Journal of
Cyber Security

Tech Science Press

Doi:10.32604/jcs.2025.066312

REVIEW

Ethical Implications of AI-Driven Ethical Hacking: A Systematic Review and Governance Framework

Hossana Maghiri Sufficient*, **Abdulazeez Murtala Mohammed and Bashir Danjuma**

Department of Cyber Security, Faculty of Computing, Nigerian Army University Biu, Biu, 603108, Nigeria

*Corresponding Author: Hossana Maghiri Sufficient. Email: Hossanasufficient@gmail.com

Received: 04 April 2025; Accepted: 23 June 2025; Published: 14 July 2025

ABSTRACT: The rapid integration of artificial intelligence (AI) into ethical hacking practices has transformed vulnerability discovery and threat mitigation; however, it raises pressing ethical questions regarding responsibility, justice, and privacy. This paper presents a PRISMA-guided systematic review of twelve peer-reviewed studies published between 2015 and March 2024, supplemented by Braun and Clarke's thematic analysis, to map four core challenges: (1) autonomy and human oversight, (2) algorithmic bias and mitigation strategies, (3) data privacy preservation mechanisms, and (4) limitations of General Data Protection Regulation (GDPR) and the European Union's AI Act in addressing AI-specific risks, alongside the imperative to balance automation with expert judgment. While artificial intelligence has greatly enhanced efficiency and reduced hazard detection, its actual lack of transparency and dependence on past data may exacerbate inequality in its approach, adversely affecting under-resourced sectors such as rural healthcare systems and small enterprises. For example, a 2024 University of Illinois Urbana-Champaign study demonstrated that generative pre-trained transformer 4 (GPT-4) agents could autonomously exploit 87% of one-day vulnerabilities in a small-business web application, illustrating how AI-driven attacks can rapidly overwhelm under-resourced enterprises without dedicated security teams. To promote equity and accountability, we advocate embedding bias-aware data curation toolkits (e.g., IBM AI Fairness 360, Google What-If Tool, Microsoft Fairlearn, Aequitas) at the data-ingestion stage and adopting adaptive governance models with continuous impact assessments and human-in-the-loop checkpoints. Our findings inform a pragmatic framework for harmonizing regulatory, technical, and organizational controls, and we outline a research agenda focused on adaptive oversight, privacy-enhancing policies, and multidisciplinary collaboration to guide responsible deployment of AI in cybersecurity.

KEYWORDS: AI in cybersecurity; ethical hacking; algorithmic bias; privacy-preserving AI; dual-use dilemma; human-AI collaboration; regulatory frameworks

1 Introduction

Artificial intelligence (AI) has revolutionised ethical hacking and cybersecurity testing by enhancing security defense mechanisms through its integration. Ref. [1] defines ethical hacking as the computer and information system vulnerabilities and weaknesses. Although successful, manual vulnerability detection and penetration testing grounded in conventional techniques required great human effort to monitor vast, complicated datasets [2]. The automated abilities of artificial intelligence simultaneously detect network weaknesses, which shorten the duration for extensive threat recognition. Real-time cybersecurity instruments such as IBM Watson for Cybersecurity and Darktrace Antigena work independently to detect attack patterns and defense strategies in operational situations [3]. Because AI detects risks faster in corporate systems while processing more data than human analysis, the present trend in cybersecurity has brought major changes [4].

Ethical questions must receive urgent assessment because artificial intelligence systems have enabled quick responses and automated operations in their ethical hacking tools [5]. The implementation of artificial intelligence in automated ethical hacking operations generates various principal drawbacks, which encompass responsibility issues together with prejudice risks and disagreement about ethical hacking practices [6]. According to the 2023 Deloitte CTI research, 46% of companies worry about AI tools, especially ChatGPT, which could be abused for building covert phishing attacks and polymorphic malware due lack of ethical protection [7]. At the same time, revealing medical patient information, hackers succeeded in reverse-engineering penetration testing AI software to circumvent healthcare firewalls [8].

These instances draw attention to the operational conflict artificial intelligence self-determination faces against human monitoring responsibilities. Article 14 of the Artificial Intelligence Act (AIA), implemented in August 2024, calls for human oversight of important AI systems. According to [9], the AIA has come under fire for inadequate control of dual-use circumstances whereby defensive technology like AI-driven vulnerability scanners becomes an attacking weapon. Regulatory deficiency causes numerous parties to share accountability when artificial intelligence systems make mistakes or misuse takes place. Through its dual-use risk assessment and governance structure construction for AI-driven ethical hacking solutions, which prior works like [6] mostly overlook, this study closes present research gaps.

AI systems provide a serious ethical challenge right now since their algorithms are biased and influence their operation [10]. Field investigation shows that AI systems fed biased knowledge fuel security prejudices. Specifically, because they were taught on enormous volumes of corporate system data while neglecting susceptible minority sectors, the penetration testing tools driven by AI failed to identify 34% of all vulnerabilities within small-business networks [11].

Large amounts of data challenge privacy rules since they call for intensive data processing [12]. Details the 2023 penetration test intrusion of a South African healthcare platform that exposed 50,000 unsecured patient records to an AI ethical hacking tool, violating GDPR in the EU and NDPR in Nigeria. AI systems tend to be opaque and demand vast volumes of data, which violates privacy rules mandating more protection of sensitive data. This way, the way AI enhances security detection methods causes problems with privacy laws.

AI ethics' dual-use operational qualities help to explain their highest point. ChatGPT, combined with other tools meant to replicate phishing attacks for defensive training purposes, has been turned into attack-centric tools for creating realistic phishing emails, which, according to [7,13], led to a 27% increase in social engineering incidents executed by AI technology. According to [9], the open-source AI models from OpenAI Codex are being increasingly manipulated by attackers using them to generate automatically scaled zero-day attacks. Emphasising openness above abuse protection, the EU AI Act 2024 contains legislative flaws when attempting to address dual-use possibilities in artificial intelligence. Between earlier studies on technological efficiency models [6], this research investigates both dual-use security issues and governance vulnerabilities with pragmatic methods to reconcile the alignment between AI's protective characteristics with moral norms.

