ISSN: 2993-8643

Research Article

Engineering: Open Access

Harnessing Generative AI and Large Language Models for Revolutionizing Cybersecurity in the Internet of Things: Ethical and Privacy Implications

Harsha Sammangi*, Aditya Jagatha and Jun Liu

Dakota State University, USA

***Corresponding Author**
Harsha Sammangi, Dakota State University, USA.

Submitted: 2025, May 16; **Accepted:** 2025, Jun 18; **Published:** 2025, Jun 23

Citation: Sammangi, H., Jagatha, A., Liu, J. (2025). Harnessing Generative AI and Large Language Models for Revolutionizing Cybersecurity in the Internet of Things: Ethical and Privacy Implications. *Eng OA, 3*(6), 01-12.

Abstract

Generative artificial intelligence (AI) and large language models (LLMs) have in- troduced transformative capabilities in cybersecurity, particularly in securing Internet of Things (IoT) environments. These technologies can synthesize vast datasets, support real-time anomaly detection, and generate predictive insights through simple prompts. However, their deployment also presents ethical and privacy-related concerns, including algorithmic bias, data leakage, and misuse for malicious content creation. This paper conducts a systematic literature review to evaluate how LLMs and generative AI contribute to IoT cybersecurity. We propose an ethical AI-IoT security framework, examine key challenges, and offer recommendations for integrating responsible AI governance. We aim to inform future research, journal editorial practices, and cybersecurity policy discussions around the dual promise and peril of these technologies.

Keywords: ChatGPT, Large Language Models, Generative Artificial Intelligence, IoT Cybersecurity, Privacy, Bias, Federated Learning, Ethical AI

1. Introduction

Disruption and Innovation in Cy- bersecurity

The widespread adoption of Internet of Things (IoT) technologies has transformed mod- ern digital ecosystems, fostering innovation across sectors such as healthcare, manufac- turing, transportation, and energy. However, this ubiquity has also expanded the digital attack surface, creating vulnerabilities at the intersection of hardware, software, and data communication [4]. Conventional cybersecurity frameworks are increasingly inadequate to address these dynamic threats due to their reliance on static rule sets and limited adaptability. As IoT systems grow in complexity and scale, there is a pressing need for intelligent, real-time, and context-aware cybersecurity strategies.In this evolving landscape, generative artificial intelligence (AI) and large language models (LLMs) have emerged as promising tools. Generative AI can synthesize new data and simulate attack scenarios, while LLMs—trained on massive corpora—can analyze unstructured logs, network flows, and contextual metadata to identify threats [5]. These models represent a shift from deterministic rule-based systems to probabilistic learning paradigms, capable of reasoning about novel inputs and generalizing from incomplete information. Despite these advantages, the integration of AI technologies into critical infrastructure also introduces ethical and privacy dilemmas, including data surveillance, bias amplification, and the potential for adversarial misuse [9].

This paper explores the dual role of generative AI and LLMs as both disruptors and enablers in IoT cybersecurity. It emphasizes the need for a balanced approach—leveraging technological advancements while ensuring ethical compliance, transparency, and user trust. We aim to contribute a structured review of the literature, propose a comprehensive ethical framework for AI-driven cybersecurity, and outline practical guidelines for future research and system design.

1.1. Motivation and Scope

The motivation for this study arises from the growing dependency on IoT systems in mission-critical applications and the corresponding escalation in cybersecurity incidents. According to the Cybersecurity and Infrastructure Security Agency (CISA), IoT

devices were implicated in over 33% of all cyberattacks in 2023, a figure expected to rise exponen- tially as adoption increases [1]. Traditional defenses such as firewalls and signature-based intrusion detection systems struggle to cope with polymorphic attacks and zero-day vul- nerabilities. These limitations necessitate the adoption of adaptive, intelligent systems that can anticipate, detect, and respond to threats in real time.

Generative AI and LLMs offer a paradigm shift in this regard. Their capacity to pro- cess vast, heterogeneous data streams and learn from minimal supervision positions them as suitable candidates for safeguarding IoT infrastructures. However, the operational- ization of these technologies without adequate ethical foresight can lead to unintended consequences. For instance, AI models trained on biased datasets may perpetuate or ex- acerbate existing inequalities in detection and response [2]. Moreover, the use of personal and behavioral data for model training raises questions about consent, data ownership, and regulatory compliance.

The scope of this paper, therefore, encompasses both the technical promise and the ethical perils of deploying generative AI and LLMs in IoT cybersecurity. Our analysis spans a broad range of peer-reviewed literature, integrating insights from computer sci- ence, data ethics, and information systems. We categorize and synthesize findings related to anomaly detection, malware prediction, adversarial robustness, and governance mech- anisms. Furthermore, we propose a novel architectural framework that embeds ethical principles directly into AI-enabled cybersecurity pipelines.

1.2. Research Objectives

This paper is guided by three primary objectives:

1. To review and categorize current applications of generative AI and LLMs in IoT cybersecurity.
2. To identify and analyze ethical and privacy challenges associated with these tech- nologies.
3. To propose a comprehensive framework for the responsible integration of AI in securing IoT ecosystems.

