identities to secure remote jobs in Western companies. They funnelled earnings back to the regime, highlighting the geopolitical implications of AI misuse [41].

Marks and Spencer experienced a major cyberattack that led to losses of £300 million. The attackers exploited social engineering tactics, deceiving IT staff into resetting passwords and gaining access to internal systems [42].

In the Arup Deepfake Scam, fraudsters used AI-generated deepfake videos to impersonate the company's CFO. They convinced an employee to transfer $25 million, demonstrating the severe financial risks associated with deepfake-enabled fraud [43].

These cases demonstrate the diverse applications of AI in cybercrime, ranging from financial fraud to corporate espionage and geopolitical manipulation.

4. The arms race: Who is ahead?

4.1. Asymmetries between attackers and defenders

The battle between cyber attackers and defenders has increasingly become asymmetric, with offensive actors often enjoying significant advantages in agility, innovation, and tactical execution. While defenders must secure all entry points across sprawling networks, attackers only need to find a single vulnerability. This enduring imbalance is further amplified by AI, which enables adversaries to automate reconnaissance, exploit chains, and evasive manoeuvres at an unprecedented scale [14].

AI lowers the barrier to entry for cybercrime by offering user-friendly toolkits, pre-trained models, and automated attack frameworks. Cybercriminals no longer require deep technical knowledge to exploit systems; instead, they can deploy generative AI to create phishing emails, generate malware, and even synthesize voices and faces [44]. This democratization of offensive capabilities threatens to outpace the capacity of defenders to adapt and respond effectively.

Meanwhile, defenders often face bureaucratic inertia, budgetary constraints, and ethical restrictions when deploying cutting-edge AI tools. Public-sector organizations and under-resourced private companies may lack the skilled personnel and access to state-of-the-art models necessary to mount adequate defences [45].

4.2. Access to computing power, data, and innovation

A major driver of the attacker-defender imbalance lies in access to computing resources, proprietary data, and innovative tools. Offensive actors - particularly state-sponsored groups and well-funded cybercriminal organizations - leverage advanced GPUs, distributed botnets, and black-market data troves to train and fine-tune AI models for malicious purposes [46]. In contrast, defenders are often limited to curated or synthetic datasets that may not reflect real-world adversarial behaviours. Defensive AI relies heavily on labelled data for training intrusion detection systems, threat classification models, and anomaly detection engines. However, attackers continually evolve their techniques, rendering historical datasets obsolete.

Innovation is also more fluid on the offensive side. Red teams and cybercriminals are not bound by the regulatory, ethical, and legal norms that constrain defensive research and deployment. This freedom allows them to experiment with risky or untested capabilities - such as generative adversarial networks (GANs) for deepfake production or reinforcement learning for malware propagation [47].

4.3. Strategic advantages of AI in offense vs. defence

AI provides several clear advantages to attackers. Offensive use cases benefit from predictive capabilities, automation, and adaptability. AI can simulate user behaviour, bypass authentication mechanisms, and disguise command-and-control traffic to mimic legitimate network flows [48]. These tactics significantly delay detection and response times. Furthermore, AI allows attackers to conduct real-time exploitation of zero-day vulnerabilities, even tailoring payloads dynamically based on system feedback. For example, polymorphic malware driven by machine learning can evade antivirus signatures and adjust in response to sandbox environments [39]. By contrast, defensive AI tools are largely reactive - focused on detection, mitigation, and response. While some progress has been made in proactive defence (e.g., predictive threat intelligence), most systems still rely on signature databases, heuristic rules, or static models that require constant retraining [49]. Additionally, explainability remains a challenge in defensive AI, making it difficult to trust and audit decisions made by autonomous systems [50].

4.4. Are defenders lagging behind?

Evidence suggests that defenders are currently lagging behind attackers in terms of operationalizing AI. Although significant investments are being made in AI-driven threat intelligence, automated SOCs (Security Operations Centers), and behaviour analytics, many tools are still in early development or deployment stages. A 2024 global survey by IBM found that while 68% of organizations have adopted AI for cybersecurity, only 24% have fully integrated these tools into their core security operations [51]. Challenges include lack of skilled personnel, difficulty integrating AI with legacy systems, and uncertainty about regulatory compliance. Moreover, there is growing concern that attackers are better

at collaborating across borders and underground communities than defenders, who are often siloed by national jurisdictions or corporate competition [18]. This lack of coordination hinders timely sharing of threat intelligence and response strategies.

