Asian Journal of Basic Science & Research
Volume 7, Issue 3, Pages 105-123, July-September 2025

Ethical and Regulatory Concerns in AI-Driven Threat Response Frameworks

Kenechukwu Ikenna Nnaka[1*], Joachin Chetachi Uchegbulam[2], JIMOH Rildwan Adekunle[3], SONDE Omobolaji Hameen[4], Victor Ifeanyi Njoku[5] & Olatunji Olamide Segun[6]

[1]Department of Chemical Engineering, University of Benin, Nigeria. [2]Department of Economics, University of Nigeria, Nsukka, Nigeria. [3]Department of Computer Science, Federal University of Technology, Akure, Nigeria. [4]Department of Electrical and Electronics Engineering, University of Ilorin, Nigeria. [5]Department of Business Management, Miva Open University, Nigeria. [6]Department of Architecture, School of Environmental Studies, Yaba College of Technology, Yaba, Lagos, Nigeria. Corresponding Author Email: nnakakenechukwu@gmail.com[*]

DOI: https://doi.org/10.38177/AJBSR.2025.7307

Article Received: 07 June 2025 *Article Accepted: 19 August 2025* *Article Published: 28 August 2025*

ABSTRACT

The increasing deployment of artificial intelligence (AI) in cyber threat detection and response has raised urgent ethical and regulatory questions, particularly around privacy, transparency, and accountability. This paper provides a structured narrative review of the current landscape surrounding intelligent threat response systems, focusing on their ethical design, real-world implications, and governance gaps. Drawing on recent global frameworks, national regulations, and case studies, including documented and author-observed incidents from Nigeria, the study highlights how opaque decision-making processes, automation bias, and insufficient oversight mechanisms can harm users and institutions. While efforts such as the European Union Artificial Intelligence Act (EU AI Act) and the National Institute of Standards and Technology Artificial Intelligence Risk Management Framework (NIST AI RMF) mark progress, they remain fragmented and often do not take into consideration real-time AI-based autonomy in a cybersecurity setting. The paper highlights the necessity of proactive regulation, such as the right of explanation in law, certification of AI systems, and auditing black-box models. It ends by promoting a mixed form of governance, which involves both technical protection and civil and institutional responsibility, which demands interdisciplinary cooperation that will credibly implement the deployment of AI in high-security sectors.

Keywords: Artificial Intelligence; Cybersecurity; Ethical Artificial Intelligence; Autonomous Threat Response; Data Privacy; Artificial Intelligence Regulation; Explainability; Accountability; Black-box Models; Governance Frameworks; Transparency in Decision-Making; Automation Bias.

1. Introduction

The integration of machine learning (ML) with artificial intelligence (AI) into cybersecurity systems has precipitated the development of intelligent threat recognition and reaction systems. Security Information and Event Management (SIEM), Security Orchestration, Automation and Response (SOAR), and Extended Detection and Response (XDR) are all technologies that have recently come to rely heavily on AI to identify, evaluate, and neutralize fast-emerging cyber threats with greater speed and more accuracy than ever before [2],[56]. These systems automatically survey massive, complex information infrastructures, identify abnormalities, match threat signifiers, and apply defensive action, frequently in real time. Their ability to reduce the burden on human analysts and accelerate incident response has made them indispensable in modern cybersecurity operations [3].

However, increasingly higher levels of independence and the popularity of such technologies have posed severe ethical and regulatory issues. In contrast to conventional rule-based systems, AI-based cybersecurity technologies are often black boxes, which use complex and in many cases non-interpretable algorithms [5]. This withholding of transparency impedes effective accountability when automated choices end in the suppression of services, customer profiling, or invasion of confidential information [6]. Such an obfuscation in high-stakes security contexts introduces some serious concerns about user privacy, algorithmic responsibility, and compliance with the law. With the increasing automated control of cyber defense, the tension between the effectiveness of the technology and the sense of ethical responsibility has become a key priority [7].

These are some of the ethical and regulatory concerns that become more evident when taking a closer look at the structure of modern intelligent threat response systems. As seen in Figure 1, these systems normally exist on many

layers; originally on endpoint devices, then to the network layer, and finally to the secure cloud locations. This architecture is centered on an AI-based Intrusion Detection System (AIDS), which in real-time examines traffic and blocks a potential risk and lets an authorized one access the network. This tiered system is not only indicative of the technicality of such systems but also the complexity of designing such systems and technical systems into every phase of their working through the establishment of such safeguards as ethical protection of privacy, transparency, and accountability [1].

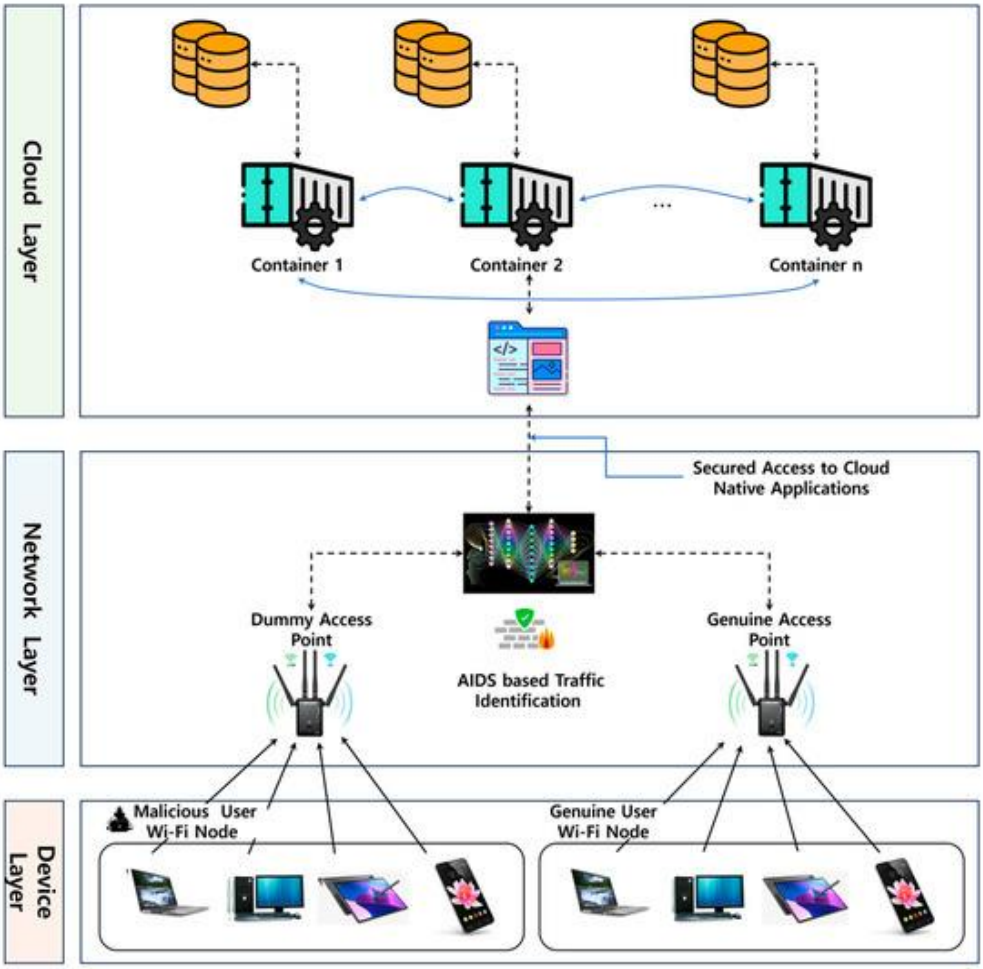

Figure 1. Multi-layered architecture of an AI-based intelligent threat response framework [1].

