Cybersecurity in the Age of AI: Challenges and Solutions

Damir Delija
Zagreb University of Applied Sciences, Croatia
Goran Sirovatka
Zagreb University of Applied Sciences, Croatia
Darko Možnik
Zagreb University of Applied Sciences, Croatia
Marinko Žagar
Zagreb University of Applied Sciences, Croatia

Abstract

The rapid advancement of artificial intelligence (AI) has brought both opportunities and challenges to cybersecurity. On one hand, AI-powered tools are revolutionizing threat detection, incident response, and predictive analytics, enabling organizations to defend against increasingly sophisticated cyberattacks. On the other hand, AI is being weaponized by malicious actors to conduct automated attacks, develop adaptive malware, and create deepfake content for social engineering. This paper explores AI's dual role in cybersecurity, highlighting its potential as a defensive tool while addressing the ethical, legal, and technical challenges posed by its misuse. Through case studies and analysis of current trends, we examine how AI is reshaping the cybersecurity landscape and discuss strategies for balancing innovation with regulation. The findings emphasize the need for collaboration between technologists, policymakers, and cybersecurity experts to harness AI's potential while mitigating its risks.

Keywords: artificial intelligence, cybersecurity, cyber threats, machine learning, deepfake

JEL classification: O33, L86, K24, D82

Paper type: Research article
Received: 27 March 2025
Accepted: 9 June 2025

DOI: 10.54820/entrenova-2025-0034

Citation: Delija, D., Sirovatka, G., Možnik, D., & Žagar, M. (2025). Cybersecurity in the Age of AI: Challenges and Solutions. ENTRENOVA - ENTerprise REsearch InNOVAtion, 11(1), https://doi.org/10.54820/entrenova-2025-0034.

Acknowledgments: The authors acknowledge the assistance of OpenAI's ChatGPT and DeepSeek (DeepSeek.com) in generating initial drafts and refining the text of this paper.

Introduction

In today's rapidly evolving digital landscape, cybersecurity has become one of the most pressing challenges of the 21st century. As organizations and individuals increasingly depend on digital technologies, the frequency and complexity of cyber threats have surged. A 2023 report by Cybersecurity Ventures (Cybersecurity Ventures. 2023) predicts that the global cost of cybercrime will reach $10.5 trillion annually by 2025, underscoring the urgent need for innovative solutions to combat these threats. At the center of this ongoing battle is artificial intelligence (AI), a transformative technology that is reshaping cybersecurity in both promising and perilous ways.

AI is a double-edged sword in cybersecurity. On the one hand, AI-driven tools enhance threat detection, incident response, and predictive analytics, enabling organizations to defend against increasingly sophisticated cyberattacks proactively. By analyzing vast amounts of data in real-time, AI-powered systems can identify anomalies, automate responses, and improve the overall efficiency of cybersecurity defenses. On the other hand, malicious actors are weaponizing AI to execute more advanced and adaptive attacks. AI-driven phishing campaigns, self-learning malware, and deepfake-enabled social engineering are being used to exploit vulnerabilities at an unprecedented scale and speed.

This duality presents both challenges and opportunities. While AI has the potential to revolutionize cybersecurity, its misuse raises significant ethical, legal, and technical concerns that must be addressed through a multidisciplinary approach. This paper explores AI's dual role in cybersecurity by examining both its transformative potential and the risks it introduces. Through real-world case studies, current trends, and emerging solutions, this study aims to provide a comprehensive understanding of how AI is reshaping the cybersecurity landscape and to propose actionable strategies for harnessing its power while mitigating its risks.

Based on the literature and preliminary findings, the following hypotheses were formulated and analyzed:

H1: AI-driven cybersecurity tools significantly reduce incident response time in comparison to traditional methods.

H2: AI-powered attacks, such as deepfake-based social engineering, exhibit higher breach success rates than traditional attacks.

H3: Implementing adversarial training increases the detection accuracy of AI-based security models.

Methodology

This study adopts a mixed-methods approach, combining qualitative case studies with quantitative data from industry reports and academic research. The primary sources are real-world cyberattacks, such as the BlackMatter ransomware attack and deepfake scams, and AI-powered security offerings like Darktrace and IBM QRadar (IBM Security, 2022), exemplifying the dual nature of AI in cybersecurity. Secondary data from reputable organizations, including IBM, MIT, and NIST, alongside peer-reviewed journals, provides an overview of the trends in AI-based threats and countermeasures. The approach spotlights a side-by-side comparison of offensive and defensive AI uses, in a context aligned with ethical and legal principles, as represented by the EU AI Act. Solutions were developed through interdisciplinary research, including technical innovations, policy recommendations, and human resource development proposals. Case studies were strategically selected for their relevance to emerging cybersecurity threats and for demonstrating AI's transformative impact on the field.

Given the central role of artificial intelligence (AI) in modern cybersecurity, this study not only explores AI-driven challenges and solutions but also integrates AI tools into the research process itself. Specifically, OpenAI's ChatGPT and DeepSeek AI's DeepSeek were employed to enhance the efficiency and depth of analysis throughout the manuscript preparation.
The use of these AI-powered language models was motivated by their ability to:

- Rapidly synthesize information from diverse cybersecurity literature, aiding in the formulation of a structured discussion.
- Improve clarity and coherence in technical explanations, ensuring accessibility for a broad academic and professional audience.
- Facilitate comparative analysis of AI-driven cybersecurity threats and defenses by identifying patterns and trends in published research.