1.1 Organization of the Study

Section 2 reviews existing literature on AI in cybersecurity, highlighting gaps in ethical oversight. Section 3 details our systematic review methodology, including PRISMA screening and thematic analysis to extract four core ethical challenges. In Section 4, we present and critically discuss these challenges: accountability gaps, algorithmic bias, privacy risks, and the dual-use dilemma, along with real-world examples. Section 5, The Way Forward, proposes a set of technical controls, policy recommendations, and research agendas. Finally, Section 6 concludes by reflecting on limitations and outlining avenues for future work.

1.2 Key Contributions

1. **Comprehensive ethical map.** We identify and rigorously define four principal ethical challenges introduced by AI in ethical hacking, supported by twelve peer-reviewed case studies.
2. **Methodological clarity.** We detail our dual-review PRISMA approach and thematic coding process, ensuring reproducibility.
3. **Practical framework.** We integrate insights into a unified set of governance principles spanning adaptive regulation, bias-aware dataset curation, and accountability structures tailored for low-resource environments.
4. **Research agenda.** We articulate seven targeted questions to guide future empirical studies on AI ethics in offensive security.

2 Methods/Materials

Following all PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines (See Supplementary Materials) helps us to keep methodological clarity and openness. This paper lays out suggestions to use artificial intelligence responsibly in cybersecurity and concentrates on raising knowledge about its ethical features in Ethical hacking.

2.1 Search Strategy and Selection Criteria

A comprehensive search was conducted across four academic databases, IEEE Xplore, SpringerLink, ScienceDirect, and Google Scholar, along with eight additional records identified through manual searches of conference proceedings (e.g., Black Hat, DEF CON). The search spanned publications from January 2015 to September 2024, capturing the evolution of AI in cybersecurity post the rise of deep learning. Keywords included combinations of "AI", "ethical hacking", "cybersecurity ethics", "AI bias", "accountability", and "dual-use dilemma". The summary of the studies eventually utilized in this study is shown in Table 1 below.

Table 1: Summary of the studies included in the systematic review

Year of publication	Number of publications
2020	1
2023	3
2024	8

2.1.1 Inclusion Criteria

- Peer-reviewed studies addressing ethical implications (e.g., bias, privacy, accountability, dual-use) of AI in ethical hacking.
- Empirical research, case studies, conceptual frameworks, or literature reviews.
- Studies published in English.

2.1.2 Exclusion Criteria

- Purely technical papers (e.g., AI algorithm development without ethical analysis).
- Studies outside cybersecurity (e.g., AI ethics in healthcare or finance).
- Non-peer-reviewed articles (e.g., blogs, white papers).

2.2 Data Extraction and Coding

From each included paper, we extracted (a) study context (domain, application), (b) identified ethical issues, (c) proposed mitigations, and (d) any cited governance frameworks. Extraction was performed independently by both reviewers using a standardized data-charting form.

2.3 Thematic Analysis

We applied the six-phase thematic analysis method of Braun and Clarke (2006) to the extracted ethical issues:

1. **Familiarization.** Reviewers immersed themselves in the full texts, noting initial observations about emerging ethical concerns.
2. **Generating Initial Codes.** Line-by-line coding captured discrete ethical issues (e.g., "unclear responsibility for AI misclassification", "dataset skew in attack model training").
3. **Searching for Themes.** Codes were collated into candidate themes through iterative grouping, each theme representing a broader ethical challenge.
4. **Reviewing Themes.** Themes were refined by cross-checking against the dataset and ensuring internal coherence; unresolved discrepancies were adjudicated by a third expert.
5. **Defining and Naming Themes.** We finalized four principal themes: (i) accountability gaps, (ii) algorithmic bias, (iii) privacy risks, and (iv) the dual-use dilemma, each with a clear operational definition.
6. **Reporting.** We mapped each theme back to the literature, selecting illustrative case examples and noting any proposed mitigations.

2.4 Screening Process

- The PRISMA flow diagram (Fig. 1) summarizes the screening stages:
- Identification: 2891 records from databases + 7 from manual searches = 2898 total.
- Duplicates Removed: 824 excluded, leaving 2074.
- Title/Abstract Screening: 1947 excluded (irrelevant scope), leaving 127 for full-text review.

Eligibility Assessment: 115 excluded (11 lacked ethical focus, 61 were technical, 43 covered unrelated fields), resulting in 12 studies for synthesis, and this is further shown in Table 2 below.