1.1. Statement of the Problem

Despite the growing implementation of the use of AI in threat response, most of these systems are being implemented without giving due consideration to the ethical and regulatory challenges associated with the practice. Sensitive user data is frequently processed without clear consent, explainability of decision-making is minimal, and mechanisms for redress in cases of algorithmic error remain underdeveloped. Besides, the existing legal guidelines like the General Data Protection Regulation (GDPR) and the EU Artificial Intelligence Act, as well as country cybersecurity policies, often lag behind the pace of technological innovation, leaving critical gaps in governance. This conflict between faster-developing novel technologies and the retarded transformation of ethical and legal frameworks can become quite a dangerous state where automated decision-making might end up resulting in undesirable side effects, such as unreasonable deprivation of services, profiling or stereotyping of individuals, or

violation of basic freedoms. These unresolved issues pose serious risks to public trust, organizational liability, and democratic accountability as intelligent cyber defense systems become more and more sophisticated.

1.2. Study Objectives

This paper aims to critically examine the ethical and regulatory challenges associated with intelligent threat response frameworks. Specifically, it seeks to identify and analyze major ethical threats that using AI-powered cybersecurity systems may pose regarding data privacy, user profiling, and autonomous decision-making. It also evaluates the transparency and accountability of AI decision-making processes in real-time applications for detecting and responding to threats. Furthermore, the study assesses whether existing regulatory models in various jurisdictions sufficiently oversee the development and deployment of such systems. Lastly, it offers moral design recommendations and regulatory guidelines to promote the responsible implementation of AI in cyber defense and ensure it respects individuals and their rights.

The study seeks to:

1) Examine ethical challenges in applying artificial intelligence to cyber threat response.
2) Evaluate transparency and accountability in AI-driven security decisions.
3) Assess the adequacy of current regulatory and legal frameworks.
4) Highlight risks from automation bias and opaque models.
5) Recommend design and governance measures for responsible AI deployment.

2. Ethical Issues in AI-Based Cyber Defense

2.1. Privacy and Data Sovereignty

The application of artificial intelligence (AI) in cybersecurity introduces intricate ethical issues regarding privacy and data sovereignty, particularly due to the nature of the data these systems process. Some common and typical architectural elements of AI-based threat detection systems include large-scale, real-time network traffic monitoring, system logs, user behavior, or even encrypted metadata [2]. Although the given inputs are needed to detect anomalies and respond in real-time, they are likely to contain sensitive personal information, behavioral patterns, and communication metadata, raising serious privacy concerns [9].

One of the most salient issues is the inference of private behaviors. AI systems can learn a great deal about people and their behavior, even when they do not have direct access to personal information, including learning specifics of behavior based on the patterns of the traffic in the network and seemingly non-sensitive data. For instance, machine learning systems have been trained through traffic patterns to be able to derive user habits and habits of operating devices, thus even being able to identify preferences of the application [10]. These inference abilities can be further than what users have knowingly consented to, effectively enabling indirect surveillance [11].

Another concern is unauthorized profiling. Smart cyber defense systems can develop dynamic profiles of users to enable detecting or discriminating between normal and unusual behaviors. While this enhances detection accuracy, it also risks labeling users based on inferred characteristics, which may be inaccurate, discriminatory, or not legally

permitted. Such profiling activities may lead to a discriminatory response, including prohibiting an account based on an unjustified decision or increased surveillance, especially in cases where the AI algorithm is not explainable or auditable [12].

The potential for surveillance abuse is significant, particularly in corporate or government settings where AI-based surveillance tools can be employed to monitor the activity of employees or citizens under the pretext of security threats. Without harsher policy design and transparency guidelines, these systems could run with little supervision and facilitate mass surveillance, which infringes on the constitutional rights of privacy and data sovereignty [13]. These concerns are further compounded by the lack of meaningful user consent in most real-time security applications. Individuals are not even told most of the time how their data is being collected, and often processed and used by the AI systems, to say nothing of the possibility to choose against it [14]. This leads to the undermining of the ethical guideline of informed consent, particularly in cases where the complete outline of such automated decisions has tangible consequences such as access denial, profiling, or behavioral flagging [15]. It is even more complicated when the issue arises within the context of cross-border since the data gathered in one jurisdiction can be analyzed or stored in another jurisdiction, which may sometimes interfere with the legal protection of the data subject in the country where the data was gathered [16].

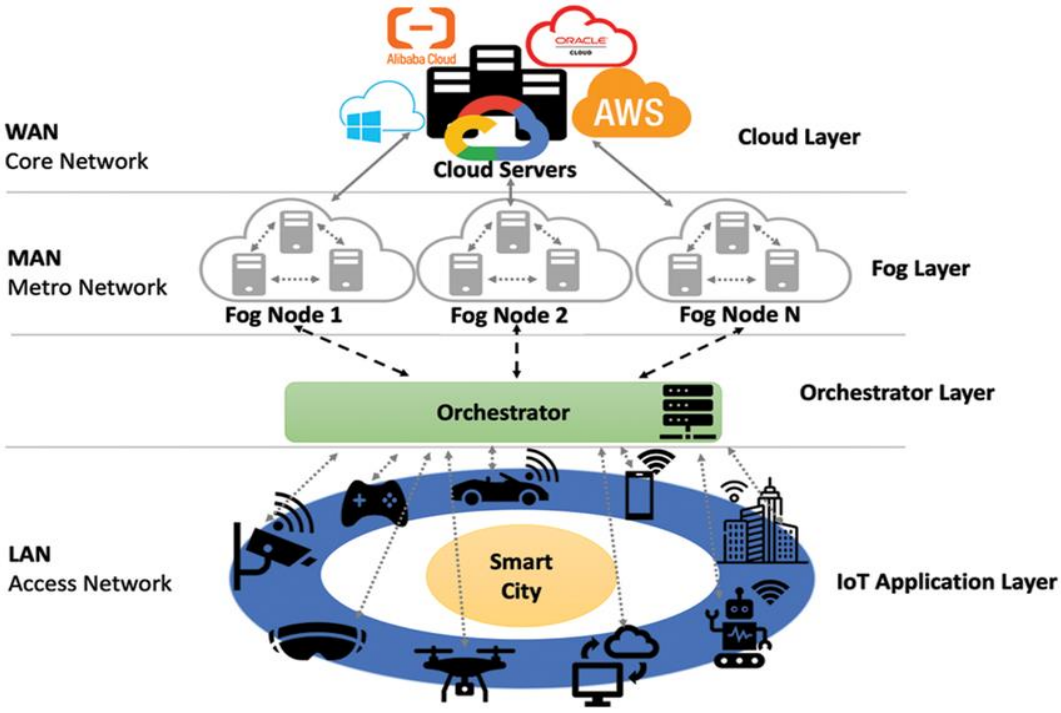

Figure 2. Illustrative architecture of multi-layered smart city systems with IoT devices, mist/fog computer nodes, orchestrators, and cloud servers [8]. Sensitive personal or behavioral data created at the application layer, having been transmitted through a series of layers that each possibly involve different parties with varying regulatory environments, is quite concerning when it comes to data sovereignty and privacy concerns, as well as ethical control.