By incorporating AI in the drafting process, this paper exemplifies the practical application of AI tools in cybersecurity research. However, all AI-generated content was critically reviewed, validated, and refined by the authors to maintain academic integrity and ensure the originality and reliability of the final work. The insights, conclusions, and arguments presented remain human-driven, with AI serving as an augmentative rather than an autonomous component of the research process.
While the study predominantly relies on secondary data and case studies, a triangulation approach was employed by cross-validating findings across multiple sources, including industry reports, peer-reviewed journals, and authoritative cybersecurity organizations (e.g., NIST, MITRE, IBM). Case selection followed the criteria of recency (2021–2024), impact (financial or operational), and representativeness (across varied sectors and attack vectors).

Limitations of This Study

While this study provides a comprehensive analysis of AI's dual role in cybersecurity, several limitations must be acknowledged:

- Scope Constraints – The study primarily focuses on case studies and industry reports from the past 4 years. Rapid advancements in AI and cybersecurity mean that new threats and defense mechanisms may emerge beyond the scope of this research.
- Dependence on Secondary Data – Many findings rely on industry reports, peer-reviewed literature, and case studies, which may introduce biases or inconsistencies in data interpretation. Access to proprietary cybersecurity threat intelligence could deepen analysis.
- Lack of Real-Time Empirical Testing – The study does not conduct live penetration testing or real-world AI attack simulations. Future research incorporating experimental validation of AI-driven cyberattacks and defenses could provide more concrete evidence.
- Regulatory Uncertainty – While the study examines frameworks like the EU AI Act, AI regulations are still evolving, and policy responses to AI-driven cybersecurity challenges may change. This study does not account for future legislative developments that could reshape AI governance.
- Ethical and Societal Implications – The paper discusses ethical concerns related to AI in cybersecurity. However, it does not explore broader societal impacts, such as AI-driven job displacement, misinformation, and surveillance risks. A more interdisciplinary approach could strengthen this dimension.
- Potential Geopolitical Bias – The research primarily focuses on Western regulatory frameworks (e.g., U.S., EU). It may not fully capture AI-cybersecurity

policies in regions such as China, Russia, and developing nations. A global comparative analysis would provide a more balanced perspective.

Research Hypotheses

Artificial intelligence (AI) is increasingly recognized as both a tool for cyber defense and a weapon for cyber offense. To systematically explore AI's dual impact in cybersecurity, this study formulates and analyzes the following research hypotheses, based on existing literature and case studies:

- H1: AI-driven cybersecurity tools significantly reduce the time required for incident detection and response compared to traditional methods.
- H2: AI-powered cyberattacks (e.g., deepfakes, adaptive ransomware) achieve higher breach success rates than conventional cyber threats.
- H3: The implementation of adversarial training techniques improves the robustness of AI-based security systems against evasion tactics.
- H4: Organizations that invest in AI-focused workforce upskilling show a measurable improvement in cybersecurity resilience.

Each hypothesis is evaluated through qualitative case study analysis and supported by secondary quantitative data from reputable cybersecurity sources. These hypotheses are not tested through direct experimentation but are validated through triangulation across documented real-world incidents, organizational reports, and scholarly findings.

Results: Key Findings on AI's Dual Role in Cybersecurity

The analysis of AI's impact on cybersecurity reveals several critical findings, categorized into defensive capabilities, offensive risks, human and regulatory implications, and solution effectiveness.

AI's Defensive Efficacy can be described in the following points in Table 1.

Table 1
AI's Defensive Efficacy

Reduction in False Positives	AI-driven tools reduced false positives by 60%, preventing multimillion-dollar losses, including mitigating a $5.2 million ransomware attack at a financial institution (Darktrace, 2023).
Enhanced Threat Detection	Predictive analytics shortened threat detection times by 40%, as demonstrated in healthcare networks leveraging AI-powered threat hunting (IBM Security, 2022; HealthITSecurity, 2023).

Source: Authors' work

AI's Offensive Risks can be summarised in Table 2.

Table 2
AI's Offensive Risks

Increased Breach Success Rates	AI-powered cyberattacks, including deepfake scams and adaptive ransomware, increased breach success rates by 300% (e.g., $243K CEO fraud and $200M ransomware campaign) Forbes, (2023); Elliptic, (2021); Kaspersky, (2022); Kaspersky, (2024).
Adversarial AI Exploits	Techniques such as data poisoning degraded threat detection accuracy by 40% in models lacking robust adversarial training (MIT Technology Review, 2023; MITRE, 2023).

Source: Authors' work

Human and Regulatory Impacts are described in Table 3.

Table 3
Human and Regulatory Impact

Faster Incident Resolution	Organizations with AI-trained cybersecurity teams resolved incidents 50% faster than those relying solely on traditional security methods (ISC², 2023; Palo Alto Networks, 2023).
Regulatory Gaps and Legal Challenges	Weak regulatory frameworks allowed 65% of AI-driven attacks to exploit jurisdictional ambiguities, underscoring the urgency for global AI governance frameworks, such as the EU AI Act (European Commission, 2023).
Adversarial Training Success	Implementing adversarial training improved threat detection resilience by 35%, reinforcing AI's ability to withstand sophisticated cyber threats (MITRE, 2023).
Workforce Development Impact	Organizations participating in public-private upskilling initiatives reduced AI cybersecurity skill gaps by 25%, strengthening their defensive posture (ISC², 2023; NSA, 2023).

Source: Authors' work

AI is transformative but imperfect, while AI enhances threat detection and response, its misuse amplifies risks, necessitating human oversight, ethical safeguards, and adversarial resilience. In asymmetric warfare, attackers can exploit AI's scalability more effectively than defenders, requiring proactive investment in adaptive AI defenses.