However, there are signs of improvement. Initiatives like MITRE's Caldera platform and the EU's AI4CYBER project aim to bolster defensive capabilities by developing open-source AI tools and fostering collaboration between researchers, governments, and industry [52]. Public-private partnerships and AI ethics frameworks may also help level the playing field over time.

5. Regulatory and ethical implications

5.1. Current regulatory landscape

As artificial intelligence becomes embedded in both cybersecurity defence and attack strategies, the regulatory and ethical implications become more urgent. While AI holds the promise of enhancing security, it also poses serious challenges that require thoughtful governance to prevent abuse, ensure accountability, and maintain public trust [53]. Globally, there is a growing momentum to regulate AI in cybersecurity, but the frameworks remain fragmented. The European Union leads with its proposed AI Act, which classifies AI systems according to risk levels and proposes strict oversight for high-risk applications, including those in critical infrastructure and security [54]. Meanwhile, the United States has adopted a more sector-specific and voluntary framework promoting innovation while emphasizing responsible use [55].

At the international level, efforts by the OECD and the UN's Group of Governmental Experts (GGE) on developments in the field of information and telecommunications in the context of international security indicate attempts to build consensus around AI norms [56]. However, enforcement and harmonization remain difficult due to divergent national interests and capabilities.

5.2. Ethical dilemmas and dual-use risks

AI in cybersecurity often involves dual-use technologies that can serve both protective and malicious purposes. Language models, for example, can power spam filters or generate phishing emails. Deep learning techniques can enhance fraud detection or produce deepfakes for identity theft [14], [18]. These dual-use potentials raise ethical concerns around the open publication of AI models and datasets. Should access to generative models be restricted? How can we balance openness in research with the risk of weaponization? These questions are central to ongoing debates in the AI ethics community [57]. Transparency and explainability are also critical. Black-box models, which are difficult to interpret, challenge the principle of accountability in cybersecurity. If an AI system flags a threat or mistakenly blocks legitimate activity, determining liability is complex [58].

5.3. Bias, discrimination, and civil liberties

Bias in AI models poses additional ethical and legal concerns. Cybersecurity tools trained on unrepresentative datasets may disproportionately misclassify behaviours from minority groups or specific regions, leading to unfair targeting or surveillance [59]. Moreover, the use of AI in surveillance and predictive policing has triggered widespread civil liberties

debates. Tools like facial recognition and predictive analytics, when used without oversight, can infringe on privacy and freedom of expression [60]. International human rights law must be integrated into the development and deployment of AI systems in cybersecurity contexts.

5.5. The need for global coordination

Given the transnational nature of cyber threats, unilateral or fragmented regulation is insufficient. A coordinated global governance framework is necessary to set standards, facilitate information sharing, and prevent arms races in AI-enabled cyber capabilities [61]. Efforts like the Global Partnership on AI (GPAI) and proposals for an international AI treaty modelled on the Geneva Conventions for cyber warfare underscore the need for collective ethical commitments [62]. Such frameworks should include mechanisms for accountability, transparency, and recourse for affected individuals.

AI in cybersecurity presents both extraordinary opportunities and unprecedented risks. As capabilities evolve, so must the ethical and regulatory frameworks that govern their use. Balancing innovation with responsibility, openness with security, and efficiency with human rights is imperative. Only through inclusive, transparent, and globally coordinated approaches can we ensure that AI serves as a tool for protection rather than oppression.

6. Future directions and recommendations

6.1. Explainable AI (XAI): Building trust in security decisions

One of the most pressing challenges in deploying artificial intelligence in cybersecurity is the opacity of many machine learning models, especially DL systems. This "black box" nature hinders trust, particularly in high-stakes security decisions where accountability is paramount. Explainable AI (XAI) seeks to mitigate this issue by making AI systems more interpretable to human users, thereby increasing transparency, trust, and adoption [58], [63].