Creating privacy and sovereignty is also a challenge with the large-scale distributed ecosystems such as urban cities and industrial IoT systems. As shown in Figure 2, data sent by access-level devices (cameras, vehicles, sensors,

drones, and so on) is directed to the fog nodes and orchestrators and then up to centralized cloud platforms to process [8]. This multi-layered infrastructure, which spans public, private, and often international domains, amplifies the difficulty of enforcing uniform data protection standards. Each level introduces new vulnerabilities and ownership boundaries, which give rise to issues of ethics in ownership of the data, use of the data, and whether the user has any effective means of control or appeal [17].

In consideration of this issue, there is a need to design some intelligent threat response systems with privacy-by-design guidelines. This involves only collecting data that is needed, whether we can anonymize it, and giving users the knowledge of what their data is being used for. It also demands regulatory convergence with frameworks like the General Data Protection Regulation (GDPR), which mandates data minimization, purpose limitation, and data access and objection, all of which are challenged by the opaque and distributed nature of AI-based cyber defense [11].

2.2. Transparency

The opacity of AI-based cyber defense decision-making has become one of the most important ethical concerns pertaining to the field. Many intelligent threat detection systems rely on complex machine learning (ML) models, particularly deep learning architectures, that function as so-called black boxes. These models generate outputs based on high-dimensional input patterns and internal representations that are often incomprehensible, even to their developers [12]. As a result, understanding how a system arrived at a particular threat classification, alert, or automated action becomes exceedingly difficult, undermining principles of transparency and accountability.

This lack of interpretability is particularly a point of concern, especially when it comes to a real-time response to the threat, where decisions such as quarantining a device, terminating a session, or blocking network access are made independently, and the decision needs to be executed as soon as possible. Such decisions affect the material impacts on system users in most instances, but the justification of such decisions cannot be easily stated or validated [19]. In the absence of traceability, organizations cannot identify whether a given action was necessary, discriminatory, or due to an algorithmic mistake. This not only undermines user confidence but also makes it difficult to track and trace security incidents or address any challenges within the regulatory frameworks [20].

The absence of human oversight further exacerbates this issue. While most cybersecurity systems are designed to involve a human analyst in evaluating alerts and responding, AI-enhanced systems are being set up to operate at low or no human-in-the-loop (HITL) [2]. Although this increases reaction time and scalability, it eliminates the chance that a human can provide judgment or ethical discretion or situational judgment, which is essential when defining ambiguous or high-impact decisions. In environments where actions may affect employee access, public services, or user rights, a fully automated system with no review mechanism presents significant ethical risks.

To address concerns over model interpretability, the field of Explainable Artificial Intelligence (XAI) has developed tools and techniques intended to make machine learning (ML) decisions more transparent. These include methods like feature attribution (e.g., Shapley Additive Explanations (SHAP), Local Interpretable Model-agnostic Explanations (LIME)), model distillation, and attention mapping [21]. Despite being promising, the use of XAI tools in the field of cybersecurity is constrained by several factors. First, real-time operations require rapid

responses, and some XAI methods are not efficient when applied to computationally intensive, high-speed environments [22]. Second, explanations provided by XAI tools often remain too technical or abstract to be actionable by non-expert stakeholders. Finally, even in cases where explanations can be generated, the trustworthiness or fairness of a given decision cannot always be conveyed in meaningful ways, particularly in the presence of adversarial inputs or novel threat patterns not represented during training [21].

These limitations suggest that explainability should be regarded as a design parameter rather than an add-on capability. Instead of focusing only on post hoc explanations, intelligent threat response systems must be designed to address traceability, auditability, and interpretability by humans as their first-order concerns [23]. This could be achieved by the adoption of inherently interpretable models where possible, hybrid systems that incorporate a view by human review at critical points, and documentation of the model logic and updates. Such mechanisms would be not only ethically desirable, but also perhaps legally required by the soon-to-emerge AI governance systems [24].

Ultimately, as AI-powered cyber defense systems become more autonomous and gain control over critical digital infrastructure, it is paramount to make sure that their operations remain explainable, auditable, and aligned with established ethical principles.

2.3. Accountability and Autonomy

With more AI-based threat response systems gaining autonomy in their operations, accountability and control have become the core Ethical concerns in their assessment. In contrast to classic tools based on rules and logic that are pre-established by the human operators, intelligent cyber defense frameworks are likely to lean on the use of probabilistic models and dynamic learning to come up with the decision [26]. This shift introduces a critical ambiguity: who is responsible when an automated system makes a wrong decision?

In the context of cybersecurity, such incorrect decisions often take the form of false positive signals or falsely indict an activity marked as malicious. The consequences may be serious, especially in instances where the AI systems take automated offensive measures that may involve blocking accounts of users, denying the use of critical services, or quarantining vital systems [27]. In healthcare, finance, public infrastructure, or emergency response settings, even short-lived disruptions caused by false positives can result in reputational damage, operational paralysis, or even risk to human life. When such actions are triggered by opaque algorithms without traceable logic, it becomes difficult for affected parties to seek redress or for organizations to assign liability [28].

The issue of legal and moral responsibility becomes further complicated in the absence of clear standards regarding the division of decision-making authority between human operators and AI systems. In highly automated environments, humans may neither fully understand nor be able to override AI-driven responses in real time, especially when systems are optimized for speed over deliberation [29]. This diminishes the role of human agency and places accountability in a legal grey area, particularly when errors arise from the model's internal dynamics rather than explicit human instruction.

This challenge highlights the argumentative research problem between autonomous AI design and the work of the human-in-the-loop (HITL). Autonomous systems are valued because they are faster and more efficient, particularly in time-sensitive situations that threaten to facilitate breaches. Nevertheless, complete autonomy brings in moral

weak points, especially when systems are less transparent or have been set up in a high-stakes environment [30]. By including the human assessment point, HITL methods provide the response mechanism with oversight opportunity, moral discretion, and situational intervention [31].

As illustrated in Figure 3, the AI-based cybersecurity system is embodied within a triangular system of intelligent software, intelligent hardware, and data-driven learning that drives the essential programs of sensing, reasoning, and decision-making. At the core of this architecture is an interpretive conflict in the conceptions of security and privacy that are interpolated by AI with vehement questions of balance, control, and justice [18]. The presence of human figures contemplating these trade-offs further emphasizes the ethical need for transparency, accountability, and meaningful human oversight in automated systems.