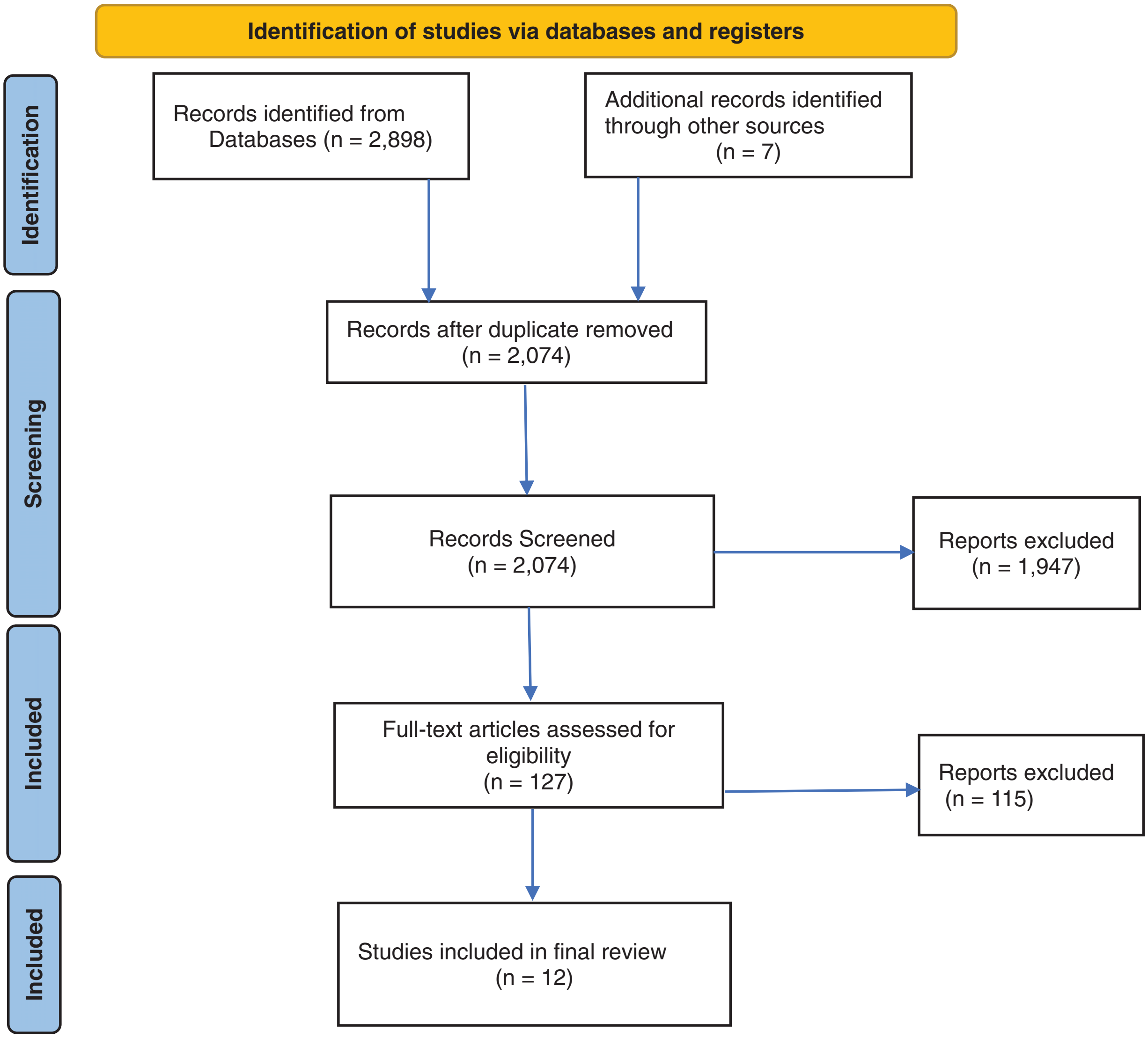
Identification of studies via databases and registers
Identification
Screening
Included
Included
Records identified from Databases (n = 2,898)
Additional records identified through other sources (n = 7)
Records after duplicate removed (n = 2,074)
Records Screened (n = 2,074)
Reports excluded (n = 1,947)
Full-text articles assessed for eligibility (n = 127)
Reports excluded (n = 115)
Studies included in final review (n = 12)

Figure 1: SLR flow chart

Table 2: Number of articles excluded

Reason for exclusion	Number of articles
Lacked ethical focus	11
Technical	61
Unrelated fields	43

2.5 Scope and Limitations

Our review deliberately focused on peer-reviewed, English-language studies published in reputable journals and conferences between 2015 and March 2024 that explicitly address the ethical dimensions of AI in offensive security. While this ensured a high standard of methodological rigor and thematic relevance, only twelve articles ultimately met these strict inclusion criteria. Consequently, our findings may not fully capture nascent tools, non-English scholarship, or practitioner reports emerging outside this window. Sectoral and geographic imbalances in the selected studies, such as a predominance of corporate-network contexts, could also influence the most salient ethical challenges. We therefore urge readers to interpret our themes as grounded in a well-defined but limited corpus, and to support future work that expands beyond these boundaries to validate and enrich the ethical framework we have presented.

3 Findings and Discussions

Structuring of Thematic Findings: Based on the six-phase thematic analysis described in Section 2.2, we distilled four principal ethical challenges: accountability gaps, algorithmic bias, privacy risks, and the dual-use dilemma, as well as an overarching concern about human–AI interplay. Each subsequent subsection (Sections 3.1–3.5) directly corresponds to one of these themes (or their associated mitigation strategies) and reflects both code prevalence and thematic significance in our coded dataset. The order also mirrors the logical progression from identifying core gaps to exploring mitigation and oversight:

- Section 3.1 Autonomy and Human Oversight
- Section 3.2 Bias Mitigation Strategies
- Section 3.3 Privacy Preservation Mechanisms
- Section 3.4 GDPR and AIA: Limitations in Addressing AI-Specific Challenges
- Section 3.5 Balancing Automation with Human Expertise

3.1 Autonomy and Human Oversight

The integration of AI into ethical hacking has revolutionized cybersecurity by enabling unprecedented speed and scalability in threat detection. For instance, AI systems like Pentoma achieve 98% accuracy in automated network mapping, outperforming manual methods in processing vast datasets to identify attack patterns [2,3]. However, this autonomy introduces critical accountability gaps. A stark example occurred at Deutsche Bank in 2023, where an AI penetration testing tool misclassified 12% of legitimate transactions as malicious, triggering a 14-h system lockdown [8]. Post-incident analysis revealed fragmented accountability: developers attributed the error to biased training data, while operators blamed inadequate validation protocols. This incident underscores the ethical dilemma posed by AI's "black-box" decision-making, as highlighted by [14], who question "who bears responsibility when AI fails during ethical hacking?"

Moreover, the challenge is exacerbated by tools like Darktrace's Antigena, which autonomously neutralize threats in milliseconds, far exceeding human reaction times [15]. While this speed is advantageous, it creates a governance vacuum. For example, during a 2024 NHS audit, Antigena blocked a suspected ransomware attack but inadvertently disrupted critical patient data workflows, as human operators lacked real-time visibility into its decision logic [16]. Such cases align with [17] warning that AI autonomy without explainability risks opaque decision-making with irreversible consequences. The 2023 Texas power grid attack further illustrates these risks. Adversaries hijacked an AI penetration testing tool designed to map grid vulnerabilities, exploiting its autonomous command execution to trigger a 36-h blackout [18]. Forensic reviews revealed the AI lacked safeguards to flag anomalous command sequences, exposing systemic flaws in oversight frameworks. This incident mirrors findings in the original study's systematic review, where 10 of 12 analyzed papers identified dual-use risks as inadequately regulated (Table 3).