In cybersecurity, XAI is essential for understanding anomaly detection, interpreting the rationale behind threat prioritization, and justifying autonomous defence actions. For instance, when an AI model flags a network behaviour as malicious, security professionals must be able to verify and understand the reasoning - especially if automated countermeasures are enacted. Without explainability, organizations risk both false positives and dangerous blind trust. Investing in interpretable models or post-hoc explanation tools like LIME and SHAP is thus a necessary direction [64].

6.2. Resilient architectures: Human-AI collaboration and fault tolerance

The integration of AI in cybersecurity must be guided by principles of resilience and human-machine teaming. AI is not infallible; adversaries can manipulate its outputs through adversarial attacks or poisoning of training data [65]. To address this, future security systems should embrace hybrid architectures - where AI enhances, rather than replaces, human judgment. Research suggests that the most robust cybersecurity frameworks will blend human intuition, contextual awareness, and strategic thinking with AI's pattern recognition and scalability [66]. This "centaur" model - where humans and machines complement each other - can improve decision quality and system fault tolerance.

Additionally, organizations should design architectures that anticipate AI failure modes and allow for graceful degradation, fail-safes, or human override in critical functions [67].

6.3. Cross-sector collaboration: Public-private partnerships and global cooperation

Addressing AI-driven threats in cybersecurity requires a coordinated effort that transcends organizational and national boundaries. Public-private partnerships (PPPs) can accelerate innovation while pooling expertise and threat intelligence. Government institutions, private sector actors, and academia must co-create platforms for sharing data, tools, and best practices [6], [55].

Several initiatives already showcase the promise of such collaboration. For instance, the Cybersecurity and Infrastructure Security Agency (CISA) in the U.S. and the European Union Agency for Cybersecurity (ENISA) have fostered information exchange networks between state agencies and private critical infrastructure providers. In parallel, technology companies like Microsoft, IBM, and Google have committed to responsible AI pledges and threat sharing initiatives [68], [69].

On a global scale, AI in cybersecurity must be framed as a transnational concern. Norms for AI in cyber conflict - such as no-target zones for critical civilian infrastructure - should be developed through multilateral dialogues at platforms like the UN Group of Governmental Experts on Developments in the Field of Information and Telecommunications in the Context of International Security (GGE) and the Global Partnership on Artificial Intelligence (GPAI).

6.4. Call for interdisciplinary research and ethical innovation

The rapid evolution of AI in cyber security demands an interdisciplinary research agenda. Legal scholars, ethicists, technologists, and sociologists must collaborate to navigate the complex moral and legal dilemmas of autonomous defence, data privacy, and dual-use risks [14], [70]. Ethical frameworks should inform both the design and deployment of AI systems to ensure they align with democratic values, human rights, and societal well-being.

Educational institutions must also expand curricula to train professionals in AI-literate cybersecurity. This includes not only engineers and analysts but also policymakers and legal experts who can bridge the gap between innovation and governance. Funding bodies and academic institutions should prioritize interdisciplinary AI research that integrates ethics, policy, and usability.

Thus, the future of AI in cybersecurity will depend not only on technical excellence but also on explainability, resilience, cooperation, and ethical foresight. By fostering transparent systems, robust human-AI teams, and inclusive cross-sector governance, we can steer AI's capabilities toward securing digital ecosystems - rather than undermining them.

7. Conclusions

The intersection of artificial intelligence and cybersecurity presents one of the most significant challenges - and opportunities - of our digital age. AI is not inherently good or

bad; rather, it is a neutral enabler, with its impact shaped by the intentions and applications of its users. On the defensive front, AI has transformed how we detect threats, analyse vast amounts of data, and respond to incidents with increased speed and accuracy. Tools like ML-powered intrusion detection systems, behavioural analytics, and predictive modelling are helping defenders stay ahead in an increasingly complex environment.