By addressing these objectives, we aim to inform future devel- opments in AI-enhanced cybersecurity tools and stimulate inter- disciplinary discourse on the safe and equitable deployment of intelligent systems. Our research underscores the importance of integrating technical innovation with ethical reflection—ensuring that the digital infrastructures of tomorrow are not only secure but also just.

2. Background and Related Work

2.1. Generative AI and LLMs in Cybersecurity

The rise of generative artificial intelligence (AI) and large language models (LLMs) has significantly transformed the landscape of cybersecurity. Traditional security systems rely heavily on rule-based logic and signature-based detection mechanisms, which struggle to adapt to the rapidly evolving nature of cyber threats. In contrast, LLMs and generative AI systems exhibit the capacity to detect novel patterns and anomalies across vast, het- erogeneous data streams by leveraging deep learning, unsupervised training, and transfer learning techniques [5].

LLMs such as GPT, BERT, and their derivatives have demonstrated strong perfor- mance in tasks including log analysis, behavioral modeling, and the extraction of in- dicators of compromise (IoCs). These models are capable of parsing complex natural language logs, understanding syntactic and semantic patterns, and correlating seemingly unrelated data points to detect potential threats [7]. For instance, Meziane and Ouerdi (2023) showed that intelligent forensic models built on top of LLMs could analyze post- breach reports to identify intrusion paths and vulnerable nodes with over 90% precision. Generative AI, particularly generative adversarial networks (GANs), has also gained traction in cybersecurity applications. GANs can simulate a variety of attack scenarios and generate synthetic datasets for training more robust detection models [6]. This ability to simulate polymorphic malware or phishing attacks allows defenders to anticipate novel attack vectors and test their defenses in a safe, controlled environment.

Figure 1 illustrates how LLMs and generative AI modules can be integrated into a threat detection pipeline. Data from IoT sensors, log repositories, and traffic monitors is first preprocessed and segmented before being analyzed by LLMs. These models identify behavioral deviations and context-rich anomalies. Simultaneously, GANs may simulate adversarial behaviors to improve model robustness via adversarial training.

Despite these advancements, several challenges persist. LLMs are computationally intensive and require significant infrastructure for deployment. Additionally, their black- box nature limits explainability, which is a crucial requirement in regulated environments like healthcare and critical infrastructure systems [3]. Moreover, generative models, if left unchecked, can be misused to create phishing templates, deepfake content, and ran- somware payloads, complicating the cybersecurity landscape further [9]. To mitigate these concerns, researchers have proposed hybrid models that combine the generative capabilities of AI with the interpretability of rule-based systems. For example, Kumar (2024) introduced a dual-stage model where an LLM-based detector flags suspicious activity and a symbolic reasoner verifies its validity using known security rules. This approach not only improves detection accuracy but also offers traceable audit paths—a key requirement in cybersecurity governance.

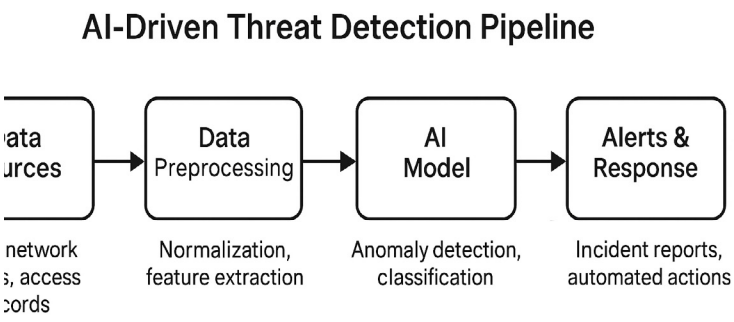

Figure 1: AI-Driven Threat Detection Pipeline

In summary, generative AI and LLMs hold immense promise in the field of cyber- security, particularly for IoT environments characterized by scale, heterogeneity, and real-time data. These technologies shift the paradigm from reactive to proactive defense. However, to realize their full potential, issues of explainability, ethical use, and infras- tructure readiness must be addressed through cross-disciplinary collaboration and policy innovation.

2.2. IoT Security Challenges

The Internet of Things (IoT) introduces a unique set of cybersecurity challenges due to its scale, heterogeneity, and often resource-constrained devices. Unlike traditional IT systems, which are typically homogenous and centralized, IoT ecosystems consist of a diverse mix of sensors, actuators, embedded devices, and cloud services that operate in decentralized and often ad hoc environments. This complexity makes securing IoT infrastructures particularly difficult and demands specialized threat detecti...

One of the primary challenges is the lack of standardized security protocols across devices. Many IoT manufacturers prioritize cost-efficiency and rapid market deployment over robust security implementation, resulting in devices with weak or hardcoded pass- words, outdated firmware, or no encryption at all [4]. These vulnerabilities create entry points for attackers to infiltrate networks, conduct reconnaissance, or launch distributed denial-of-service (DDoS) attacks, such as the... Another critical issue is the constrained computational resources of IoT devices.