Collaboration is key through cross-sector partnerships, regulatory harmonization, and global AI governance. Such efforts are necessary to thwart AI-powered threats and achieve responsible AI adoption.

Discussion

This chapter discusses the dual role of AI in cybersecurity, including its use as a defensive mechanism and an offensive tool, and the general implications of this duality.

AI as a Defensive Tool

The ability of artificial intelligence to analyze large datasets, recognize patterns, and enable automated responses has made it invaluable in modern cybersecurity approaches. Its main uses in defense include topics listed in Table 4.

Table 4
AI as a Defensive Tool

Threat Detection and Incident Response	AI-powered cybersecurity systems significantly improve real-time threat detection by spotting unusual patterns in network activity. For instance, Darktrace's Antigena uses advanced machine learning algorithms to detect anomalies that may indicate a breach, flagging them based on deviations from established behavioral baselines (Darktrace, 2023). In a similar vein, IBM's QRadar applies AI to correlate security events across diverse systems, helping reduce false positives and speed up incident response times (IBM Security, 2022).
Predictive Analytics	AI models trained on historical cyberattack patterns are increasingly capable of predicting and preventing emerging threats. For example, Google's Chronicle leverages machine learning to scan dark web activity and identify indicators of potential ransomware campaigns before they materialize (Cybersecurity Dive, 2023). Likewise, the MITRE ATT&CK Framework incorporates AI-driven simulations of adversarial behavior, enabling organizations to proactively identify and address vulnerabilities before they are exploited (MITRE, 2023).
Automation of Security Operations	AI streamlines cybersecurity workflows by automating repetitive tasks such as log analysis and patch management. Platforms like Palo Alto Networks' Cortex XSOAR employ natural language processing (NLP) to triage alerts, allowing security analysts to focus on high-priority threats (Palo Alto Networks, 2023). According to Gartner (2023), organizations implementing AI-driven Security Orchestration, Automation, and Response (SOAR) systems have reduced their average incident resolution time by 80%.
Enhanced Authentication and Fraud Prevention	AI-powered behavioral biometrics enhance authentication by analyzing keystroke dynamics, mouse movements, and other user behaviors. Mastercard's NuDetect, for instance, employs AI to detect anomalies in user interactions, reducing account takeover fraud by 80% (Mastercard, 2022).

Source: Authors' work

AI as an Offensive Weapon

While AI strengthens cybersecurity defenses, its exploitation by malicious actors presents unprecedented risks. Cybercriminals are leveraging AI to conduct automated attacks, create adaptive malware, and manipulate digital identities through deepfakes.

In automated, scalable attacks, AI enables cybercriminals to launch large-scale, hyper-personalized attacks with minimal effort. DeepPhish, an AI-driven phishing tool, scrapes social media profiles to craft convincing emails, increasing phishing success rates by 300% (Barracuda Networks, 2023). Additionally, researchers have demonstrated how AI-generated malware, created through models like ChatGPT, can evade traditional signature-based detection systems (Check Point Research, 2023).

AI-powered malware can dynamically adjust its behavior to avoid detection, making it adaptive and employing evasion tactics. DeepLocker, developed as a proof-of-concept by IBM, uses ML to remain dormant until it identifies a specific target, such as via facial recognition, before activating (IBM Research, 2023). Similarly, reinforcement learning enables malware to test various evasion strategies against security sandboxes and adapt in real time to bypass defenses (McAfee, 2023).

Deepfake technology has introduced new avenues for social engineering attacks. AI-generated deepfake audio and video have been used to impersonate executives and authorize fraudulent transactions. In 2023, a UK energy firm fell victim to a deepfake CEO voice scam, resulting in a $243,000 financial loss (Forbes, 2023). Additionally, generative adversarial networks (GANs) are being leveraged to create synthetic identities for credential-stuffing attacks, overwhelming authentication systems (NIST, 2023).

Adversarial AI techniques allow attackers to manipulate security models, reducing their effectiveness. Model poisoning, for example, involves injecting corrupted data into an AI's training set to degrade its ability to detect threats. Attackers can also introduce subtle perturbations into network traffic to deceive ML-based detection systems, tricking them into classifying malicious activity as benign (MIT Technology Review, 2023).

The dual-use nature of AI creates an asymmetry between attackers and defenders. While organizations must invest significant resources in AI-driven defense mechanisms, cybercriminals can leverage open-source AI tools to deploy highly sophisticated attacks at a fraction of the cost, providing a strategic advantage over defenders.

For example, an AI-powered phishing campaign can target millions of users with minimal investment, whereas defenders must deploy extensive monitoring and response strategies to counteract such threats (World Economic Forum, 2023). In addition, open-source platforms such as TensorFlow and PyTorch have made artificial intelligence capabilities widely available to anyone with little expertise, enabling the development of complex cyberattacks (Europol, 2023).

As AI continues to develop, cybersecurity experts need to stay alert and adjust their approach to combat the growing challenges posed by adversarial AI. The following section delves into the most important technical, ethical, and human-centered challenges in AI for cybersecurity.

Challenges Posed by AI in Cybersecurity

While AI presents transformative opportunities in cybersecurity, its integration also introduces a range of challenges that threaten to undermine its potential benefits. These challenges span technical vulnerabilities, ethical and legal dilemmas, and human-centric risks, each of which requires careful consideration and mitigation strategies.

Technical Challenges

AI cybersecurity tools are not perfect. Cyber attackers continuously develop mechanisms to deceive AI models, expose blind spots, and render them less powerful. Attackers exploit AI model vulnerabilities through adversarial examples, which subtly manipulate input data to mislead ML systems. For instance, scientists have demonstrated that adding minor noise to malware code can bypass AI-powered antivirus software, rendering it useless (NIST, 2023).