Yet the same capabilities are being exploited by malicious actors. From highly personalized phishing campaigns powered by language models to the manipulation of audio and video for deception, AI is giving rise to a new breed of cyberattacks that are faster, more convincing, and harder to detect. Self-learning malware and adaptive obfuscation techniques illustrate how attackers are rapidly innovating, often with fewer constraints than their defensive counterparts. This dynamic creates an asymmetry where defenders must anticipate a broader range of threats without equivalent resources or agility. Addressing these challenges requires more than technical innovation; it calls for robust governance frameworks, ethical foresight, and sustained collaboration across sectors. Explainable AI (XAI) is one promising avenue for improving transparency and trust in automated security decisions. Resilient architectures - designed for failure tolerance and adaptive learning - must also become the norm rather than the exception. In parallel, interdisciplinary research that bridges law, ethics, computer science, and public policy is essential for anticipating the long-term implications of AI-driven security tools.

Ultimately, the future of AI in cybersecurity hinges on our collective ability to balance innovation with responsibility. If we act decisively - investing in secure-by-design systems, fostering global cooperation, and embedding ethical principles at the core of AI development - we can steer this powerful technology toward a safer, more secure digital ecosystem. Resilient architectures - designed for failure tolerance and adaptive learning - must be complemented by human expertise to ensure reliable responses to evolving threats. The hybrid human-AI model is not simply a technical preference but a strategic necessity in managing the complexity and unpredictability of cyber conflict.

Moreover, regulatory frameworks must evolve in tandem with technological advances, fostering international cooperation and embedding ethical principles at their core. Only through collaborative global efforts can we set meaningful boundaries that deter malicious AI use while encouraging innovation for public good.

Education and interdisciplinary research will be equally crucial in preparing the next generation of cybersecurity professionals, policymakers, and ethicists. As AI technologies become more pervasive, these stakeholders will need the skills and insights to anticipate risks, mitigate harms, and leverage AI responsibly.

In conclusion, artificial intelligence in cybersecurity offers a transformative opportunity to strengthen digital defences and promote safer cyberspace. Yet, this potential will only be realized if we balance technological prowess with governance, transparency, ethical integrity, and international solidarity. By doing so, we can ensure that AI remains a force for protection, trust, and resilience in the face of emerging digital challenges.

References

[1] S. Russell and P. Norving, "Artificial Intelligence: A Modern Approach," *Pearson,* vol. 4, 2021.

[2] A. L. Buczak and E. Guven, "A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection," *IEEE Communications Surveys & Tutorials,* vol. 18, no. 2, pp. 1153-1176, 2016.

[3] C. Chio and D. Freeman, "Machine learning and security: Protecting systems with data and algorithms," *O'Reilly Media,* 2018.

[4] Deloitte, "CISO's Guide: Using AI for Cyber Defense," *Risk & Compliance Journal, WSJ ,* 2024.

[5] IBM QRadar, "Advanced threat detection with IBM QRadar SIEM," 2023.

[6] enisa, "AI cybersecurity challenges," *Threat Landscape for Artificial Intelligence,* 2020.

[7] J. P. Anderson Co., "Computer security threat monitoring and surveillance," 1980.

[8] R. Sommer and V. Paxson, "Outside the Closed World: On Using Machine Learning for Network Intrusion Detection," *IEEE Symposium on Security and Privacy,* pp. 305-316, 2010.

[9] I. Goodfellow, Y. Bengio and A. Courville, "Deep Learning," *MIT Press,* 2016.

[10] N. Shone , T. N. Ngoc, V. D. Phai and S. Qi, "A Deep Learning Approach to Network Intrusion Detection," *IEEE Transactions on Emerging Topics in Computational Intelligence,* vol. 2, no. 1, pp. 41-50, 2018.

[11] H. Martin , J. Komarkova , E. Bou-Harb and P. Celeda, "Survey of Attack Projection, Prediction, and Forecasting in Cyber Security," *IEEE Communications Surveys & Tutorials,* vol. 21, no. 1, pp. 640-660, 2019.

[12] S. Li, "Federated learning for cybersecurity applications: A survey," *Journal of Network and Computer Applications,* 2020.