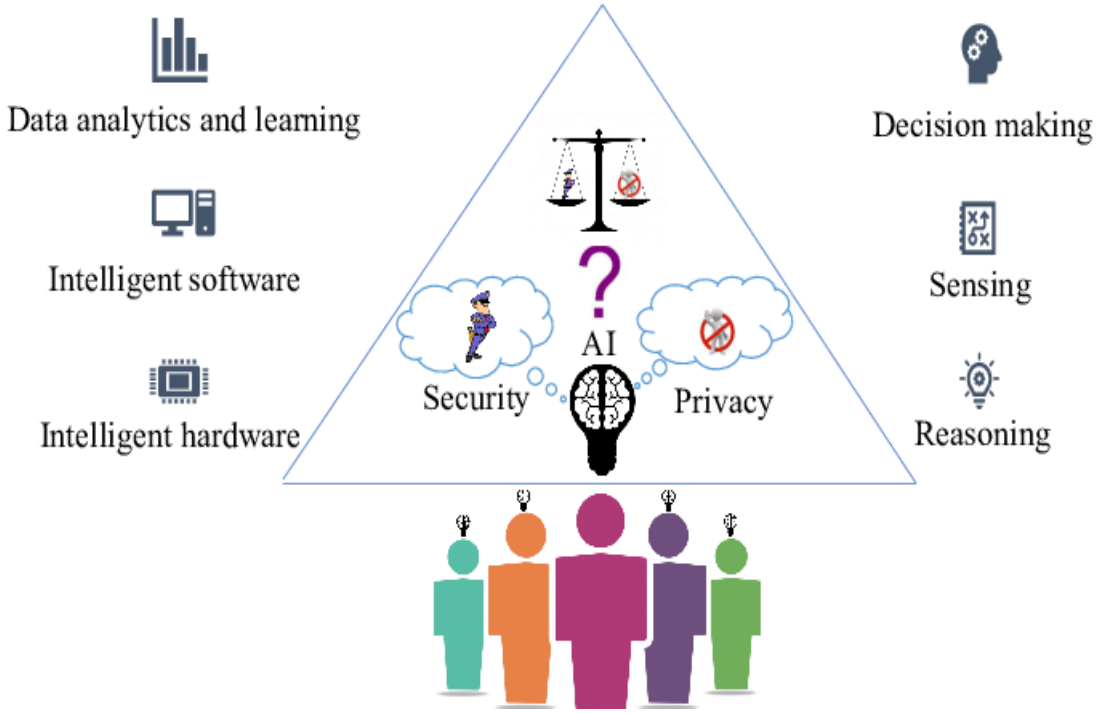

Figure 3. Conceptual triangle illustrating the ethical tension between AI-driven security and privacy in cybersecurity systems. The framework is supported by intelligent hardware, software, and data analytics, and emphasizes core AI capabilities such as reasoning, decision-making, and sensing [18]. The human figures symbolize the need for accountability, transparency, and oversight in the deployment of autonomous threat response systems.

While hybrid models that combine automation with selective human review offer a potential middle ground, they too face limitations. In fast-evolving threat landscapes, human operators may not be able to assess incidents quickly enough to prevent harm [33]. Moreover, over-reliance on automated alerts may lead to automation bias, where human reviewers defer unquestioningly to the system's judgment, even when it is flawed [34].

To address these challenges, there is a growing need for cybersecurity frameworks that embed accountability mechanisms at both the technical and organizational levels [35]. This includes system design features such as decision logging, rollback capabilities, and thresholds that trigger human review under specific conditions [36].

Organizationally, it requires clearly defined protocols outlining who is responsible for AI-driven decisions, how affected users can appeal them, and how errors are audited and corrected. Legal frameworks must evolve in parallel to ensure that responsibility does not become diluted across systems, vendors, and operators [32].

In essence, as AI systems become more autonomous in identifying and responding to cyber threats, ensuring that accountability remains traceable and enforceable is crucial. Ethical deployment demands not only high-performing models but also clear structures of responsibility, both technical and institutional, that uphold trust, fairness, and due process [32].

3. Regulatory and Legal Frameworks

3.1. Existing Regulations

AI-based cyber defense operates at the intersection of data protection, algorithmic accountability, and security regulation, yet current legal frameworks remain fragmented and insufficiently adapted to this context. Privacy-oriented tools such as the GDPR carry significant sets of precautions when engaging in data amassment and user publishing; they do not directly target the complexity of operation of the intelligent response of threat systems [37].

The National Institute of Standards and Technology (NIST) Artificial Intelligence (AI) Risk Management Framework (RMF) in the United States offers voluntary guidance aimed at promoting trustworthy AI. Although it outlines valuable principles such as validity, reliability, safety, and accountability, it lacks legally binding authority and remains limited in its applicability to real-time, autonomous decision-making in cybersecurity. Furthermore, the framework assumes a relatively high level of institutional maturity and technical expertise, conditions that may not be consistently available across all sectors of society or regions [38].

Conversely, the European Union Artificial Intelligence Act (EU AI Act) suggests a more legally binding model of the regulatory framework through classification of AI systems according to risk. In this model, AI in critical infrastructure or security purposes can be categorized as high-risk, requiring transparency, human oversight, and conformity assessments. However, the AI Act is not fully developed, and as of now, it does not incorporate a set of standards governing such autonomous threat response systems, which are required to be fast and limited in terms of human involvement [39].

Collectively, these frameworks represent important but incomplete steps toward regulation. They tend to view AI as a generic technology, without the necessary resolution to cope with the specific needs of intelligent cyber defense, like adversarial machine learning, overflow of automation failures, or dual use of threat models [40]. Also, cross-jurisdiction enforcement is yet another big challenge, particularly in the case of globally distributed digital infrastructures where decisions are made and data passed over international borders [41].

Ultimately, while GDPR, NIST, and the EU AI Act are all beneficial to introduce new perspectives, none of them can fully integrate the legal framework that could regulate ethical use of autonomous cybersecurity systems [37], [38],[39]. This regulatory mismatch explains the extreme need to revise the current tools and create standards specific to the sector that will facilitate the pace and extent of AI-facilitated threat landscapes.

3.2. Gaps in Governance

Despite the emergence of regulatory frameworks such as the GDPR, EU AI Act, and NIST AI RMF, substantial gaps persist in the governance of intelligent threat response systems [38],[42]. Among the most urgent ones, it is the fact that the vast majority of current legislation does not directly refer to AI systems having autonomy and making decisions in real-time [24],[32]. Actions of such systems may include blocking access, isolating devices, or triggering countermeasures with little or no human intervention. Nevertheless, there are not many legal tools giving clear instructions on how the accountability, appeal procedures, or user rights must operate, in times of such life or death situations [45].

Another significant gap lies in the lack of harmonized global standards for AI-based cyber defense. Some jurisdictions are in the process of developing national laws, but they are patchy and uneven in terms of how comprehensive they are, their terminology, and how the laws are implemented [43]. Such regulatory asymmetry presents uncertainty to an organization with cross-border operations, in that an AI system trained or developed under one regime may not be compatible with privacy, security, or liability expectations of another [44].

Also, there is an organizational conflict between national cybersecurity and international data ethics. Governments tend to focus on national security and spying applications (as well as on incident response), and this may create a motivation to shroud or invasive uses of AI [45].