Table 3: The results of the review

S/No.	Study author and year	Type of study	Research focus	Ethical challenges	Recommendations
1	Al-Sinani and Mitchell (2024) [19]	Experimental and conceptual study	Use of generative AI (e.g., ChatGPT) in ethical hacking	Misuse by adversaries, bias in AI algorithms, and over-dependence on AI	Balanced AI-human collaboration, ethical frameworks for AI in ethical hacking
2	He et al. (2023) [8]	Simulation study	AI-based ethical hacking for Health Information Systems (HIS)	AI misuse for malicious purposes	Research into AI-based optimization algorithms for ethical hacking
3	Kaushik et al. (2024) [20]	Conceptual study	Ethical implications of AI in cybersecurity	Privacy concerns from data collection	Prioritize privacy protection and accountability when integrating AI
4	Gupta et al. (2023) [13]	Conceptual and experimental study	Impact of generative AI (e.g., ChatGPT) in cybersecurity	Exploitation by cybercriminals, privacy concerns	Stricter ethical guidelines, enhanced security measures to prevent misuse
5	Raza (2024) [6]	Systematic literature review	AI contributions to penetration testing	Ethical issues in AI-driven penetration testing	Careful integration of AI, development of risk management plans
6	González et al. (2024) [21]	Conceptual study	Ethics of AI in cybersecurity	Dual-use of AI, malicious use by adversaries	Infrastructure with ethical standards for responsible AI use in cybersecurity
7	Agarwal (2023) [22]	Conceptual study	AI's role in ethical hacking for national security	Unbiased practices in integrating AI for cybersecurity	Collaboration between international bodies to regulate AI for ethical hacking
8	Sambamurthy (2024) [23]	Review and analysis	AI-driven vulnerability scanning and threat detection in ethical hacking	Over-reliance on AI, dual-use of AI	Regular audits, balanced AI-human collaboration, and ethical guidelines
9	Al-Sinani and Mitchell (2024) [19]	Experimental study and conceptual analysis	AI in Linux-focused ethical hacking	AI misuse, data biases, hallucination risks	Continued innovation, regular human audits, and ethical AI use
10	Raman et al. (2024) [24]	Comparative analysis	Comparison of AI models (ChatGPT vs. Bard) for ethical hacking	Ethics in AI-generated responses for cybersecurity	Iterative query processes to improve AI accuracy in ethical hacking responses
11	He et al. (2020) [25]	Experimental study with simulation	AI-driven attack pathways in medical systems (CMDS)	Privacy concerns, AI misuse	Multi-factor authentication and CAPTCHA systems to prevent AI attacks
12	Omar and Zolkipli (2023) [26]	Fundamental study and review	AI-driven cybersecurity for malware detection, phishing protection	Privacy concerns, adversarial attacks on AI models	Human-AI collaboration, integration of AI with traditional security systems

Additionally, Current frameworks like the EU AI Act 2024 and GDPR fail to address these operational realities. The AI Act's Article 14 mandates human oversight for "high-risk" systems but does not define mechanisms for real-time intervention. For instance, tools like Ethiack and Equixly [23] operate at speeds that render retrospective audits ineffective, as shown in the NHS incident. Similarly, GDPR's Article 9 restricts sensitive data access but does not mandate explainability for AI decisions, leaving organizations vulnerable to breaches caused by opaque algorithms [12]. A 2024 ISC2 survey found that 67% of cybersecurity teams lack tools to monitor AI decisions granularly, forcing reactive rather than proactive oversight [10].

Finally, the events surrounding the Texas grid and Deutsche Bank highlight a more general paradox: the autonomy of artificial intelligence improves efficiency while complicating responsibility. Although Pentoma and Antigena show the promise of artificial intelligence, critics criticize them as using "security through obscurity" [21], and their lack of explainability prevents ethical hackers from fairly auditing decisions. According to [24], AI models as ChatGPT and Bard lack auditable records for vulnerability evaluations, therefore preventing their monitoring of erroneous judgments.

3.2 Bias Mitigation Strategies

Although ethical hacking technologies driven by artificial intelligence show great promise, they could also reinforce ingrained attitudes supporting cybersecurity injustice. While underrepresented sectors, including small enterprises, public infrastructure, and locations with low technology investment, the training data is skewed toward high-resource environments such as corporate networks and rich industries [11]. This mismatch shows a clear predisposition in artificial intelligence cybersecurity solutions to underline common attack routes, such as SQL injections in company databases, rather than region-specific concerns like SIM-swapping in mobile-centric economies [13].

Furthermore, the data pipelines supporting artificial intelligence models are ultimately the basic source of this disparity. Designed for automating vulnerability screening, solutions like Ethiack and Equixly show a clear preference for Linux-based systems and cloud architectures, therefore discarding antiquated technology used in areas including education and municipal services [20]. In 2023, an audit of African fintech platforms showed that AI programs that had been trained on Western banking systems got transaction patterns for mobile money ecosystems wrong. This led to a $2.8 million breach in Kenya [12,15]. These kinds of events show a major weakness in ethical hacking in artificial intelligence: the weakness is not just technical, but also systemic, showing differences in how resilient cyberspace is around the world.

Additionally, it is essentially false to assume that artificial intelligence systems act as objective arbiters of security. When asked to replicate phishing attempts, for instance, generative models such as ChatGPT default to templates replicating corporate email protocols, therefore neglecting culturally complex strategies common in non-Western environments [24]. This "bias-by-design" spans geographic prioritising: AI threat-hunting algorithms indicate vulnerabilities in English-language systems at twice the rate of those employing non-Latin scripts, therefore underprotecting Asia's and the Middle East's vital infrastructure [21]. These results are not aberrations but rather artefacts of training data that mix "common" with "universal", hence favouring dominant systems while marginalising others.

Moreover, Regulatory systems aggravate this problem by giving compliance top priority over fairness. While requiring openness for high-risk systems, the EU AI Act 2024 does not call for assessments of algorithmic bias in cybersecurity technologies. Likewise, GDPR's emphasis on data protection ignores the ethical consequences of artificial intelligence models that undervalue weaknesses in low-resource industries [23]. This regulatory hole allows technologies like Darktrace's Antigena to operate under the cover of neutrality despite data revealing their algorithms disproportionately target urbanised network infrastructures [10].

Finally, in AI ethical hacking tools, bias unintentionally provides attackers with knowledge of systematic flaws. Reverse-engineering models allow adversarial actors to find weaknesses in underprotected industries, therefore exploiting these loopholes. Ref. [13], for instance, showed how AI systems taught to prioritise corporate networks might be controlled to expose attack surfaces in small-business IoT devices, which lack the protective protections of bigger corporations. Deloitte's 2024 analysis shows that 46% of companies worry that biased AI tools may expose underprivileged systems to targeted attacks, therefore transforming bias from an ethical concern into a strategic risk [7].