[13] Y. Liu, "Reinforcement learning based cybersecurity defence," *IEEE Transactions on Cybernetics,* 2021.

[14] M. Brundage and e. al., "The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation," *Computer Science, Artificial Intelligence, arxiv ,* 2018.

[15] A. Soare , "The strategic role of artificial intelligence in strengthening cybersecurity resilience," *Sustainability Challenges in Romania's Development,* no. 2, pp. 35-45, 2023.

[16] D. Tanasescu and A. Petrescu, "Socio-technical challenges of algorithmic decision-making in critical infrastructures," *Sustainability Challenges in Romania's Development,* no. 1, pp. 22-31, 2023.

[17] M. Dumitru , "Responsible innovation in AI-driven digital ecosystems: A policy perspective," *Sustainability Challenges in Romania's Development,* no. 2, pp. 55-64, 2023.

[18] M. Taddeo and L. Floridi, "How AI can be a force for good," *Science,* vol. 361, no. 6404, pp. 751-752, 2018.

[19] A. Kumar, "AI-powered phishing detection and mitigation techniques: A review," *Cybersecurity,* 2022.

[20] I. Virtosu and M. Goian, "Disinformation using artificial intelligence technologies – akey component of Russian hybrid warfare," *Smart Cities International Conference (SCIC) Proceedings,* vol. 11, pp. 197-222, 2023.

[21] Kaspersky;, "The evolution of AI threats: Deepfakes and beyond," 2023.

[22] R. Calo, "Artificial intelligence policy: A primer and roadmap," *U.C. Davis Law Review,* vol. 51, no. 2, 2018.

[23] L. Floridi and e. al, "AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations," *Minds and Machines,* vol. 28, no. 4, pp. 689-707, 2018.

[24] OECD, "Recommendation of the Council on Artificial Intelligence," 2019.

[25] Future of Life Institute, "AI governance and safety," 2023.

[26] S. Islam, "A survey on deep learning techniques for intrusion detection systems," *Journal of Network and Computer Applications,* vol. 127, pp. 25-48, 2019.

[27] J. Rundle, "The AI Effect: Amazon Sees Nearly 1 Billion Cyber Threats a Day," *Cybersecurity, WSJ,* 2024.

[28] Z. Cai, J. Han and X. Zhang, "Detecting advanced persistent threats using machine learning," *IEEE Access,* vol. 8, 2020.

[29] G. Kim, S. Lee and S. Kim, "A novel hybrid intrusion detection method integrating anomaly," *Expert Systems with Applications,* vol. 41, no. 4, pp. 1690-1700, 2016.

[30] A. Pillai, "AI-based threat intelligence: A survey," *Journal of Cyber Security Technology,* vol. 4, no. 3, pp. 163-191, 2020.

[31] Business Insider, "SplxAI's pitch deck: AI startup simulates cyberattacks in real time," 2024.

[32] J. Fier , "The future of cyber security: 2022 Predictions by Darktrace," *Darktrace Blog,* 2022.

[33] TitanHQ, "How AI Is Allowing Cybercriminals to Launch Sophisticated Cyber-attacks," 2025.

[34] CyberPinnacle, "AI-Powered Phishing Attacks: Navigating Evolving Cybersecurity Threats," 2024.

[35] C. Stupp, "Fraudsters Used AI to Mimic CEO’s Voice in Unusual Cybercrime Case," *Cybersecurity, WSJ,* 2019.

[36] B. Cruz, "Scammers are using AI to bilk victims out of millions: Here’s how to protect yourself," *New York Post,* 2025.

[37] A. Payuyo, "AI-Enhanced Cyberattacks: Understanding and Combating Evolving Threats," *AvePoint,* 2024.

[38] SOCRadar, "Examples of AI-Assisted Cyber Attacks," *Digital Risk Protection ,* 2024.

[39] Perception Point, "AI malware: Types, real life examples, and defensive measures," 2023.

[40] L. H. Newman and M. Burgess, "Authorities Carry Out Elaborate Global Takedown of Infostealer Heavily Used by Cybercriminals," *Security, Wired ,* 2025.