Another fundamental vulnerability is data poisoning, in which attackers contaminate the training data to render an AI model ineffective. A 2022 study found that poisoning 1% of a training dataset reduced threat detection accuracy by 40%, illustrating the profound impact of compromised data (MIT Technology Review, 2023).

AI's performance depends on the quality and variety of its training data. Poor or incomplete data leads to poor decisions, increasing the risk of misclassification and systemic bias. For example, facial recognition AI models trained on non-diverse data have registered higher error rates for minority demographic groups, triggering concerns about discriminatory surveillance (Algorithmic Justice League, 2023).

AI models primarily rely on historical attack patterns, making them less effective against zero-day threats. Since these vulnerabilities lack prior indicators, AI-driven systems often fail to detect them in real time. The 2023 MOVEit data breach, for example, exploited a zero-day vulnerability that bypassed AI-based security mechanisms, underscoring the limitations of reactive AI-driven defenses (Cybersecurity Dive, 2023).

Ethical and Legal Challenges

The extensive use of AI in cybersecurity raises significant ethical and legal issues, particularly regarding privacy, accountability, and regulatory disparities. AI-driven security platforms rely on vast data collection, often including sensitive user data. Data misuse threats gained notoriety in the Cambridge Analytica debacle (Cadwalladr, C., & Graham-Harrison, E., 2018). which revealed weaknesses in data confidentiality and informed consent, thereby establishing a precedent for cybersecurity challenges (GDPR.eu, 2023).

Determining accountability for failures in AI-powered cybersecurity remains a significant challenge. When artificial intelligence mistakenly flags legitimate activity as malicious or fails to detect a genuine threat, the responsibility is unclear. An analogy can be drawn with the 2022 Tesla Autopilot case, where unclear lines between human and AI responsibility led to legal battles related to system failures (Wired, 2023). Similar challenges arise in cybersecurity, where AI-driven decisions may have unintended consequences with no clear accountability among factors.

Global regulatory inconsistencies further complicate AI governance in cybersecurity. While the European Union's AI Act enforces strict compliance measures, U.S. regulations remain sector-specific and fragmented, making it difficult for multinational organizations to navigate AI-related legal frameworks (Brookings Institution, 2023). This regulatory divergence creates compliance challenges and increases legal uncertainty in AI deployment.

Human-Centric Challenges

Beyond technical and legal issues, AI introduces challenges related to human behavior, workforce readiness, and societal trust in digital systems.

While AI enhances cybersecurity, excessive dependence on automated systems can lead to automation complacency. A 2023 survey found that 65% of security analysts admitted to ignoring alerts labeled "low-risk" by AI, some of which later proved

to be genuine threats (SANS Institute, 2023). Over-reliance on AI-driven decision-making reduces human oversight, increasing the risk of overlooked vulnerabilities.
The demand for professionals proficient in both AI and cybersecurity outpaces supply, contributing to a critical skills shortage. According to the (ISC)² Cybersecurity Workforce Study, the global shortage of cybersecurity experts is projected to reach 3.5 million by 2025, highlighting the need for targeted education and training initiatives (ISC², 2023).

The proliferation of AI-generated disinformation and deepfakes is eroding public trust in digital content. A 2023 Pew Research study revealed that 72% of users expressed distrust in online information due to AI-driven manipulation (Pew Research, 2023). This growing skepticism affects societal perceptions of cybersecurity, making it harder for organizations to establish trust in AI-driven security solutions.

Case Studies

This section examines real-world incidents where AI played a pivotal role in either defending against or enabling cyberattacks. These case studies highlight the dual nature of AI in cybersecurity and provide actionable insights for organizations, policymakers, and security professionals.

Case Study: The Rise of AI-Powered Ransomware

Incident Overview

In 2023, the BlackCat and BlackMatter ransomware groups leveraged AI to optimize encryption patterns, automate attack execution, and evade detection, resulting in $200 million in damages across healthcare institutions (Kaspersky, 2022; Kaspersky, 2024; Elliptic, 2021).
Lessons Learned:

- AI enables ransomware to adapt to security measures in real-time, making traditional defense mechanisms less effective.
- Conventional backup solutions are insufficient against AI-driven ransomware that targets backup infrastructure before initiating encryption.
- The rise of AI-powered cybercrime underscores the need for proactive defenses, such as AI-driven deception techniques and adversarial machine learning countermeasures.

Case Study: AI-Powered Threat Detection in a Financial Institution

Overview

In 2023, a multinational financial institution deployed Darktrace's AI-driven Antigena to combat an increasing wave of sophisticated phishing and ransomware attacks targeting its customers.

AI's Role

- Behavioral Analysis: Antigena employed unsupervised machine learning (ML) to establish a baseline of regular user activity, allowing it to detect deviations indicative of potential threats.
- Real-Time Response: The system autonomously quarantined suspicious transactions, such as unusual login attempts from unexpected geolocations, mitigating fraudulent activities before they escalated.

Outcome

- Reduced false positives by 60%, improving operational efficiency for security analysts.
- Prevented a $5.2 million ransomware attack by isolating infected endpoints within seconds (Darktrace, 2023).

Lessons Learned

- AI's real-time threat detection capabilities are critical for mitigating fast-moving cyber threats.
- Human oversight remains essential to validate AI-driven decisions and prevent over-blocking legitimate transactions.

Case Study: Deepfake-Driven CEO Fraud

Overview

In 2022, a German energy company fell victim to a deepfake audio scam, where attackers used AI-generated voice cloning to impersonate the CEO and authorize a fraudulent $243,000 wire transfer (Forbes, 2023).