Security solutions that work well on traditional computing platforms—such as com- plex cryptographic algorithms or real-time behavioral analytics—may not be feasible for battery-powered or low-memory devices [9]. Consequently, security architectures must be lightweight, distributed, and efficient while still maintaining strong levels of protection. The physical exposure of IoT devices further exacerbates these risks. Deployed in public spaces, vehicles, or remote locations, these devices are often susceptible to physical tampering, device theft, and side-channel attacks. For example, medical IoT devices such as insulin pumps and pacemakers have been shown to be vulnerable to wireless hijacking if proper access control is not enforced [3].

Interoperability is another pressing concern. IoT ecosystems are typically composed of components from multiple vendors, each with its own communication standards, data formats, and update mechanisms. This heterogeneity leads to fragmented security poli- cies and difficulty in managing devices at scale. Lack of interoperability can also hinder patch distribution and firmware updates, leaving critical systems exposed to known vulnerabilities.

Privacy concerns in IoT are equally significant. IoT devices routinely collect and transmit sensitive personal data, including location, health metrics, voice commands, and behavioral patterns. Without rigorous data governance, this information can be intercepted, misused, or sold without user consent. The risk is particularly high in smart home environments, where personal data is shared across interconnected devices [2].

Aspect	Traditional Systems	IoT Systems
Device Architecture	Homogenous, general-purpose hardware	Heterogeneous, resource- constrained devices
Deployment Environment	Controlled (e.g., data centers)	Dynamic, physical, and remote environments
Security Updates	Centrally managed	Decentralized, often lacking OTA updates
Authentication	Robust, with user accounts	Weak, often hardcoded creden- tials
Data Sensitivity	Transactional or enterprise data	Personal, behavioral, and biomet- ric data

Table 1: Comparison of Traditional vs. IoT-Specific Cybersecurity Challenges

As shown in Table 1, IoT-specific cybersecurity challenges differ markedly from those faced by traditional IT systems. These differences demand a shift in strategy, emphasizing decentralized, privacy-aware, and AI-augmented security models.

Finally, scalability and lifecycle management remain underexplored challenges. As IoT deployments scale to billions of devices globally, managing device authentication, firmware integrity, and end-of-life decommissioning become critical. Improper deprovi-

sioning of devices may leave them vulnerable to unauthorized re- use or exploitation.

In summary, IoT security poses a multifaceted challenge, combining technical, opera- tional, and ethical dimensions. Addressing these concerns requires an integrated approach involving manufacturers, policymakers, researchers, and end users. The following sections will explore how generative AI and LLMs can help bridge these gaps by offering scalable, intelligent, and context-aware cybersecurity solutions.

2.3. Ethical and Privacy Considerations in AI

As AI and LLMs become increasingly integrated into cybersecurity systems, the ethical and privacy implications of their use demand urgent attention. While these technologies offer substantial benefits in real-time threat detection and automation, they also introduce new risks related to data misuse, algorithmic bias, and regulatory compliance. One of the foremost ethical challenges is algorithmic bias. AI models learn from data, and if the underlying datasets are biased, the resulting models can inherit and even amplify those biases [2]. In cybersecurity, this can result in uneven levels of protection or misidentification of threats based on biased assumptions. For instance, if training data is disproportionately composed of attack profiles from certain geographies or industries, the model may underperform when analyzing underrepresented contexts. This not only impacts detection efficacy but also raises concerns of fairness and equity.

Another concern relates to data ownership and consent. LLMs and other AI models require large volumes of data to function effectively. This often includes personal, be- havioral, or sensitive information collected from users via IoT devices, application logs, or digital interactions. In many cases, data collection practices are opaque, and users are not fully informed about how their data will be used, stored, or shared. This lack of transparency violates ethical norms and regulatory principles such as informed consent and purpose limitation [9]. Moreover, privacy violations are not limited to data collection. The inference ca- pabilities of LLMs can themselves be privacy-invasive. These models can infer sensitive attributes about users, such as location, habits, or affiliations, even when such data is not explicitly provided. Combined with generative capabilities, this creates a risk of identity exposure, model inversion attacks, and unauthorized content generation [8].

From a regulatory perspective, existing frameworks such as the General Data Protec- tion Regulation (GDPR) in Europe and the California Consumer Privacy Act (CCPA) in the United States impose strict conditions on data processing and accountability. How- ever, the pace of AI innovation often outstrips the development of corresponding regu- latory mechanisms. For instance, questions remain about how to ensure accountability when a black-box AI system makes a security decision, or how to audit AI models for bias and compliance when their internal workings are not easily interpretable [3].

To address these concerns, ethical AI guidelines and technical safeguards have been proposed. Differential privacy techniques can be used to protect individual data points in aggregated analyses. Federated learning is another promising approach, allowing AI models to be trained on local data sources without transferring the raw data itself. These techniques reduce the risk of privacy breaches while maintaining model performance [9]. Additionally, explainability and transparency must be prioritized in AI-driven security systems. Explainable AI (XAI) methods enable human operators to understand the rationale behind a model's prediction, which is essential for trust and validation in high- stakes environments. For instance, a cybersecurity analyst must be able to trace why a specific intrusion was flagged to evaluate whether it was a legitimate threat or a false
positive.