[41] L. Doye, "Fears North Korean spies are posing as IT workers to infiltrate Western companies & earn cash for Kim’s warped regime," *The Sun,* 2025.

[42] Financial Times, "In cyber attacks, humans can be the weakest link," 2025.

[43] A. Talapatra, "The Rise of AI-Driven Cyber Attacks: Implications for Modern Security," *Radware,* 2025.

[44] T. B. Brown, B. Mann, N. Ryder and e. al, "Language Models are Few-Shot Learners," *Computer Science > Computation and Language, arxiv ,* 2020.

[45] CISA, "AI and cybersecurity: Opportunities and challenges for defenders," 2023.

[46] M. Kumar, J. Chen and S. Patel, "AI in cyber offense and defense: A Comparative Review," *Journal of Cybersecurity,* vol. 9, no. 1, pp. 34-49, 2023.

[47] U.S. Department of Homeland Security, "The role of AI in the future of cyber threats," 2023.

[48] N. Guansiracusa, How Algorithms Create and Prevent Fake News: Exploring the Impacts of Social Media, Deepfakes, GPT-3, and More, vol. 1, Apress, 2021.

[49] Deloitte, "AI-driven cybersecurity: Moving from reactive to predictive defense," 2024.

[50] R. Shokri, M. Stronati, C. Song and V. Shmatikov, "Membership Inference Attacks against Machine Learning Models," *IEEE Symposium on Security and Privacy,* 2017.

[51] IBM, "AI Adoption in cybersecurity: Global status report," 2024.

[52] MITRE Caldera, "Adversary emulation for defensive testing," 2023.

[53] A. Jobin, M. Ienca and E. Vayena, "The global landscape of AI ethics guidelines," *Nature Machine Intelligence ,* vol. 1, no. 9, pp. 389-399, 2019.

[54] European Commission, "Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts," 2021.

[55] NIST, "AI Risk Management Framework," *Information Technology Laboratory.*

[56] United Nations Digital Library , "Developments in the field of information and telecommunications in the context of international security : resolution," 2022.

[57] R. Binns, "Fairness in Machine Learning: Lessons from Political Philosophy," *Journal of Machine Learning Research, Conference on Fairness, Accountability, and Transparency,* p. 149–159, 2018.

[58] D. V. Finale and B. Kim, "Towards A Rigorous Science of Interpretable Machine Learning," *Statistics > Machine Learning, arxiv ,* 2017 .

[59] K. Crawford, Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence, vol. 74, Yale Univeristy Press, 2022, pp. 61-62.

[60] S. Feldstein, "The Global Expansion of AI Surveillance," *Carnegie Endowment for International Peace,* 2019 .

[61] L. Floridi and J. Cowls, "A Unified Framework of Five Principles for AI in Society," *Harvard Data Science Review,* vol. 1, no. 1, 2019.

[62] OECD, "Global Partnership on Artificial Intelligence".

[63] D. Gunning , M. Stefik, J. Choi and T. Miller, "XAI -explainable artificial intelligence," *Science Robotics,* vol. 4, no. 37, 2019.

[64] S. Lundeberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," *Computer Science>Artificial Intelligence, arxiv,* vol. 30, 2017.

[65] B. Battista and F. Roli, "Wild Patterns: Ten Years After the Rise of Adversarial Machine Learning," *Computer Science > Computer Vision and Pattern Recognition,* vol. 84, pp. 317-331, 2018.

[66] M. Turek, "AI-human teaming for national security," Journal of Defense Modeling and Simulation," vol. 19, pp. 361-369, 2022.

[67] D. Amodei, C. Olah, S. Jacob , P. Christiano, J. Schulman and D. Mane, "Concrete Problems in AI Safety," *Computer Science > Artificial Intelligence, arxiv ,* 2016.

[68] Microsoft, "AI security principles," 2021.

[69] Google AI, "Our AI Principles".

[70] B. D. Mittelstadt, P. Allo, M. Taddeo, S. Watchter and L. Floridi, "The ethics of algorithms: Mapping the debate," *Big Data & Society,* vol. 3, no. 2, 2016.