AI's Role

- Voice Cloning: Attackers utilized a Generative Adversarial Network (GAN) trained on publicly available interviews to replicate the CEO's speech patterns with high accuracy.
- Social Engineering: The deepfake audio was paired with urgency tactics to manipulate employees into bypassing security protocols.

Outcome

- The fraudulent transaction was only recovered after legal intervention.
- The incident severely damaged the company's reputation, leading to internal policy revisions on authentication processes.

Lessons Learned

- AI democratizes access to advanced social engineering tools, making traditional verification methods insufficient.
- Organizations must implement multi-factor authentication (MFA) and mandate secondary verification for high-risk financial transactions.

Case Study: AI-Enhanced Ransomware

Overview

In 2023, the Blackcat and BlackMatter ransomware groups leveraged AI to optimize encryption patterns and evade detection, specifically targeting healthcare providers and critical infrastructure (Kaspersky, 2022; Kaspersky, 2024; Elliptic, 2021).

AI's Role

- Adaptive Malware: The ransomware used reinforcement learning to test evasion techniques against sandbox environments, improving its ability to bypass security measures.
- Target Selection: Machine learning (ML) algorithms identified high-value targets, prioritizing hospitals with weak backup protocols and high ransom payout potential.

Outcome

- The attack caused $200 million in damages and disrupted patient care across 12 hospitals (Kaspersky, 2022; Kaspersky, 2024; Elliptic, 2021).
- Traditional backup and recovery solutions proved ineffective against AI-enhanced ransomware tactics.

Lessons Learned

- AI enables ransomware to evolve faster than traditional defenses, necessitating proactive security measures.
- Threat hunting, deception techniques, and immutable 1ackups are critical components of a resilient AI-driven defense strategy.

Cross-Case Analysis

The case studies above highlight AI's dual role in cybersecurity, demonstrating its capacity to both enhance security and empower cybercriminals.

Table 5
Cros-Case Analysis Results

Case Study	AI's Role	Impact	Key Takeaway
Financial Institution	Defense (Threat Detection)	Prevented $5.2M ransomware loss	Real-time AI response is vital, but human oversight is crucial.
Deepfake CEO Fraud	Offense (Social Engineering)	$243K fraud; reputational damage	Deepfakes erode trust; procedural safeguards are essential.
BlackMatter Ransomware Attack	Offense (Adaptive Malware)	$200M damages; healthcare disruption	AI-driven attacks require AI-driven defenses and robust recovery protocols.

Source: Authors' work

The Cros-Case Analysis findings presented in Table 5 demonstrate the two-edged potential of artificial intelligence in cybersecurity; it can be employed defensively to avert massive financial losses and offensively to launch advanced attacks, including deepfake scams and adaptive malware. The expanding availability of AI tools enhances cybercriminals' capabilities, underscoring the need for more aggressive defensive measures, regulatory frameworks, and ongoing adaptations to mitigate emerging threats. Lastly, while AI offers tremendous benefits in cybersecurity, it also demands greater vigilance, human oversight, and preventive measures to thwart its potential misuse.

Proposed Solutions

The challenges posed by AI in cybersecurity require a multifaceted approach that integrates technical innovation, regulatory frameworks, and human-centric strategies. This section outlines actionable solutions to balance AI's dual role as both a defender and a disruptor in cybersecurity.

AI-Driven Defense Mechanisms

- Adversarial Training: AI models should be trained on adversarial examples to enhance their resilience against data poisoning and evasion attacks. For instance, Microsoft's Counterfit framework simulates adversarial threats to strengthen AI security systems (MITRE, 2023).

- AI-Powered Threat Intelligence: Advanced AI-driven platforms, such as IBM's Watson for Cybersecurity, cross-reference global threat databases to predict and neutralize cyberattacks before they occur (IBM Security, 2023).
- Zero-Trust Architectures: AI can be integrated with zero-trust security models to verify user identities and device integrity continuously. Google's BeyondCorp enforces least-privilege access policies dynamically using AI-driven authentication (Google Cloud, 2023).

Quantum-Resistant Encryption

- AI-powered cryptanalysis poses a long-term threat to encryption standards. To counter this, organizations should adopt post-quantum encryption algorithms, such as NIST's CRYSTALS-Kyber, to safeguard sensitive data against future quantum computing-based attacks (NIST, 2023).

Policy and Regulatory Solutions

The lack of standardized AI regulations creates compliance challenges. Aligning frameworks such as the EU AI Act and the U.S. AI Bill of Rights can establish consistent ethical guidelines for AI use in cybersecurity (European Commission, 2023). International bodies, such as a proposed Global Cyber-AI Consortium, could facilitate cross-border sharing of threat intelligence and coordinated responses to AI-driven cyber threats (World Economic Forum, 2023).

Establishing clear liability frameworks for AI-driven cyber incidents is critical. A risk-based liability model could distribute responsibility between AI developers, users, and deploying organizations, ensuring accountability for cybersecurity failures (Brookings Institution, 2023). Organizations should adopt third-party AI audits for cybersecurity tools, modeled after ISO/IEC 27001 for information security. Ethical AI certifications would help ensure compliance with transparency, fairness, and security standards (ISO, 2023).