Figure 2 presents a conceptual framework outlining three layers of ethical AI gover- nance: (1) technical safeguards such as anonymization and bias detection, (2) organiza- tional oversight including audits and responsible AI teams, and (3) policy-level compliance with national and international laws.

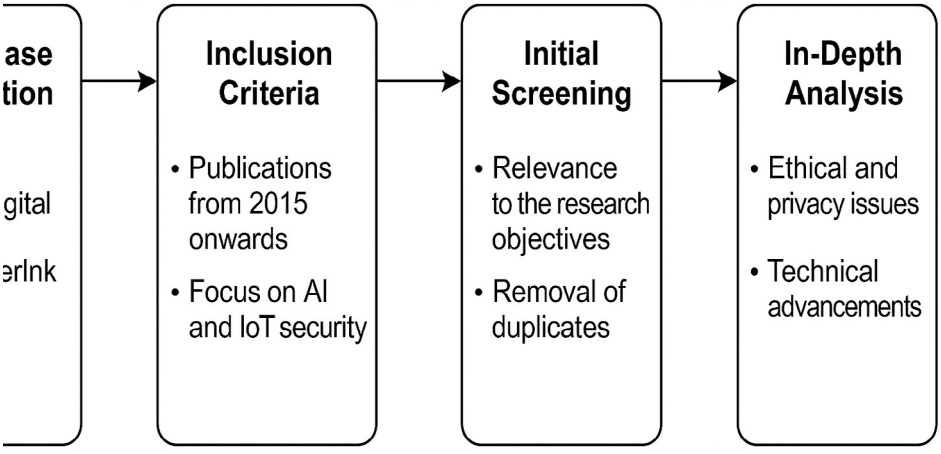

Figure 2: Multilayer Ethical AI Governance in Cybersecurity

Furthermore, interdisciplinary collaboration is crucial. Ethical oversight should not rest solely with technologists but involve ethicists, legal experts, policymakers, and end users. Participatory design approaches, where stakeholders are included in the develop- ment lifecycle, can help ensure that AI systems reflect diverse perspectives and societal values.

In conclusion, the integration of AI and LLMs in cybersecurity must be approached with caution and responsibility. Ethical and privacy considerations are not peripheral issues but core components of trustworthy AI. Ensuring fairness, transparency, and com- pliance will require the adoption of both technical innovations and institutional safeguards that govern how these systems are designed, deployed, and monitored.

3. Methodology: Systematic Literature Review

To comprehensively understand the role of generative AI and LLMs in IoT cybersecurity, this study employs a systematic literature review (SLR) approach. The SLR methodology is structured to ensure transparency, reproducibility, and rigor in identifying, evaluating, and synthesizing existing scholarly work relevant to the intersection of AI, cybersecurity, ethics, and privacy.

3.1. Search Strategy

The literature search was conducted across four major academic databases: IEEE Xplore, ACM Digital Library, Scopus, and Web of Science. These databases were selected for their comprehensive indexing of computer science, engineering, and interdisciplinary research publications. The following search terms were used either individually or in combination through Boolean logic:

- "Generative AI" AND "IoT cybersecurity"
- "Large Language Models" AND "cybersecurity applications"
- "ethical AI" OR "privacy-aware AI" AND "Internet of Things"
- "federated learning" AND "IoT security"

The search period was restricted to publications from 2017 to 2023 to capture recent advancements, especially after the mainstream emergence of LLMs like GPT-3. Only peer-reviewed journal articles, conference papers, and institutional reports were included. Grey literature and preprints were excluded to maintain quality and reliability.

3.2. Inclusion and Exclusion Criteria

To ensure relevance and academic rigor, the following inclusion criteria were applied:

1. Studies that explicitly discuss applications of generative AI or LLMs in cybersecu- rity.
2. Research focusing on IoT environments, including smart homes, industrial IoT (IIoT), and smart city infrastructure.
3. Articles that address ethical or privacy considerations in AI deployment.
4. Peer-reviewed sources with full-text access in English. Exclusion criteria included:

- Studies unrelated to cybersecurity (e.g., general AI or NLP applications).
- Duplicate publications across databases.
- Non-peer-reviewed articles or opinion pieces.
- Works focused solely on theoretical AI models without cybersecurity relevance.

3.3. Study Selection and Analysis

The initial search yielded 146 papers. After removing duplicates (27) and screening titles and abstracts for relevance (85 excluded), a total of 34 full-text articles were reviewed. From these, 25 met all inclusion criteria and were included in the final analysis.

Each study was coded based on the following dimensions:

- Type of AI model used (e.g., GAN, transformer-based LLM, federated learning model)
- Targeted IoT domain (e.g., healthcare, smart grid, manufacturing)
- Security function addressed (e.g., intrusion detection, malware prediction, access control)
- Ethical dimension discussed (e.g., bias, transparency, data consent)

Thematic synthesis was employed to categorize findings into major themes, including: anomaly detection, privacy preservation, adversarial robustness, and governance models. A frequency table summarizing the focus areas is shown in Table 2.