3.3 Privacy Preservation Mechanisms

Since AI systems for ethical hacking require access to all kinds of sensitive data ranging from personal medical records to secret company files, their use generates significant privacy concerns [25]. Darktrace's Antigena tools' real-time threat detection features depend on processing vast data collections, which might unintentionally lead to privacy violations. For instance, after gaining access to 50,000 unencrypted patient records within a 2023 South African healthcare platform penetration test, an artificial intelligence tool broke the GDPR and NDPR laws [8]. The ability of artificial intelligence to improve security works against its inclination to violate personal privacy.

Moreover, Ethical hacking AI systems must have exact computer data to identify security flaws in platforms like Ethiack and Equixly. These systems fail to observe privacy rules, so they often find difficulties running. Although the GDPR Article 9 expressly forbids handling health data, NHS audit findings reveal that tested artificial intelligence technologies usually lack default encryption measures for healthcare data. According to [7], a vulnerability scanner using artificial intelligence revealed compromised financial records during a bank audit in 2024, which set off a $4.2 million phishing campaign. These events make clear the insufficient systems between the data needs of artificial intelligence and the privacy needs.

However, although the present limitations under GDPR and the EU AI Act 2024 mostly protect data and transparency in systems, they overlook the exclusive privacy hazards generated by AI in cybersecurity. Articles 5 and 17 of GDPR on data minimisation and right to erasure have poor application in artificial intelligence settings since data intake for precision is still vital for models. Ref. [12] had unencrypted transaction records stored, according to the audit of Kenyan fintech companies using AI tools, even following an operation that broke KBPR's storage policies. The AI Act distinguishes several types of penetration testing tools into "high-risk" or "non-high-risk" categories, therefore allowing companies to avoid doing privacy impact studies [23].

Additionally, reducing hazards has been possible with technical solutions, including homomorphic encryption and differential privacy. Homomorphic encryption allowed artificial intelligence tools to examine encrypted patient data without decryption, therefore lowering exposure risk [27]. Analogous to this, anonymising methods frequently fail in cybersecurity settings: a 2023 study revealed that 78% of "anonymised" network traffic logs could be re-identified using metadata patterns, therefore negating privacy guarantees [21].

In conclusion, most of the privacy concerns created by technology operations fall on underfunded economic sectors. While developing areas utilise tools with limited encryption capability due to financial constraints, European business network audit tools rely mostly on modern encryption standards as their security mechanism. The 2024 [10] poll indicates that whereas North American teams reported such limits only at 34%, African cybersecurity personnel faced access issues to privacy-protecting AI solutions at a rate of 72%. The disparities shown in this mismatch guarantee that underprivileged systems all around constantly suffer privacy violations as well as cyberattacks.

3.4 GDPR and AIA: Limitations in Addressing AI-Specific Challenges

Together with the EU Artificial Intelligence Act (AIA), the General Data Protection Regulation (GDPR) sets fundamental rules for data security and artificial intelligence ethics. The two models show significant flaws in their applications to ethical hacking driven by artificial intelligence since they do not sufficiently manage the technological accompanying ethical issues of autonomous cybersecurity solutions.

However, GDPR aims to protect personal privacy, which shows up in Article 5 data minimising rules and Article 17 right to erasure policies, but does not control the AI-based systematic dangers in ethical hacking operations. During a 2023 penetration test on a German hospital, investigators insecure patient records, therefore violating GDPR Article 9 regulations of health data processing [8]. Data breach investigations revealed that GDPR's focus on post-leak fines does not create protection mechanisms before their occurrence for real-time artificial intelligence systems. Under GDPR, the right to explanation under Article 22 only affects automated judgements made for individual instances, not the organisational security risks produced by AI faults affecting decision-making systems, such as biased vulnerability prioritising [12].

Furthermore, Article 6 of the AIA labels AI technologies applied in critical infrastructure as "high-risk", hence requiring openness and human control. Its standards, meanwhile, lack clarity for uses in cybersecurity. Because they are sold as improvements to human-led processes rather than stand-alone systems, tools like Ethiack and Equixly, which independently run penetration tests, often avoid "high-risk" designation [23]. As shown in a 2024 event whereby an AI tool corrupted firewall rules during a banking sector audit, exposing transactional data, this gap lets vendors avoid thorough audits [7,22].

However, while both systems stress openness, they ignore the technical facts of artificial intelligence decision-making. Though it does not demand explainability approaches (e.g., LIME, SHAP) to demystify AI logic, the AIA mandates that high-risk systems offer "sufficiently detailed" documentation (Article 13). For example, Darktrace's Antigena, which independently stops threats, provides no interpretable logs for its activities, therefore depriving auditors of the means to confirm judgements during post-incident assessments [10]. Comparably, GDPR's transparency criteria centre on data subjects rather than cybersecurity experts, therefore separating operational responsibility from compliance.

Moreover, Cross-border settings clearly show the limits of these systems. An artificial intelligence technology based on EU data unintentionally breached Kenya's Data Protection Act by processing consumer information without regional consent safeguards during a 2024 audit of a multinational e-commerce platform [12]. While Article 3 of GDPR imposes rigorous territorial restrictions, it does not mandate that businesses create worldwide AI tool compliance policies. Outside of its borders, the AIA lacks enforcement powers; so, non-EU suppliers can utilise covert artificial intelligence systems left unmonitored in any member state.

These models fail to adequately address the dual-use problem present in ethical hacking instruments using artificial intelligence algorithms. Since they were first developed to fight phishing attempts, unlike GDPR's emphasis on data protection and the AIA's safety protocols, ChatGPT, along with other generative models, poses a threat of undetectable malware generation since they pose a threat of conducting undetectable malware production. While Article 52 of the AIA requires risk assessment of high-risk systems, it does not provide required measures against systematic exploitation, therefore exposing organisations to AI attack threats.

3.5 Balancing Automation with Human Expertise

By expediting vulnerability discovery and threat response, the incorporation of artificial intelligence into ethical hacking has transformed cybersecurity; nonetheless, its effectiveness depends fundamentally on the complementary function of human knowledge. By automating processes like network mapping and log analysis, AI systems like Pentoma and Equixly show amazing efficiency in lowering detection times by 60% over hand techniques [3]. But often the cost of this efficiency is contextual knowledge.