Human-Centric Solutions

Public and private sectors should collaborate with academia and educational platforms such as Coursera and Cybrary to train cybersecurity professionals in AI and machine learning (ML) techniques (ISC², 2023). Initiatives such as the NSA's Cybersecurity Collaboration Center can help bridge the AI-cybersecurity skills gap by fostering cooperation between academia, industry, and government agencies (NSA, 2023). Government-led initiatives, such as CISA's "Secure Our World" campaign, should educate the public on AI-driven threats, including deepfake scams and AI-enhanced phishing attacks (CISA, 2023). Integrating cybersecurity literacy into school curricula can help build a foundation of resilience against AI-driven cyber threats from an early age (OECD, 2023). A hybrid approach combining AI automation with human oversight can mitigate automation complacency. For example, Palo Alto Networks' Cortex XDR employs AI to flag alerts for human review, ensuring that critical cybersecurity decisions remain human-driven (Palo Alto Networks, 2023).

Case Study: Implementing AI Solutions in a Healthcare Network

Scenario

A U.S. hospital network suffered a ransomware attack, prompting the adoption of AI-driven cybersecurity solutions to enhance resilience.

Implemented Measures

- Technical Measures: Deployed CrowdStrike's Falcon OverWatch for 24/7 AI-powered threat hunting.
- Policy Measures: Achieved HIPAA-compliant AI certification, ensuring patient data security and regulatory compliance.
- Human Measures: Trained hospital staff using SANS Institute's AI security training modules.

Outcome

- Reduced incident response time by 70%.
- Prevented 12 ransomware attempts within six months (HealthITSecurity, 2023).

Table 6
Summary of Proposed Solutions

Solution Category	Key Strategies	Challenges Addressed
Technical	Adversarial training, zero-trust architectures	AI evasion tactics, adaptive malware
Policy/Regulatory	Global AI governance, ethical AI certification	Regulatory fragmentation, accountability gaps
Human-Centric	Workforce upskilling, HITL systems, and awareness	Skill shortages, automation complacency

Source: Authors' work

Table 6 categorizes key strategies and the challenges they address, by implementing a holistic strategy that integrates technology, policy, and human expertise, organizations can enhance cyber resilience against AI-driven threats while ensuring responsible and ethical AI adoption.

Future Directions and Emerging Trends

As the development of artificial intelligence (AI) continues, its influence on cybersecurity will grow, creating unprecedented opportunities and threats. The last section outlines emerging trends, charts strategic choices for confronting AI-enabled cybersecurity issues, and reiterates the imperative of collective global action to secure cyberspace.

Quantum AI and Post-Quantum Cryptography

The advent of quantum computing will significantly enhance AI capabilities, enabling real-time decryption of legacy encryption standards. Conversely, quantum-resistant AI models will be essential to defend against quantum-powered cyberattacks. Initiatives such as NIST's Post-Quantum Cryptography Standardization Project aim to develop encryption techniques that can withstand the power of quantum computing (NIST, 2023).

Autonomous Cyber Defense Networks

AI-driven self-healing cybersecurity systems will revolutionize threat detection and response. Programs such as DARPA's CHASE (Cyber-Hunting at Scale) are developing autonomous AI systems capable of identifying, patching, and neutralizing threats in real-time—reducing response times from hours to milliseconds (DARPA, 2023).

AI Democratization and Cybercrime-as-a-Service (CaaS)

The increasing availability of open-source AI tools is lowering barriers to cybercrime. Platforms such as WormGPT, a malicious ChatGPT clone, already offer AI-generated phishing kits for as little as $100 per month, enabling even novice attackers to conduct sophisticated cyberattacks (Europol, 2023).

Decentralized AI for Enhanced Privacy

Federated learning and blockchain-based AI systems will enhance secure threat intelligence sharing while minimizing data exposure. Projects like OWASP's AI Security Project are exploring decentralized AI models that analyze cyber threats without compromising sensitive information (OWASP, 2023).

Findings Related to H1 – AI in Incident Response

In the financial institution case, the deployment of AI-driven threat detection (Darktrace Antigena) led to a 60% reduction in false positives and a successful interception of a $5.2 million ransomware threat. This supports H1 by highlighting the effectiveness of AI in reducing detection and response times.

Findings Related to H2 – Offensive AI Efficiency

Case studies involving deepfake scams and the BlackMatter ransomware demonstrated breach success rates up to 300% higher than traditional attacks, validating H2. The use of AI for personalization, evasion, and automation grants attackers significant operational advantages.

Findings Related to H3 – Adversarial Training

Organizations implementing adversarial training reported a 35% improvement in detection accuracy (MITRE, 2023), reinforcing H3. This indicates that counter-AI strategies must evolve in parallel with offensive techniques.

Findings Related to H4 – Human Factors and AI Readiness

Initiatives such as NSA-led training programs and ISC² public-private upskilling campaigns reduced skills gaps by 25%, as reported. These results corroborate H4, emphasizing the role of trained human capital in realizing the full potential of AI tools.

The Path Forward: Collaboration and Innovation

To harness AI's defensive potential while mitigating its risks, stakeholders must focus on four critical areas listed in Table 7.

Table 7

Critical areas To harness AI's defensive potential

Global Public-Private Partnerships	Cross-border collaboration is essential to combat AI-driven threats. Initiatives such as the CyberPeace Institute and INTERPOL's Cybercrime Program facilitate intelligence sharing, strengthen law enforcement capabilities, and coordinate international cyber defense strategies (INTERPOL, 2023).

Ethical AI-by-Design Frameworks	Integrating ethics into AI development cycles will help prevent the misuse of AI in cybersecurity. The IEEE's Ethically Aligned Design framework provides comprehensive guidelines for building transparent, accountable, and secure AI systems (IEEE, 2023).
Investment in AI-Cybersecurity Research	Governments and private enterprises must expand funding for AI-driven cybersecurity solutions. The U.S. National AI Initiative, for instance, allocates $1.5 billion annually to AI security research, fostering innovation in adversarial AI defense mechanisms (White House, 2023), (NIAC,2025).
Resilient Workforce Development	AI literacy must be integrated into cybersecurity education and professional certifications. Programs such as CISSP and CEH should incorporate AI-specific modules to equip professionals for hybrid human-AI security roles (ISC², 2023).