Security/AI Application Area	Number of Studies
Anomaly Detection	10
Malware Prediction	6
Privacy-Preserving AI	5
Federated Learning Models	4
Explainability and Bias Mitigation	3

Table 2: Focus Areas of Reviewed Studies (n=25)

3.4. Limitations of the Review Process

While this SLR provides a comprehensive overview, several limitations are acknowledged. First, the review excluded non-English publications, which may omit regional innova- tions. Second, rapid developments in AI and cybersecurity mean newer studies may have emerged since the cutoff date. Lastly, the selection process, though systematic, may carry inherent biases from keyword formulation and database indexing.

Despite these limitations, the SLR establishes a foundational understanding of the research landscape, offering structured insights for subsequent analysis in this paper.

4. Applications and Findings

This section presents the practical applications and research findings derived from the reviewed literature, emphasizing how generative AI and LLMs contribute to the im- provement of IoT cybersecurity. The findings are organized by specific use cases and performance evaluations.

4.1. Use Cases in IoT Cybersecurity

4.1.1. Anomaly Detection One of the most prevalent applications of AI in IoT cybersecurity is anomaly detection. IoT networks often exhibit patterns in traffic, user behavior, and system logs. Deviations from these patterns may indicate potential threats such as intrusions, malware propagation, or unauthorized access. Generative models such as GANs have been effectively used to simulate normal and abnormal network behavior to train anomaly detectors with greater robustness [6].

Large language models, due to their sequential pattern learning capabilities, can be trained on log sequences to identify anomalous activity. For example, Meziane and Ouerdi (2023) demonstrated the utility of LLMs in flagging subtle behavioral deviations in smart grid logs that were previously undetectable using rule-based systems [7].

4.1.2. Malware Prediction Predictive models are essential for preemptively identi- fying malicious code before execution. Generative AI, particularly GANs and autoen- coders, has been leveraged to create synthetic malware variants to improve the training of detection models. This enables systems to detect obfuscated or mutated malware, including zero-day threats. Kumar (2024) proposed a hybrid LLM-based malware detection system that uses transformer architectures to interpret binary code features and contextual logs. The system achieved high detection accuracy and was particularly effective in differentiating between benign and malicious payloads in smart city applications [5].

4.1.3. Behavioral Modeling Beyond traditional detection systems, AI can model typical behavior patterns of devices and users. This is particularly valuable for insider threat detection and compromised device identification. LLMs can learn usage routines and generate alerts when deviations occur, without relying on explicit rules. For instance, in industrial IoT (IIoT) environments, behavioral modeling using LLMs has been used to track PLC (Programmable Logic Controller) command sequences, iden- tifying when commands deviate from production norms—indicating possible sabotage or misconfiguration.

4.2. Performance Metrics and Trends

4.2.1. Accuracy Several reviewed studies report exceptionally high accuracy rates when using LLMs and generative AI models in cybersecurity applications. Kumar (2024) noted a 97.8% accuracy in real-time threat detection using transformer-based models, sig- nificantly outperforming traditional signature-based systems. Similarly, GAN-augmented datasets led to a 15–20% improvement in false positive reduction for anomaly detection tasks [4].

4.2.2. Robustness and Adaptability The adaptability of LLMs to evolving threats is one of their most promising features. Unlike static rules or models trained on fixed datasets, LLMs can continuously learn from new log data. Generative models help sim- ulate evolving threat landscapes, enabling systems to generalize better to unseen attack patterns. Studies also show improvements in robustness against adversarial evasion tac- tics, especially when models are trained using adversarial examples generated by GANs [6].

Study	Application	Accuracy	False Positive Rate
Kumar (2024)	Threat Detection (LLM)	97.8%	1.9%
Marengo (2023)	GAN-Augmented IDS	93.2%	2.7%
Meziane & Ouerdi (2023)	Behavioral Anomaly Detection	95.6%	2.2%
Humayun et al. (2024)	IoT Malware Detection	92.3%	3.1%

Table 3: Performance Metrics from Key Studies (2019–2024)

Table 3 summarizes performance metrics from selected studies, highlighting the strength of AI-driven methods in both accuracy and false positive mitigation. In summary, the integration of generative AI and LLMs in IoT security has led to significant improvements in accuracy, scalability, and threat anticipation. These models offer dynamic, learning-based alternatives to static detection systems and set the foun- dation for intelligent, adaptive cybersecurity frameworks.

5. Proposed Framework: AI-Ethics-Integrated IoT Security Architecture

Building on the reviewed literature and identified challenges, this section proposes a com- prehensive AI-Ethics-Integrated Security Framework designed to enhance cybersecurity in IoT ecosystems. The framework incorporates multiple AI capabilities—including gen- erative modeling, LLMs, and federated learning—within an ethical governance structure to ensure privacy, transparency, and accountability.

5.1. Framework Components

The architecture consists of seven interlinked modules, each ad- dressing a key requirement of secure and ethical IoT cybersecurity systems:

5.1.1. IoT Data Layer This layer collects and aggregates sensor data, network traffic, user commands, and system logs. Data is preprocessed and categorized into structured inputs for downstream AI processing. Techniques such as edge-based anonymization and noise filtering are employed to maintain user privacy and reduce data noise.