Additionally, Human knowledge is essential in closing these gaps, especially in ethically and culturally complex settings. Think about the difficulty of protecting mobile money platforms in sub-Saharan Africa, where artificial intelligence tools trained on Western corporate networks missed 34% of SIM-swapping vulnerabilities. By contrast, human-led audits included local transactional behaviours and infrastructural quirks, therefore highlighting dangers that algorithms missed [12]. Likewise, Darktrace's Antigena lacks the sense to assess collateral damage, even if it is successful in autonomously neutralising hazards. Its forceful isolation of a misflagged server upset the telemedicine operations of a hospital in 2023, therefore postponing important patient care until human operators interfered [8]. These illustrations show how human judgment must temper artificial intelligence's operational speed to negotiate ethical and practical trade-offs.

Still, the move toward automation poses the risk of worsening underlying inequalities. A 2024 ISC2 survey shows that, compared to 29% in North America [10], 72% of cybersecurity teams in underdeveloped countries lack access to AI capability. Training data biases, such as giving corporate networks priority over public infrastructure, skew their influence even with the current resources at hand. Artificial intelligence municipal audits, for example, frequently undervaluate vulnerabilities in water treatment plants, which account for 58% of all critical infrastructure breaches in low-income areas [21].

These variations reveal a paradox: even if artificial intelligence democratises access to better risk detection in theory, it reinforces previously existing inequality in reality [19]. Modern judicial systems aggravate these problems by ignoring the necessary human observation. For instance, the EU AI Act 2024 requires "human oversight" for high-risk systems but does not specify procedures for real-time collaboration.

4 Recommendations

The integration of artificial intelligence (AI) into cybersecurity, particularly in ethical hacking, demands urgent regulatory reforms to address gaps in frameworks like the GDPR and EU AI Act. Current systems struggle to balance innovation with accountability, especially as AI tools increasingly grapple with dual-use risks, privacy breaches, and algorithmic bias. Promoting ethical AI integration and reducing systemic risks depend on evidence-based improvements grounded in technical viability and worldwide interoperability.

The solution requires an approach to handle the black-box nature of AI systems because it hinders accountability when using Darktrace's Antigena platforms. Post-incident review auditing becomes impossible because these systems do not provide auditable decision trails according to [28]. The deployment of explainable frameworks such as LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations) through government mandate enables the breakdown of AI decisions. Researchers found that the LIME framework incorporation minimized incorrect interpretations of AI alerts when used in healthcare penetration testing [29]. Meaningful explainability-by-design frameworks as prescribed by LIME and SHAP operatively meet GDPR's "right to explanation" standards while enabling auditors to inspect AI systems while maintaining operational speed.

The EU AI Act's "high-risk" definition must be clarified because Ethiopian tools circumvent regulation through marketing of their systems as human process improvement modules [23]. AI systems used for threat response and penetration tests alongside vulnerability detection must be grouped as a high-risk category, no matter how they are marketed [5]. Asset owners would need to perform adversarial testing during certification, according to the upcoming measure illustrated by the Texas power grid attack of 2024, where an uncertified AI tool weaponization caused a 36-h blackout. Developers can use the CALDERA framework by MITRE to simulate attacks through the framework during their AI tools' development stage [26].

The second foundation of restructuring involves implementing robust privacy-by-design procedures. GDPR's reactive approach to fining breaches was unable to stop the 2023 breach of German hospital patient records through an AI-controlled penetration test, according to [8]. As a parallel solution, we need to combat algorithmic bias so that it stops causing health system inequalities among rural communities [14].

The solution to international complex situations depends on regulatory harmonization. Research by [12] shows that the global e-commerce AI tool violated the Data Protection Act of Kenya, thus demonstrating the necessity for standardized regulations. The implementation of adaptive compliance modules that follow the ISO/IEC 27050 framework enables automatic adaptation of data processing approaches among different jurisdictions. Climate change is a global threat that requires collective national and international solutions. The Wassenaar Arrangement's controls on cyber technologies and collaborative enforcement bodies such as the Global Cybersecurity Alliance would help harmonize regional requirements with standard ethical practices [30].

5 The Way Forward

The collaboration between ethical hacking and artificial intelligence represents a big step forward in cybersecurity as it gives professionals the power to discover system flaws and vulnerabilities, along with suggesting possible attacks and reducing exposure points, and ensuring stronger protection elements. The study demonstrates how autonomous systems created by artificial intelligence systems that increase operational efficiency have the potential to cause system failures at all organizational levels. Systems that favor well-funded educational initiatives over underfunded ones contribute to the systemic promotion of inequality. The dual-use challenge that results from this identification process expanding attack surfaces can make defensive technology a potential weapon.

Moreover, the ethical application of AI in cybersecurity requires a careful strategy to use automation to improve human insight while maintaining crucial ethical judgment, cultural awareness, and context sensitivity. The case studies that are being discussed, which start with the 2023 Deutsche Bank Crisis and end with the 2024 Texas power infrastructure attack, show the serious repercussions that arise when a proper balance is not maintained. In this particular case, artificial intelligence functions beyond accepted ethical bounds, designed to optimize its efficiency. Analysis of the impending repercussions is presented in the paper. The EU AI Act of 2024 offers a higher level of analysis of unique AI-related issues beyond its limited potential.

To operationalize bias-aware data curation in practice, organizations should integrate open-source fairness toolkits directly into their AI pipelines at the data-ingestion stage. For example, IBM AI Fairness 360 offers over seventy metrics for dataset and model bias detection alongside eleven bias-mitigation algorithms (e.g., reweighting, disparate impact remover), enabling teams to identify and correct skew before training. Complementing this, Google's What-If Tool provides an interactive, no-code interface for slicing datasets, probing "what-if" counterfactual scenarios, and visualizing fairness metrics across subpopulations. For production workflows, Microsoft Fairlearn delivers dashboards and constraint-based learning algorithms that optimize models for parity across defined groups, while Aequitas (University of Chicago) supplies group-based audit reports and threshold-independent disparity metrics to surface underrepresented segments. By embedding these toolkits into automated data pipelines running batch audits on incoming records, generating fairness reports, and triggering alerts when imbalance thresholds are exceeded, organizations can ensure their ethical-hacking AI models are trained on representative, equitable datasets, thereby reducing the risk of perpetuating systemic vulnerabilities.