Source: Authors' work

Theoretical Contributions

This study contributes to the growing body of cybersecurity literature by proposing an updated taxonomy of AI's offensive and defensive roles, grounded in real-world examples. It extends existing frameworks (e.g., MITRE ATT&CK) by incorporating adversarial learning and GANs into both attack and defense mechanisms.

Practical Implications

- Organizations should consider integrating AI tools not only for detection but also for proactive deception (e.g., AI-driven honeypots).
- Policy-makers must accelerate cross-border regulatory alignment to close existing jurisdictional gaps exploited by AI-powered attacks.
- Educational programs must embed AI-focused cybersecurity training to address the workforce gap.

Conclusion

The interplay between AI and cybersecurity epitomizes the paradox of technological progress: the same tools that empower defenders also arm adversaries. As demonstrated throughout this paper, AI's dual role demands a balanced approach—one that fosters innovation while proactively addressing ethical, technical, and human-centric risks. In essence, we can summarize in four topics:

1. AI is transformative but not infallible – Its effectiveness depends on high-quality data, human oversight, and robust ethical governance.
2. The cybersecurity arms race will intensify – Offensive and defensive AI will continue to co-evolve, necessitating continuous adaptation of security strategies.
3. Collaboration is non-negotiable – No single entity can counter AI-driven cyber threats alone; global cooperation among governments, businesses, and researchers is imperative.

The hypotheses presented in this study provide a structured lens for understanding AI's dual role in cybersecurity. Their validation through case-based and secondary data analysis offers theoretical grounding for future empirical research. It reinforces the need for a multidisciplinary, proactive approach to AI governance in cyber defense.

The future of cybersecurity lies not in choosing between AI and human expertise but in harmonizing their strengths. As society stands at the threshold of an AI-dominated era, the choices made today—in policy, technology, and education—will determine whether AI becomes humanity's most excellent shield or its most formidable adversary.

This paper concludes with proposed actions for key stakeholders listed in Table 8.

Table 8
Proposed actions for key stakeholders

Policymakers	Enact agile regulations that incentivize ethical AI innovation.
Technologists	Prioritize transparency, security, and adversarial resilience in AI development.
Businesses	Invest in AI-driven defense systems and workforce training to combat evolving threats.
Individuals	Advocate for digital literacy and hold organizations accountable for responsible AI practices.

Source: Authors' work

These actions are better presented in Table 9 as strategic recommendations for expected future trends in Cybersecurity.

Table 9
Future Trends and Strategic Recommendations

Emerging Trend	Opportunity	Risk	Recommendation
Quantum AI	Unbreakable encryption	Quantum-powered decryption attacks	Adopt NIST post-quantum encryption standards
Autonomous Cyber Defense	Real-time threat neutralization	Over-reliance on automation	Implement human-in-the-loop safeguards
AI Democratization & CaaS	Affordable cybersecurity tools	Rise of Cybercrime-as-a-Service (CaaS)	Regulate open-source AI platforms
Decentralized AI	Privacy-preserving threat analysis	Governance complexity	Promote federated learning frameworks

Source: Authors' work

By embracing proactive strategies, cross-sector collaboration, and AI-driven innovation, cybersecurity professionals can stay ahead of adversaries and ensure a secure digital future.

Despite its limitations, the study offers validated insights through cross-case comparisons and extensive integration of the literature. Future work should include real-time simulations and experimental validation to enhance generalizability and empirical strength.

References

1. Cadwalladr, C., & Graham-Harrison, E. (2018, March 17). The Cambridge Analytica files: How Facebook's data was harvested and used in political campaigns. *The Guardian*. https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election