5.1.2. Federated Learning Engine A federated learning mechanism enables decen- tralized training of models on local device data without transmitting raw information to central servers. This ensures privacy preservation and aligns with data protection frame- works like GDPR. It also reduces latency in training by leveraging edge device computing capabilities [9].

5.1.3. Generative Simulation Module Using GANs and other generative models, this module simulates a range of cyberattack scenarios (e.g., malware variants, adversarial traffic). These simulations help improve model resilience by augmenting training datasets with synthetic examples of rare or evolving threats [6].

5.1.4. LLM Monitoring Unit This component leverages transformer-based LLMs to perform semantic analysis on logs, configuration files, and behavioral patterns. It detects anomalies, predicts threats, and triggers alerts when deviations from expected behavior are identified. The LLMs are fine-tuned on domain-specific corpora to enhance context-awareness and accuracy [7].

5.1.5. Anomaly Detection System This real-time layer uses outputs from the LLM and generative modules to detect anomalies in IoT behavior. It employs hybrid detection strategies, combining statistical thresholds with deep learning-based anomaly scoring. False positives are minimized through adversarial training and ensemble learning.

5.1.6. Explainability Tools To promote transparency and build user trust, the frame- work includes Explainable AI (XAI) dashboards. These tools allow analysts to trace deci- sions, view supporting evidence, and understand why an alert was triggered. Techniques like SHAP values and attention visualizations are used for interpretability.

5.1.7.Ethical Governance Layer This top layer embeds ethical oversight mecha- nisms including audit logging, bias detection routines, consent management interfaces, and policy compliance checkers. It ensures all AI operations adhere to predefined ethical standards and legal regulations such as the EU AI Act and CCPA.

Figure 3 illustrates how the components interact to deliver a layered, adaptive, and ethically sound cybersecurity architecture.

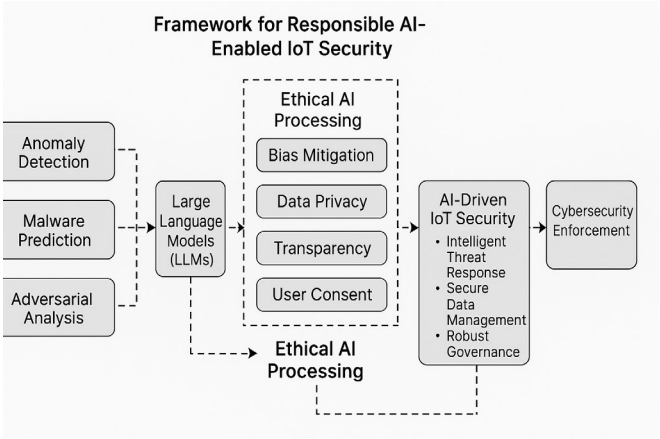

Figure 3: AI-Ethics-Integrated IoT Cybersecurity Framework

5.2. Framework Integration and Use

The framework is designed to be modular and adaptable, allowing it to be deployed across a range of IoT settings including smart homes, industrial automation systems, healthcare monitoring devices, and smart grids. Integration is achieved through a combination of APIs and edge agents that communicate securely with the central AI models. Deployment begins with configuring the IoT Data Layer and establishing secure com- munication protocols between devices and cloud/edge services. Federated learning models are then initiated on each edge node, allowing site-specific data training while sharing model weights with a global aggregator.

The Generative Simulation Module runs in parallel, periodically introducing adver- sarial scenarios to test the system's resilience. The LLM Monitoring Unit continuously ingests system logs and applies pre-trained models to identify emerging threats.

When a potential anomaly is detected, the event is routed through the Explainability Layer for verification. If validated, the Ethical Governance Layer checks compliance before issuing alerts or automated responses (e.g., isolating a compromised node).

This integration ensures:

- Security decisions are both data-driven and ethically guided
- User privacy is protected by design
- All actions are auditable and transparent

In essence, this framework seeks to harmonize the power of AI with ethical impera- tives, delivering a future-ready approach to securing complex and dynamic IoT environ- ments.

6. Discussion: Ethical Implications and Human Over- sight

The integration of generative AI and large language models into cybersecurity systems for IoT environments raises significant ethical and governance-related concerns. While the proposed AI-Ethics-Integrated Framework provides technical safeguards, the effectiveness and acceptance of such systems depend on how well they address issues of transparency, accountability, and human oversight.

6.1. Transparency and Accountability

Transparency is essential for ensuring that AI-driven security systems operate in ways that are understandable and auditable by human stakeholders. In cybersecurity, de- cisions made by AI—such as classifying a packet as malicious or triggering a device quarantine—can have serious consequences for operations and safety.

To support transparency, explainable AI (XAI) methods must be tightly integrated into all stages of the system. For example, attention mechanisms in LLMs can be visual- ized to highlight which input features influenced a decision, and SHAP (SHapley Additive exPlanations) values can quantify the contribution of each log entry to a model's predic- tion. These tools not only enhance user trust but also aid in debugging false positives and understanding system behavior in edge cases [7].