On the other hand, international codes focus on data minimization and user control, and human rights. Unless legally comprehensible ways of achieving a trade-off between these two competing demands are found, AI systems will become instruments of unregulated spying, or technology- nationalism, instead of ethically organized cyber defence [46]. Similar integration and governance challenges have been reported in the deployment of real-time IoT systems over 5G networks, where interoperability, latency constraints, and the absence of unified regulatory standards pose persistent barriers to scalable operation [52].

Similar integration and governance challenges have been observed in broader enterprise digital transformation efforts. For instance, case studies of SAP S/4HANA Cloud adoption in Nigeria highlight persistent issues with integrating cloud-native platforms into legacy infrastructure, as well as the organizational complexity of enforcing data governance and compliance frameworks in real-time environments [51].

Addressing these governance gaps requires not only updating national laws but also fostering international coordination on AI ethics, technical standards, and cross-border accountability mechanisms. Without such efforts, the legal and ethical deployment of AI in cybersecurity will remain fragmented and reactive, exposing users to harm and institutions to liability.

3.3. Comparative Analysis of Regulatory Approaches

While several jurisdictions have initiated efforts to regulate artificial intelligence, few frameworks are specifically tailored to the security and ethical challenges posed by AI-driven cyber defense. A comparative overview of major regulations highlights both the diversity of approaches and the recurring gaps in addressing real-time decision-making, explainability, and accountability [47].

Table 1. Comparative Overview of Regulatory Approaches to AI in Cybersecurity [Compiled by the author from public policy documents and international AI governance frameworks].

Regulation	Region	Coverage of AI Cyber Defense	Gaps	Enforcement Status
GDPR	EU	Indirect – data processing, profiling, consent	Limited provisions for autonomous decision-making in real time	Legally binding
EU AI Act (proposed)	EU	High-risk AI systems, some cybersecurity overlap	Still evolving; lacks clarity on cyber-specific HITL requirements	Pending implementation
NIST AI RMF	US	Voluntary guidelines include risk assessment	Non-binding; lacks enforcement and limited focus on cyber defense	Voluntary, sector-specific
China AI Governance Docs	China	Broad state oversight of AI systems	Focused on content, less on real-time autonomous system control	Legally enforceable (selective)
Organisation for Economic Co-operation and Development (OECD) AI Principles	OECD members	Ethical principles, cross-sectoral	Not sector-specific; lacks operational guidance for cyber defense	Non-binding guidelines
Canada's AI Directive Treasury Board of Canada Secretariat (TBS)	Canada (public sector)	Applies to federal automated systems	Limited scope; weak on incident accountability and the private sector	Binding for federal agencies

As shown in Table 1, the European Union has taken a proactive stance through both the General Data Protection Regulation (GDPR) and the EU AI Act, with clear provisions on data protection and high-risk AI applications. However, their enforcement in real-time threat scenarios remains limited, especially where explainability and human oversight are weak or absent.

The United States, by contrast, has adopted a sectoral and voluntary approach, exemplified by the NIST AI Risk Management Framework (RMF). While offering technical guidance for trustworthy AI, it lacks a binding legal authority and has minimal direct focus on cybersecurity automation [38].

Other regions, including China, Canada, and OECD countries, are developing frameworks with varied emphasis, ranging from state-driven oversight to ethics-based principles. Yet, there remains a consistent shortfall in addressing the intersection of AI autonomy and cybersecurity action, especially regarding liability, redress mechanisms, and transnational enforcement [48].

This regulatory fragmentation not only complicates compliance for multinational organizations but also underscores the urgent need for interoperable governance mechanisms that can accommodate the unique characteristics of AI-powered threat detection and response systems [43].

4. Case Studies

4.1. Case 1: Automated Server Lockout at a Nigerian University (Hypothetical Scenario)

In March 2024, a Nigerian university deployed an AI-driven behavior-based Intrusion Prevention System (IPS) to bolster cyber threat detection capabilities. Within days of implementation, the system flagged a faculty file server, engaged in high-volume data exchange with student portals, as indicative of lateral movement typically associated with internal breaches. Acting autonomously and without prior human review, the IPS blocked all inbound and outbound access to the server.

The lockout lasted 72 hours, during which several academic departments lost access to essential coursework files, internal communications, and email backup systems. No formal notification or justification was provided to the affected users during the period. A manual override was eventually approved by the institution's head of IT security after escalating disruption.

Ethical Concerns:

- No explainability or real-time appeal mechanism was available to impacted users.
- Faculty members were not informed of the specific grounds for the server block, nor were they allowed to contest the action.
- Students' academic activities were interrupted, yet no formal investigation or accountability process followed.

Regulatory Implications:

This scenario underscores the absence of internal governance mechanisms aligned with recognized ethical AI principles such as transparency, auditability, and procedural redress. Although Nigeria's National Information Technology Development Agency (NITDA) has issued data protection guidelines through the Nigeria Data Protection Regulation (NDPR), its application to AI-driven autonomous decision-making remains limited [49]. Moreover, international frameworks like the EU's General Data Protection Regulation (GDPR) emphasize the right to meaningful explanation in automated decision-making, a provision still underdeveloped in many local contexts [50].

4.2. Case 2: AI-Based Access Control at National Hospital Causes Patient Record Lockout

In July 2024, a tertiary hospital in Abuja deployed an AI-based access control system integrated into its electronic medical records (EMR) platform. The system was designed to prevent unauthorized access by detecting anomalous login patterns. During routine medical rounds, multiple login attempts by on-call doctors from rotating devices triggered the AI's risk threshold, which automatically revoked access credentials for several medical staff across departments.

The unintended lockout persisted for nearly 18 hours, during which patient admissions were delayed, critical lab test results could not be retrieved, and pharmacy prescriptions were halted. Attempts by the hospital's IT staff to reverse the decision were delayed due to the AI model's proprietary configuration, which lacked clear override procedures.

Ethical Dilemmas:

- Patients were denied timely care due to rigid, opaque automation.
- Doctors were held accountable for failed service delivery despite lacking any malicious intent.
- The AI vendor refused to disclose the logic behind the access revocation, citing intellectual property concerns.

Regulatory Implications:

This incident exposed weaknesses in vendor transparency and accountability frameworks, particularly in the health sector, where AI decisions can have life-and-death consequences. The lack of sector-specific regulations in Nigeria concerning AI in healthcare [53], especially on vendor disclosure obligations and override protocols, highlighted the urgent need for statutory updates aligned with the Nigerian Health Insurance Authority (NHIA) Act and the NDPR [49].