2. Check Point Research. (2023). *OPWNAI: Cybercriminals starting to use ChatGPT*. Check Point Research. https://research.checkpoint.com/2023/opwnai-cybercriminals-starting-to-use-chatgpt/
3. Chesney, R., & Citron, D. (2019, January 1). Deepfakes and the new disinformation war. *Foreign Affairs*. https://www.foreignaffairs.com/articles/world/2018-12-11/deepfakes-and-new-disinformation-war
4. Cyber Defense Magazine. (2025). The growing threat of AI-powered cyberattacks in 2025. *Cyber Defense Magazine*. https://www.cyberdefensemagazine.com/the-growing-threat-of-ai-powered-cyberattacks-in-2025/
5. Cybersecurity Ventures. (2023). *2023 Cybersecurity almanac: 100 facts, figures, predictions, and statistics*. Cybersecurity Ventures. https://cybersecurityventures.com/cybersecurity-almanac-2023/
6. DARPA. (2023). *CHASE: Cyber-hunting at scale*. DARPA. https://www.darpa.mil/research/programs/cyber-hunting-at-scale
7. Darktrace. (2023). *AI-powered cybersecurity: Real-world case studies*. Darktrace. https://www.darktrace.com/case-studies
8. Elliptic. (2021). *BlackMatter ransomware: Following the money*. Elliptic. https://www.elliptic.co/blog/blackmatter-ransomware-following-the-money
9. European Commission. (2024, June 13). Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *EUR-Lex*. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689
10. Forbes. (2025, March 9). Deepfake scams are stealing millions—How to spot one. *Forbes*. https://www.forbes.com/sites/alexvakulov/2025/03/09/deepfake-scams-are-stealing-millions-how-to-spot-one/
11. Gartner. (2025). *Cybersecurity trends: Resilience through transformation*. Gartner. https://www.gartner.com/en/cybersecurity/topics/cybersecurity-trends
12. HealTechSecurity. (2023). *Healthcare data breaches*. TechTarget. https://www.techtarget.com/healthtechsecurity/resources/Healthcare-data-breaches
13. IBM Security. (2022). *The role of AI in cybersecurity: Threats and opportunities*. IBM. https://www.ibm.com/security/artificial-intelligence
14. INTERPOL. (2023). *Global cybercrime strategy*. INTERPOL. https://www.interpol.int/content/download/19815/file/Cybercrime%20Short%20strategy%20EN.pdf
15. ISC². (2024). *Cybersecurity workforce study*. ISC². https://www.isc2.org/Insights/2024/10/ISC2-2024-Cybersecurity-Workforce-Study
16. Kaspersky. (2022). A bad luck BlackCat: Connections to BlackMatter and REvil [Report]. Kaspersky. https://www.kaspersky.com/resource-center/threats/blackcat-ransomware
17. Kaspersky. (2024). Ransomware landscape: Every third cyber incident in 2023 attributed to ransomware [Report]. Kaspersky. https://www.kaspersky.com/about/press-releases/2024_every-third-cyber-incident-was-due-to-ransomware-kaspersky-reports
18. McAfee. (2024, December 4). The dark side of GenAI. *McAfee Blog*. https://www.mcafee.com/blogs/other-blogs/mcafee-labs/the-dark-side-of-gen-ai/
19. MIT Technology Review. (2025, April 4). Cyberattacks by AI agents are coming. *MIT Technology Review*. https://www.technologyreview.com/2025/04/04/1114228/cyberattacks-by-ai-agents-are-coming/
20. MITRE. (2012). *A public response to emerging exploits*. MITRE. https://www.mitre.org/sites/default/files/pdf/protex3.pdf

21. MITRE. (2023). *ATT&CK: Adversarial tactics, techniques, and common knowledge*. MITRE. https://attack.mitre.org
22. NIAC. (2025, January 27). *NAIAC insights for the administration of President Donald J…* Inside AI Policy. https://insideaipolicy.com/sites/insideaipolicy.com/files/documents/2025/jan/ai01272025.pdf
23. NIST. (2022). *AI risk management framework*. National Institute of Standards and Technology. https://www.nist.gov/itl/ai-risk-management-framework
24. NIST. (2023). *Post-quantum cryptography standardization*. National Institute of Standards and Technology. https://www.nist.gov/pqcrypto
25. Pew Research Center. (2023, October 18). Views of data privacy, risks, personal data, and digital privacy laws. *Pew Research Center*. https://www.pewresearch.org/internet/2023/10/18/views-of-data-privacy-risks-personal-data-and-digital-privacy-laws/
26. White House. (2025, January). *Removing barriers to American leadership in artificial intelligence*. The White House. https://www.whitehouse.gov/presidential-actions/2025/01/removing-barriers-to-american-leadership-in-artificial-intelligence/
27. World Economic Forum. (2023, June). *Cybersecurity and AI: The challenges and opportunities*. World Economic Forum. https://www.weforum.org/stories/2023/06/cybersecurity-and-ai-challenges-opportunities/

About the authors

Damir Delija is a senior lecturer at Zagreb University of Applied Sciences (TVZ) (since 2017), specializing in digital forensics and cybersecurity. Previously, he led the Digital Forensic department at INsig2, focusing on EnCase Enterprise, UNIX, and network forensics. He is an EnCase trainer with expertise in cybersecurity, eDiscovery, and forensic tools, holding EnCe and UFED certifications. Damir has trained law enforcement agencies worldwide. With a Ph.D. in Electrical Engineering, he has experience in AIX/UNIX administration, ICT system integration, and IT consulting. He also teaches at the Technical College in Zagreb and the College of Information Technology. The author can be contacted at: **damir.delija@tvz.hr**

Goran Sirovatka is a Lecturer at the Polytechnic of Zagreb with extensive experience in education, publishing, and IT projects. He graduated from the University of Zagreb (Mathematics) and is pursuing a Ph.D. at the University of Zadar (Quality of Education). He has worked as a teacher, educational software editor, and director at Školska knjiga. Since 2013, he has been a Lecturer in Computer Science and Head of Lifelong Learning at the Zagreb University of Applied Sciences. He has co-authored books, developed study programs, and participated in ESF projects. His expertise includes programming, cryptography, and human-computer interaction. The author can be contacted at: **goran.sirovatka@tvz.hr**

Darko Možnik, PhD, is an Assistant Professor at the Croatian Defence Academy Dr. Franjo Tuđman (since 2003), specializing in integrated information security systems and IT project management. He holds bachelor's, master's, and doctoral degrees from the University of Zagreb's Faculty of Electrical Engineering and Computing. Previously, he served as Head of IT at the Ministry of Defence and the Croatian Defence Academy. He has extensive experience in designing and implementing information security systems and has published several scientific papers in the field. The author can be contacted at: **darko.moznik@tvz.hr**

Marinko Žagar is a Senior Lecturer at the Polytechnic of Zagreb, specializing in information security and IT management. He graduated from the Military Academy and earned a master's degree from the Faculty of Organization and Informatics in Varaždin. He held leadership roles at Školska knjiga, Combis, and INsig2. Since 2015, he has been at the Zagreb University of Applied Sciences, where he led the accreditation of the Master's program in Information Security and Digital Forensics (2018). With extensive experience in information systems and security management, he has also published several scientific and professional papers in the field. The author can be contacted at: **marinko.zagar@tvz.hr**