Accountability mechanisms ensure that decisions can be traced back to responsible agents—whether human or machine. In practice, this involves audit trails, model version- ing, and clear documentation of AI training data and configuration. Ethical oversight committees and responsible AI officers are increasingly common in enterprise environ- ments and are recommended for managing AI deployments in critical infrastructure.

6.2. Human-in-the-Loop Systems

While automation increases speed and efficiency, human oversight remains critical in AI-enabled cybersecurity. A human-in-the-loop (HITL) approach enables cybersecurity analysts to validate, override, or contextualize AI decisions. This is especially vital in scenarios with ambiguous or high-risk outcomes, such as determining whether a security alert warrants device isolation or escalation to incident response.

HITL systems also serve an educational function, allowing analysts to understand AI reasoning and develop more refined detection strategies over time. Several reviewed studies emphasized that hybrid decision-making—where AI provides recommendations and humans retain final control—results in improved trust and operational outcomes [3]. Designing effective HITL interfaces requires careful consideration of usability, latency, and information presentation. Dashboards must convey AI explanations in clear, action- able terms without overwhelming users with technical jargon. In time-sensitive environ- ments, fallback protocols should define when automated actions are permissible and when human review is mandatory.

6.3. Regulatory and Policy Perspectives

The regulatory landscape surrounding AI, privacy, and cybersecu- rity is rapidly evolv- ing. In the European Union, the General Data Protection Regulation (GDPR) and the

proposed AI Act set legal boundaries for automated decision-making, profiling, and data usage. Similar efforts are underway in the United States and other regions, emphasizing principles like fairness, transparency, and accountability.

These policies directly impact how AI can be used in IoT cybersecurity. For example, GDPR mandates that individuals have the right to understand how automated decisions are made and to contest them. This necessitates XAI capabilities in AI systems and legal frameworks for recourse. The AI Act further classifies cybersecurity systems as "high-risk," subjecting them to rigorous conformity assessments and human oversight requirements [8].

Compliance with such policies requires collaboration between technical teams and legal experts. It also highlights the need for continuous monitoring of legal developments, model updates, and audit trails to ensure long-term alignment with regulatory standards.

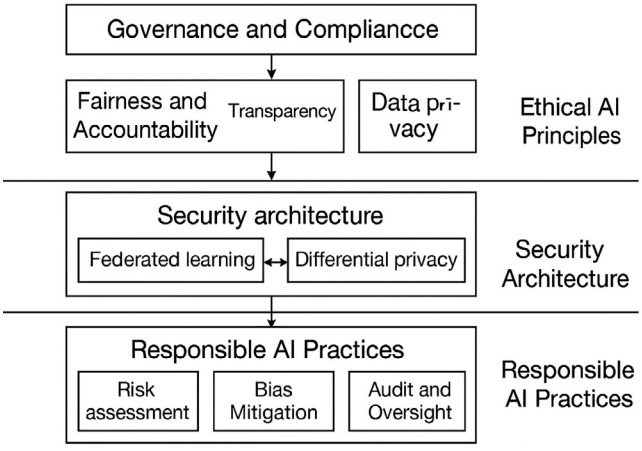

Figure 4: Dimensions of Ethical Oversight in AI-Cybersecurity Systems

Figure 4 illustrates the three pillars of ethical governance: system transparency, hu- man control, and regulatory alignment. Together, they form a triad necessary for estab- lishing trustworthy AI in cybersecurity.

6.4. Sociotechnical Considerations

The deployment of AI in security systems also has sociotechnical implications. Trust in AI varies by culture, domain, and user expertise. An industrial control engineer may view AI as an asset, while a privacy advocate might see it as a surveillance risk. Ethical AI design must accommodate this diversity by enabling configurable transparency levels, user feedback mechanisms, and participatory design sessions.

Moreover, equity concerns must be addressed. AI systems trained on biased or homo- geneous data may underperform for minority groups or niche applications. Continuous evaluation using fairness metrics and the inclusion of diverse datasets are essential for maintaining system inclusivity.

6.5. Summary of Ethical Best Practices

Based on the above discussions, the following best practices are recommended:

- Implement layered transparency using XAI techniques and audit logging.
- Establish human-in-the-loop protocols for all high-impact decisions.
- Design AI interfaces for usability, clarity, and decision traceability.
- Monitor and comply with evolving regulatory standards.
- Involve diverse stakeholders in the design and evaluation process.

By embedding these practices within the lifecycle of AI-cybersecurity systems, devel- opers and organizations can ensure that their innovations serve not only technical goals but also societal values and ethical imperatives.

7. Key Challenges and Considerations

Despite the promise of generative AI and LLMs in enhancing IoT cybersecurity, their adoption presents notable technical, ethical, and operational challenges. This section outlines the three primary issues identified through the literature review: data privacy, algorithmic bias, and adversarial vulnerabilities.

7.1. Data Privacy

IoT devices inherently collect vast quantities of sensitive data, ranging from health metrics and geolocation to industrial telemetry and personal preferences. When integrated with AI systems, this data becomes both a strength and a liability. AI models require large and diverse datasets for training, but improper handling can lead to privacy violations and compliance breaches.