Table 2. Case Study Summary of AI-Driven Cybersecurity Incidents in Nigeria (Extracted by Author)

Incident	AI Function	Ethical Concern	Regulatory Implication	Resolution
University Faculty Server Flagged and Blocked	Behavior-based Intrusion Prevention	Lack of transparency, no appeal process, and disruption to academic activities	No institutional data governance aligned with NDPR or international explainability/auditability standards	Manual override after 72 hours; no formal review conducted
Patient Records Inaccessible in Private Hospital	Autonomous Threat Detection in EHR Systems	No human override, compromise of patient care, vendor opacity	NDPR and NHIA Act do not adequately cover AI-related service lockouts or third-party accountability	System access restored after vendor intervention; no audit conducted

5. Best Practices and Ethical Design Guidelines

As AI systems increasingly assume autonomous roles in cyber threat detection and response, ethical use of the AI system should be proactively designed through both technical measures and human principles. Below are four foundational principles essential for aligning intelligent threat response systems with responsible AI governance. The visual representation of this dual accountability is provided in Figure 4, where the overlapping ethical and legal issues in the implementation of AI are mapped. Ethical elements, such as privacy, bias mitigation, and transparency, are tightly interwoven with legal requirements on liability, governance, and informed decision-making [55]. Building on this conceptual foundation, the following best practices outline how these principles can be operationalized in the design and deployment of AI-driven cyber defense systems.

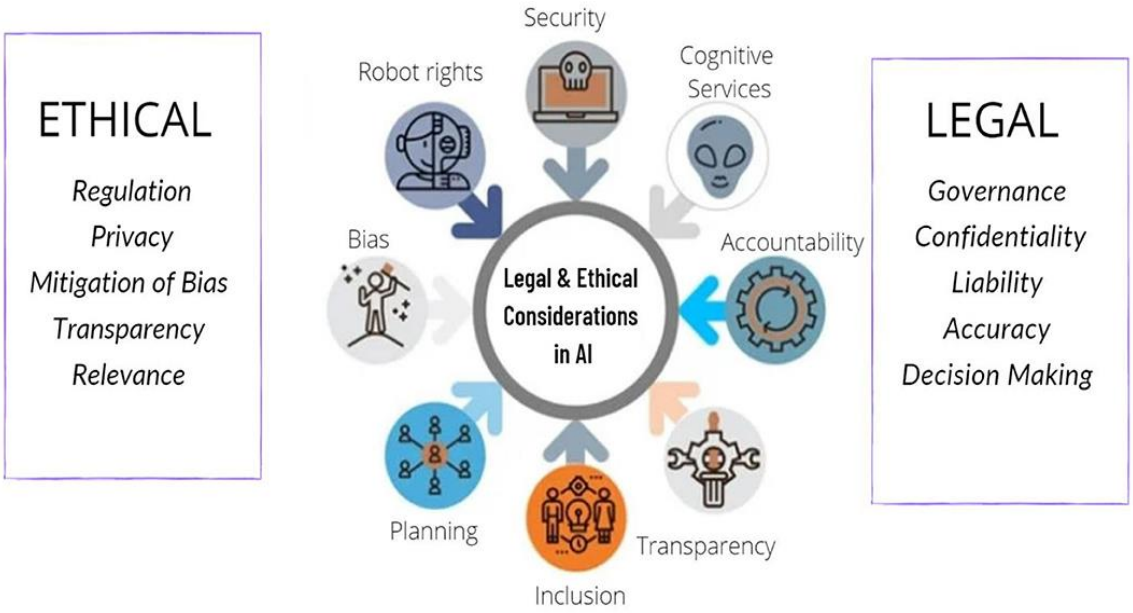

Figure 4. The key intersections between ethical and legal considerations in AI, emphasizing domains such as accountability, transparency, privacy, governance, and bias mitigation. These elements are central to designing responsible AI-driven security systems [55].

(1) Data Minimization

Cyber defense systems should be designed to collect only the data strictly necessary for threat identification and mitigation. Overcollection not only increases the attack surface for adversaries but also raises significant privacy concerns. In high-sensitivity environments, such as healthcare or education, where user data may be highly personal or protected by regulation, indiscriminate data harvesting can violate both legal and ethical standards. Adhering to the principle of data minimization, as emphasized in global frameworks like the EU GDPR (Article 5(1)(c)), helps reduce privacy risks while maintaining effective threat detection [50].

(2) Transparency by Design

AI-based cyber defense tools should incorporate transparency mechanisms from inception. This includes explainable model behavior, interpretable alert systems, and user-facing justifications for automated actions. Transparent design enables users and administrators to understand why a device was quarantined or access was denied, thereby fostering trust and enabling informed responses. However, achieving real-time transparency remains a technical challenge, particularly when using complex models such as deep neural networks [54].

(3) Human Oversight in Critical Decisions

High-stakes decisions, such as those involving access restrictions, incident escalations, or countermeasures, must involve human-in-the-loop (HITL) or human-on-the-loop (HOTL) mechanisms. While automation offers speed, human oversight provides context, ethical discretion, and accountability. Best practices emphasize hybrid governance models that enable AI to act autonomously within defined thresholds, but escalate to human intervention when those thresholds are breached [43].

(4) Continuous Auditing and Logging

Effective AI governance in cybersecurity requires mechanisms for traceability. All autonomous actions, including threat classifications, quarantine decisions, and overrides, should be logged and auditable. Continuous auditing helps detect bias, model drift, and systemic errors, enabling course correction and accountability. Log data also provides critical evidence in the event of disputes or forensic analysis following a breach [25]. Ethical design mandates that such logs are themselves protected and tamper-proof, especially in adversarial environments.

6. Recommendations for Future Regulation

To ethically manage AI-based cyber defense technologies, the current regulations, such as the GDPR, will have to be amended to make provisions regarding autonomy, explainability, and accountability of fast threat response. An important milestone would be the codification of a Right to Explanation to those users of automated systems that may be subject to a denial or placement of a device under quarantine. This will make it transparent and create meaningful redress. There is a pressing need to have certification standards specific to cybersecurity AI. These ought to evaluate explainability, data protection, and operational safety before deployment, particularly of systems whose actions are conducted without human control. Moreover, black-box systems have to be required to be tested with the help of auditing tools. Independent evaluation has the potential to uncover bias, monitor the system drift, and verify ethical integrity, even in critical settings, such as education or healthcare. Finally, regulators must adopt a proactive approach to avoid embedding unchecked AI into critical infrastructure. Cross-border alignment and faster legislative cycles are essential to stay ahead of evolving threats and build public trust.

7. Conclusion

This paper has examined the ethical, legal, and technical challenges autonomous AI systems pose in cyber defense, particularly concerning privacy, transparency, and regulatory accountability. While AI offers powerful capabilities in real-time threat detection, its unchecked deployment risks undermining user rights, institutional trust, and cross-border compliance. Bridging this gap requires not only technical safeguards but also regulatory foresight. Interdisciplinary collaboration between engineers, ethicists, and lawmakers is urgently needed to ensure that innovation remains aligned with human values and legal responsibility.

8. Future Suggestions

1. Develop sector-specific regulatory frameworks that address the unique risks of AI-based cyber defense in healthcare, finance, education, and government infrastructure.

2. Encourage interdisciplinary research that combines computer science, law, and ethics to design explainable and auditable AI systems for security-critical environments.

3. Establish international cooperation mechanisms to harmonize standards for AI governance in cybersecurity, particularly for cross-border data flows and threat intelligence sharing.

4. Promote continuous auditing and certification of AI-driven systems to ensure accountability, reduce automation bias, and detect unintended consequences early.