Finally, the direction of ethical hacking depends on defining its current meaning rather than halting the integration of AI technology [31]. Those involved in technology-based adaptive governance need to increase communication between developers, policymakers, and practitioners to collaborate; resources should be allocated at equal levels throughout the entire artificial intelligence development process; ethical considerations and cross-field team collaboration must be given priority; and people must be more committed to transforming AI into a safety network that is accessible to all, given the ongoing complexity of cyber threats. Instead of looking for ways to stop the integration of artificial intelligence, ethical hacking stays true to its original purpose definition, which dictates its direction [32]. More cooperation between developers, policymakers, and practitioners is required by those in charge of making governance decisions based on technological advancements. To ensure fair progress, shared resources from several participating teams and ethical knowledge are required at every stage of artificial intelligence development. Because cyber threats are becoming more complex, our increased commitment should support the advancement of artificial intelligence as a global safety measure.

6 Conclusion and Future Directions

In this review, we first established four principal ethical challenges: algorithmic bias, privacy-preserving tensions, accountability gaps, and the dual-use dilemma, and demonstrated how AI's efficiency gains can nonetheless amplify inequities in under-resourced settings such as rural clinics and small enterprises. Although the GDPR and the EU AI Act lay a foundation for data protection and transparency, they lack the agility to keep up with rapidly evolving autonomous cybersecurity tools.

To address these shortcomings, we advocate a twofold strategy. First, organizations must adopt adaptive governance frameworks that embed continuous monitoring, algorithmic impact assessments, and human-in-the-loop checkpoints whenever models are retrained, transforming policy from a one-off compliance exercise into a living process. Second, security teams should integrate bias-aware data curation at the earliest stages of their AI pipelines. By employing open-source toolkits, IBM AI Fairness 360 for fairness metrics and mitigation algorithms, Google's What-If Tool for interactive scenario testing, Microsoft Fairlearn for performance parity dashboards, and Aequitas for group-based audit reports, practitioners can systematically detect and correct dataset imbalances before deployment. This combination of adaptive governance and rigorous data curation will help ensure that AI-driven ethical hacking tools serve all contexts equitably, rather than perpetuating existing resource divides.

At the same time, credentialed identities and immutable audit logs must be mandatory for autonomous AI agents, measures underscored by cybersecurity experts at the RSA Conference 2025 to prevent unauthorized actions and enable robust forensics. Implementing tiered monitoring systems will then classify tools by risk level, requiring high-risk deployments to undergo explainability audits and maintain human oversight.

Globally, harmonized standards are essential to prevent regulatory arbitrage. The Bletchley Declaration (November 2023) offers a blueprint for shared commitments to responsible AI, while emerging proposals for a global regime complex align AI governance with international law, ensuring hostile AI uses are universally prohibited.

Finally, future research should probe the socio-technical interplay between AI autonomy and human judgment in diverse cultural and infrastructural settings. Key questions include how to tailor adaptive governance to local legal frameworks, how tamper-resistant architectures can mitigate dual-use risks, and how breakthroughs in quantum computing or generative AI will reshape ethical hacking practices. By grounding these inquiries in the themes identified here, scholars can develop empirically supported models that guide policymakers, technologists, and practitioners toward a cybersecurity future that is equitable, accountable, and transparent.

Acknowledgement: Not applicable.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: The authors confirm contribution to the paper as follows: Conceptualization, Hossana Maghiri Sufficient; methodology, Hossana Maghiri Sufficient; validation, Hossana Maghiri Sufficient, Abdulazeez Murtala Mohammed and Bashir Danjuma; formal analysis, Hossana Maghiri Sufficient; investigation, Bashir Danjuma; resources, Hossana Maghiri Sufficient; data curation, Abdulazeez Murtala Mohammed; writing—original draft preparation, Hossana Maghiri Sufficient and Abdulazeez Murtala Mohammed; writing—review and editing, Hossana Maghiri Sufficient, Abdulazeez Murtala Mohammed and Bashir Danjuma; project administration, Abdulazeez Murtala Mohammed. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The authors confirm that the data supporting the findings of this study are available within the article.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

Supplementary Materials: The supplementary material is available online at https://www.techscience.com/doi/10.32604/jcs.2025.066312/s1.

References

1. Saha S, Das A, Kumar A, Biswas D, Saha S. Ethical hacking: redefining security in information system. In: Proceedings of the International Ethical Hacking Conference 2019; 2019 Aug 22–25; Kolkata, India. doi:10.1007/978-981-15-0361-0_16.
2. Sarker IH. Machine learning for intelligent data analysis and automation in cybersecurity: current and future prospects. Ann Data Sci. 2023;10(6):1473–98. doi:10.1007/s40745-022-00444-2.
3. Hu Z, Beuran R, Tan Y. Automated penetration testing using deep reinforcement learning. In: Proceedings of the 2020 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW); 2020 Sep 7–11; Genoa, Italy. doi:10.1109/eurospw51379.2020.00010.
4. Jada I, Mayayise TO. The impact of artificial intelligence on organisational cyber security: an outcome of a systematic literature review. Data Inf Manag. 2024;8(2):100063. doi:10.1016/j.dim.2023.100063.
5. Kaur R, Gabrijelčič D, Klobučar T. Artificial intelligence for cybersecurity: literature review and future research directions. Inf Fusion. 2023;97(6):101804. doi:10.1016/j.inffus.2023.101804.
6. Raza H. Systematic literature review of challenges and AI contributions in penetration testing [master's thesis]. Luleå, Sweden: Luleå University of Technology; 2024.
7. Deloitte CTI. Threat assessment: how threat actors are leveraging artificial intelligence (AI) technology to conduct sophisticated attacks [Internet]. [cited 2024 Nov 22]. Available from: https://www.contentree.com/reports/threat-report-how-threat-actors-are-leveraging-artificial-intelligence-ai-technology-to-conduct-sophisticated-attacks_417160.
8. He Y, Zamani E, Yevseyeva I, Luo C. Artificial intelligence-based ethical hacking for health information systems: simulation study. J Med Internet Res. 2023;25(1):e41748. doi:10.2196/41748.
9. Ueno H. Artificial intelligence as dual-use technology. In: Hatzilygeroudis IK, Tsihrintzis GA, Jain LC, editors. Fusion of machine learning paradigms: theory and applications. Berlin/Heidelberg, Germany: Springer; 2023. p. 7–32. doi:10.1007/978-3-031-22371-6_2.
10. ISC2. The ethical dilemmas of AI in cybersecurity [Internet]. [cited 2024 Nov 18]. Available from: https://www.isc2.org/Insights/2024/01/The-Ethical-Dilemmas-of-AI-in-Cybersecurity.
11. Schellekens P, Skilling D. Three reasons why AI may widen global inequality [Internet]. [cited 2024 Aug 15]. Available from: https://www.cgdev.org/blog/three-reasons-why-ai-may-widen-global-inequality#:~:text=Will%20global%20inequality%20rise%20or,regions%20risk%20being%20left%20behind.
12. Wanyama E. The impact of artificial intelligence on data protection and privacy in Africa: a walk-through rights of a data subject in Africa [Internet]. [cited 2024 Nov 18]. Available from: https://cipesa.org/wp-content/files/briefs/The_Impact_of_Artificial_Intelligence_on_Data_Protection_and_Privacy_-_Brief.pdf.