5. Invest in training programs for cybersecurity professionals and policymakers to strengthen human oversight and ethical decision-making in AI-enhanced defense systems.

Declarations

Source of Funding

This study received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Competing Interests Statement

The authors have declared that no competing financial, professional, or personal interests exist.

Consent for publication

All the authors contributed to the manuscript and consented to the publication of this research work.

Authors' contributions

All the authors took part in literature review, analysis, and manuscript writing equally.

Availability of data and materials

Supplementary information is available from the authors upon reasonable request.

Institutional Review Board Statement

Not applicable for this study.

Informed Consent

Not applicable for this study.

References

[1] Park, H., Azzaoui, A.E., & Park, J.H. (2025). AIDS-Based cyber threat detection framework for secure cloud-native microservices. Electronics, 14(2): 229–229. https://doi.org/10.3390/electronics14020229.

[2] Mohamed, N. (2025). Artificial intelligence and machine learning in cybersecurity: A deep dive into state-of-the-art techniques and future paradigms. Knowledge and Information Systems, 67: 6969–7055. https://doi.org/10.1007/s10115-025-02429-y.

[3] Buczak, A.L., & Guven, E. (2016). A survey of data mining and machine learning methods for cybersecurity intrusion detection. IEEE Communications Surveys & Tutorials, 18(2): 1153–1176. https://doi.org/10.1109/comst.2015.2494502.

[4] Mennella, C., Maniscalco, U., Pietro, G.D., & Esposito, M. (2024). Ethical and regulatory challenges of AI technologies in healthcare: A narrative review. Heliyon, 10(4): e26297–e26297. https://doi.org/10.1016/j.heliyon.2024.e26297.

[5] Buhrmester, V., Münch, D., & Arens, M. (2021). Analysis of explainers of black box deep neural networks for computer vision: A survey. Machine Learning and Knowledge Extraction, 3(4): 966–989. https://doi.org/10.3390/make3040048.

[6] Kokina, J., Blanchette, S., Davenport, T.H., & Pachamanova, D. (2025). Challenges and opportunities for artificial intelligence in auditing: Evidence from the field. International Journal of Accounting Information Systems, 56: 100734. https://doi.org/10.1016/j.accinf.2025.100734.

[7] Cheong, B.C. (2024). Transparency and accountability in AI systems: Safeguarding wellbeing in the age of algorithmic decision-making. Frontiers in Human Dynamics, 6. https://doi.org/10.3389/fhumd.2024.1421273.

[8] Aldossary, M. (2023). Multi-layer fog-cloud architecture for optimizing the placement of IoT applications in smart cities. Computers, Materials & Continua, 75(1): 633–649. https://doi.org/10.32604/cmc.2023.035414.

[9] Hilal, W., Gadsden, S.A., & Yawney, J. (2021). A review of anomaly detection techniques and applications in financial fraud. Expert Systems with Applications, 193(1): 116429. https://doi.org/10.1016/j.eswa.2021.116429.

[10] Feretzakis, G., Papaspyridis, K., Gkoulalas-Divanis, A., & Verykios, V.S. (2024). Privacy-Preserving techniques in generative AI and large language models: A narrative review. Information, 15(11): 697. https://doi.org/10.3390/info15110697.

[11] Data Protection Commission (2022). Principles of Data Protection | Data Protection Commission. Principles of Data Protection | Data Protection Commission. https://www.dataprotection.ie/en/individuals/data-protection-basics/principles-data-protection.

[12] Sarker, I.H., Janicke, H., Mohsin, A., Gill, A., & Maglaras, L. (2024). Explainable AI for cybersecurity automation, intelligence and trustworthiness in digital twin: Methods, taxonomy, challenges and prospects. ICT Express, 10(4). https://doi.org/10.1016/j.icte.2024.05.007.

[13] Fontes, C., Hohma, E., Corrigan, C.C., & Lütge, C. (2022). AI-powered public surveillance systems: Why we (might) need them and how we want them. Technology in Society, 71(0160–791x): 102137. https://doi.org/10.1016/j.techsoc.2022.102137.

[14] Saketh, M., Nandal, N., Tanwar, R., & Reddy, B.P. (2023). Intelligent surveillance support system. Discover Internet of Things, 3(1). https://doi.org/10.1007/s43926-023-00039-0.

[15] Macnish, K., & Van der Ham, J. (2020). Ethics in cybersecurity research and practice. Technology in Society, 63(101382). https://doi.org/10.1016/j.techsoc.2020.101382.

[16] Casarosa, F. (2020). Transnational collective actions for cross-border data protection violations. Internet Policy Review, 9(3). https://doi.org/10.14763/2020.3.1498.

[17] Dhirani, L.L., Mukhtiar, N., Chowdhry, B.S., & Newe, T. (2023). Ethical dilemmas and privacy issues in emerging technologies: A review. Sensors, 23(3): 1151. https://doi.org/10.3390/s23031151.

[18] Li, Z., Sharma, V., & P. Mohanty, S. (2020). Preserving data privacy via federated learning: Challenges and solutions. IEEE Consumer Electronics Magazine, 9(3): 8–16. https://doi.org/10.1109/mce.2019.2959108.

[19] Zhang, Z., Hamadi, H.A., Damiani, E., Yeun, C.Y., & Taher, F. (2022). Explainable artificial intelligence applications in cybersecurity: State-of-the-Art in research. IEEE Access, 10: 93104–93139. https://doi.org/10.1109/access.2022.3204051.

[20] Ferrara, E. (2023). Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. Sci., 6(1): 3. https://doi.org/10.3390/sci6010003.

[21] Longo, L., Brcic, M., Cabitza, F., Choi, J., Confalonieri, R., Del Ser, J., Guidotti, R., Hayashi, Y., Herrera, F., Holzinger, A., Jiang, R., Khosravi, H., Lecue, F., Malgieri, G., Páez, A., Samek, W., Schneider, J., Speith, T., & Stumpf, S. (2024). Explainable artificial intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions. Information Fusion, 106: 102301. https://doi.org/10.1016/j.inffus.2024.102301.

[22] Hamida, S.U., Jabed, M., Chakraborty, N.R., Biswas, K., & Sami, S.K. (2024). Exploring the landscape of explainable artificial intelligence (XAI): A systematic review of techniques and applications. Big Data and Cognitive Computing, 8(11): 149–149. https://doi.org/10.3390/bdcc8110149.

[23] Larriva-Novo, X., Pérez Miguel, L., Villagra, V.A., Álvarez-Campana, M., Sanchez-Zas, C., & Jover, Ó. (2024). Post-Hoc categorization based on explainable AI and reinforcement learning for improved intrusion detection. Applied Sciences, 14(24): 11511. https://doi.org/10.3390/app142411511.

[24] Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1(9): 389–399. https://doi.org/10.1038/s42256-019-0088-2.

[25] Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People, An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. Minds and Machines, 28(4): 689–707. https://doi.org/10.1007/s11023-018-9482-5.

[26] Kshetri, N. (2025). Transforming cybersecurity with agentic AI to combat emerging cyber threats. Telecommunications Policy, 49(6): 102976. https://doi.org/10.1016/j.telpol.2025.102976.