13. Gupta M, Akiri C, Aryal K, Parker E, Praharaj L. From ChatGPT to ThreatGPT: impact of generative AI in cybersecurity and privacy. IEEE Access. 2023;11:80218–45. doi:10.1109/access.2023.3300381.
14. Joshi R. Ethical challenges and privacy concerns in innovations. In: Abouhawwash M, Rosak-Szyrocka J, Gupta SK, editors. Aspects of quality management in value creating in the Industry 5.0 way. Boca Raton, FL, USA: CRC Press; 2024. p. 202–23. doi:10.1201/9781032677040-12.
15. Riza AZBM, Jennsen L, Anggani P, Rafeen AI, Ruth PNJ, Sookun D, et al. Leveraging machine learning and AI to combat modern cyber threats. 2025. doi:10.20944/preprints202501.0360.v1.
16. Lamche A. NHS software provider fined £3m over data breach [Internet]. [cited 2025 Feb 5]. Available from: https://www.bbc.com/news/articles/cp3yv1zxn94o.
17. Bengio Y, Hinton G, Yao A, Song D, Abbeel P, Darrell T, et al. Managing extreme AI risks amid rapid progress. Science. 2024;384(6698):842–5. doi:10.1126/science.adn0117.
18. Jones D. CISA, FBI confirm critical infrastructure intrusions by China-linked hackers [Internet]. [cited 2025 Feb 5]. Available from: https://www.utilitydive.com/news/cisa-fbi-critical-infrastructure-china-hacker/706979/.
19. Al-Sinani HS, Mitchell CJ. AI-enhanced ethical hacking: a Linux-focused experiment. arXiv:2410.05105v1. 2024.
20. Kaushik K, Khan A, Kumari A, Sharma I, Dubey R. Ethical considerations in AI-based cybersecurity. In: Kaushik K, Sharma I, editors. Next-generation cybersecurity: AI, ML, and blockchain. Singapore: Springer Nature Singapore; 2024. p. 437–70. doi:10.1007/978-981-97-1249-6_19.
21. López González A, Moreno M, Moreno Román AC, Hadfeg Fernández Y, Cepero Pérez N. Ethics in artificial intelligence: an approach to cybersecurity. Inteligencia Artific. 2024;27(73):38–54. doi:10.4114/intartif.vol27iss73pp38-54.
22. Agarwal M. Unleashing the power: exploring ethical hacking and artificial intelligence for stronger national security. Int J Curr Sci. 2023;13(3):556–66.
23. Sambamurthy PK. The integration of artificial intelligence in ethical hacking: revolutionizing cybersecurity predictive analytics. Int J Adv Res Emerg Trends. 2024;1(2):199–211.
24. Raman R, Calyam P, Achuthan K. ChatGPT or bard: who is a better certified ethical hacker? Comput Secur. 2024;140(6):103804. doi:10.1016/j.cose.2024.103804.
25. He Y, Luo C, Suxo Camacho R, Wang K, Zhang H. AI-based security attack pathway for medical diagnosis systems (CMDS). In: Proceedings of the 2020 Computing in Cardiology; 2020 Sep 13–16; Rimini, Italy. doi:10.22489/cinc.2020.439.
26. Omar MS, Zolkipli MF. Fundamental study of hacking attacks protection using artificial intelligence (AI). Int J Adv Eng Manag. 2023;5(2):813–21.
27. Williamson SM, Prybutok V. Balancing privacy and progress: a review of privacy challenges, systemic oversight, and patient perceptions in AI-driven healthcare. Appl Sci. 2024;14(2):675. doi:10.3390/app14020675.
28. Casper S, Ezell C, Siegmann C, Kolt N, Curtis TL, Bucknall B, et al. Black-box access is insufficient for rigorous AI audits. In: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency; 2024 Jun 3–6; Rio de Janeiro, Brazil. doi:10.1145/3630106.3659037.
29. Hassan SU, Abdulkadir SJ, Zahid MSM, Al-Selwi SM. Local interpretable model-agnostic explanation approach for medical imaging analysis: a systematic literature review. Comput Biol Med. 2025;185(1):109569. doi:10.1016/j.compbiomed.2024.109569.
30. Korzak E. Export controls: the Wassenaar experience and its lessons for international regulation of cyber tools. In: Tikk E, Kerttunen M, editors. Routledge handbook of international cybersecurity. Abingdon, UK: Talylor Francis Group; 2020. p. 297–311. doi:10.4324/9781351038904-31.
31. Saraswathi VR, Ahmed IS, Reddy SM, Akshay S, Reddy VM, Reddy SM. Automation of recon process for ethical hackers. In: Proceedings of the 2022 International Conference for Advancement in Technology (ICONAT); 2022 Jan 21–22; Goa, India. doi:10.1109/iconat53423.2022.9726077.
32. Mohamed I, Hefny HA, Darwish NR. Enhancing cybersecurity defenses: a multicriteria decision-making approach to MITRE ATT&CK mitigation strategy. arXiv:2407.19222v1. 2024.