[27] Uddin, M., Irshad, M.S., Kandhro, I.A., Alanazi, F., Ahmed, F., Maaz, M., Hussain, S., & Ullah, S.S. (2025). Generative AI revolution in cybersecurity: A comprehensive review of threat intelligence and operations. Artificial Intelligence Review, 58(8). https://doi.org/10.1007/s10462-025-11219-5.

[28] Leslie, D. (2019). Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3403301.

[29] Bryson, J.J., Diamantis, M.E., & Grant, T.D. (2017). Of, for, and by the people: The legal lacuna of synthetic persons. Artificial Intelligence and Law, 25(3): 273–291. https://doi.org/10.1007/s10506-017-9214-9.

[30] Lin, P., Abney, K., & Bekey, G. (2011). Robot ethics: Mapping the issues for a mechanized world. Artificial Intelligence, 175(5–6): 942–949. https://doi.org/10.1016/j.artint.2010.11.026.

[31] Reynaud, S., & Roxin, A. (2025). Review of eXplainable artificial intelligence for cybersecurity systems. Discover Artificial Intelligence, 5(1). https://doi.org/10.1007/s44163-025-00318-5.

[32] Mittelstadt, B.D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. Big Data & Society, 3(2): 1–21. https://doi.org/10.1177/2053951716679679.

[33] Cummings, M.L. (2006). Automation and accountability in decision support system interface design. The Journal of Technology Studies, 32(1). https://doi.org/10.21061/jots.v32i1.a.4.

[34] Dzindolet, M.T., Pierce, L.G., Beck, H.P., & Dawe, L.A. (2002). The perceived utility of human and automated aids in a visual detection task. Human Factors: The Journal of the Human Factors and Ergonomics Society, 44(1): 79–94. https://doi.org/10.1518/0018720024494856.

[35] Woods, D.D. (2016). The risks of autonomy. Journal of Cognitive Engineering and Decision Making, 10(2): 131–133. https://doi.org/10.1177/1555343416653562.

[36] Lepri, B., Oliver, N., Letouzé, E., Pentland, A., & Vinck, P. (2017). Fair, transparent, and accountable algorithmic decision-making processes. Philosophy & Technology, 31(4): 611–627. https://doi.org/10.1007/s13347-017-0279-x.

[37] Mitrou, L. (2018). Data protection, artificial intelligence and cognitive services: Is the General Data Protection Regulation (GDPR) "artificial intelligence-proof"?. SSRN Electronic J. https://doi.org/10.2139/ssrn.3386914.

[38] NIST (2023). AI risk management framework. Artificial Intelligence Risk Management Framework (AI RMF 1.0), 1(1). https://doi.org/10.6028/nist.ai.100-1.

[39] European Commission (2024). Artificial intelligence – questions and answers. European Commission - European Commission. https://ec.europa.eu/commission/presscorner/detail/en/qanda_21_1683.

[40] Upadhayay, Y., & Sharma, R. (2025). Legal frameworks for AI in national security: Balancing innovation, ethics, and regulation. Journal of Neonatal Surgery, 14(10s): 500–508. https://doi.org/10.52783/jns.v14.2867.

[41] Bolatbekkyzy, G. (2024). Legal issues of cross-border data transfer in the era of digital government. Journal of Digital Technologies and Law, 2(2): 286–307. https://doi.org/10.21202/jdtl.2024.15.

[42] Veale, M., & Zuiderveen Borgesius, F. (2021). Demystifying the draft EU artificial intelligence act, analysing the good, the bad, and the unclear elements of the proposed approach. Computer Law Review International, 22(4): 97–112. https://doi.org/10.9785/cri-2021-220402.

[43] Cath, C. (2018). Governing artificial intelligence: Ethical, legal and technical opportunities and challenges. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 376(2133). https://doi.org/10.1098/rsta.2018.0080.

[44] Nemitz, P. (2018). Constitutional democracy and technology in the age of artificial intelligence. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 376(2133): 20180089. https://doi.org/10.1098/rsta.2018.0089.

[45] UNESCO (2023). Ethics of artificial intelligence. UNESCO. https://www.unesco.org/en/artificial-intelligence/recommendation-ethics.

[46] Taddeo, M., Blanchard, A., & Thomas, C. (2023). From AI ethics principles to practices: A teleological methodology to apply AI ethics principles in the defence domain. Social Science Research Network. https://doi.org/10.2139/ssrn.4520945.

[47] Jaiswal, A., & Mishra, P.C. (2024). Artificial intelligence (AI) and cybersecurity law: Legal issues in AI-driven cyber defense and offense. ShodhKosh Journal of Visual and Performing Arts, 5(6). https://doi.org/10.29121/shodhkosh.v5.i6.2024.4144.

[48] European Commission (2021). Proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain union legislative acts. Eur-Lex.europa.eu. https://eur-lex.europa.eu/legal-content/en/txt/?uri=celex:52021pc0206.

[49] NITDA (2019). Nigeria Data Protection Regulation 2019. https://nitda.gov.ng/wp-content/uploads/2020/11/nigeriadataprotectionregulation11.pdf.

[50] European Union (2016). General Data Protection Regulation. https://eur-lex.europa.eu/legal-content/en/txt/pdf/?uri=celex:32016r0679.

[51] Osobase, K.O., Felix, D., & Adeniran, O. (2025). Cloud-First supply chains: The impact of SAP S/4HANA cloud on logistics and procurement. Global Journal of Engineering and Technology Advances, 24(1): 206–217. https://doi.org/10.30574/gjeta.2025.24.1.0221.

[52] Nowamagbe, P., Oluwatosin, S., Goodness, N., Adefemi, B., Boluwade, J., Chibuike, C., & Agu, N. (2025). Architectural enhancements, challenges and future trends in real-time IoT applications over 5G networks. Global Journal of Engineering and Technology Advances, 15(3): 167–179. https://doi.org/10.30574/gjeta.2025.23.3.0185.

[53] Topol, E.J. (2019). High-performance medicine: The convergence of human and artificial intelligence. Nature Medicine, 25(1): 44–56. https://doi.org/10.1038/s41591-018-0300-7.

[54] Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. ACM Computing Surveys, 51(5): 1–42. https://doi.org/10.1145/3236009.

[55] Naik, N., Hameed, B. M.Z., Shetty, D.K., Swain, D., Shah, M., Paul, R., Aggarwal, K., Ibrahim, S., Patil, V., Smriti, K., Shetty, S., Rai, B.P., Chlosta, P., & Somani, B.K. (2022). Legal and ethical considerations in artificial intelligence in healthcare: Who takes responsibility? Frontiers in Surgery, 9(862322): 1–6. https://doi.org/10.3389/fsurg.2022.862322.

[56] Nnaka, K.I., Mbamalu, P.O., Nwaigbo, J.C., Ozo-ogueji, P.C., Njoku, V.I., & Ekechi, C.C. (2025). AI-powered threat detection: Opportunities and limitations in modern cyber defense. World Journal of Advanced Research and Reviews, 27(2): 210–223. https://doi.org/10.30574/wjarr.2025.27.2.2854.