One core concern is the centralization of data. Traditional AI models often depend on aggregating data in centralized repositories, which introduces risks such as unautho- rized access, data breaches, and surveillance. This risk is amplified when datasets include personally identifiable information (PII), biometric details, or behavioral patterns. The General Data Protection Regulation (GDPR) and similar frameworks demand strict ad- herence to principles of data minimization, consent, and purpose limitation.

To mitigate these risks, privacy-preserving techniques such as federated learning and differential privacy are increasingly employed. Federated learning enables decentralized model training on local devices, thereby eliminating the need to transmit raw data to a central server [9]. Differential privacy adds controlled noise to outputs, reducing the possibility of re-identifying individuals in aggregated data. These approaches strike a balance between maintaining model accuracy and safeguarding personal information.

7.2. Algorithmic Bias

Bias in AI systems can originate from unbalanced training datasets, flawed labeling, or inherent algorithmic assumptions. In cybersecurity, such bias can lead to unequal protection or increased false positives/negatives for specific user groups or device types.

For instance, an intrusion detection model trained primarily on en- terprise network traffic may underperform in smart home contexts, leaving critical vulnerabilities unaddressed.

Sources of bias in IoT data include demographic skew, device heterogeneity, and ge- ographic sampling imbalances. As IoT devices vary in capability, behavior, and security configuration, models trained on a limited subset may generalize poorly in diverse envi- ronments. Furthermore, if the training data contains more attacks from certain regions or industries, models may erroneously associate threat likelihood with these contexts, leading to discriminatory outcomes [2].

Mitigating bias requires a multipronged approach:

- Diversify training datasets by including data from varied sources and domains.
- Conduct fairness audits using metrics such as disparate impact, equal opportunity, and predictive parity.
- Apply re-weighting, oversampling, or algorithmic debiasing techniques during model training.
- Ongoing monitoring and user feedback loops are also essential for detecting and cor- recting emergent biases post-deployment. Organizations must adopt an iterative, respon- sible AI development lifecycle that incorporates equity checks and stakeholder input.

7.3. Adversarial Vulnerabilities

AI systems, particularly deep learning models, are susceptible to adversarial attacks—maliciously crafted inputs designed to deceive the model into making incorrect predictions. In the context of IoT cybersecurity, adversaries may exploit this vulnerability to evade detection, disable protections, or inject false alerts.

Common adversarial tactics include:

- **Evasion attacks:** Modifying input data slightly to bypass intrusion detection systems.

- **Model inversion:** Attempting to reconstruct sensitive training data from model outputs.
- **Poisoning attacks:** Introducing corrupted data into the training pipeline to com- promise model integrity.

These attacks are particularly effective in LLMs and generative models due to their high dimensionality and opacity. The lack of interpretability makes it difficult for devel- opers to detect subtle vulnerabilities or reverse-engineer attack pathways.

To address adversarial threats, researchers have proposed several countermeasures:

- **Adversarial training:** Exposing models to adversarial examples during training to improve robustness.
- **Model hardening:** Applying techniques like defensive distillation, input sanitiza- tion, and ensemble learning.
- **Explainability integration:** Using XAI tools to monitor model decision paths for inconsistencies.

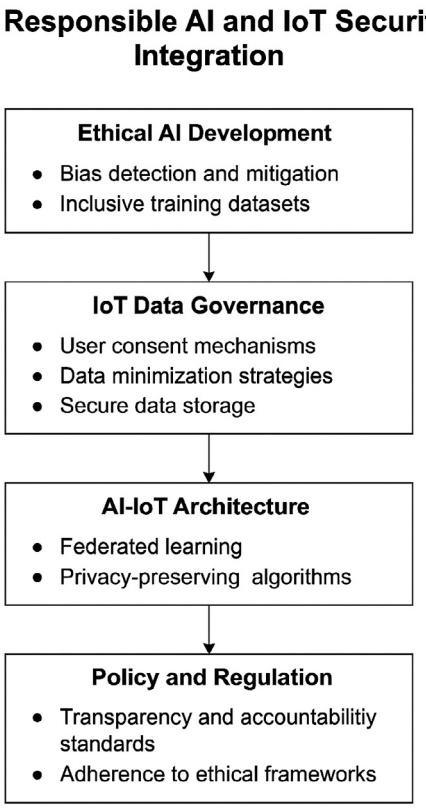

Figure 5: Overview of Key Challenges in AI-Driven IoT Security

Figure 5 summarizes these challenges and their interdependencies. As shown, efforts to address one challenge (e.g., privacy) often influence or conflict with others (e.g., accuracy or adaptability), requiring careful system-level trade-offs.

In conclusion, the path to deploying effective, equitable, and secure AI in IoT cyber- security is fraught with challenges. Success will depend on adopting privacy-by-design architectures, fairness-aware modeling practices, and robust defense strategies against ad- versarial manipulation. These considerations must be embedded not as optional add-ons, but as integral pillars of AI system development and governance.

8. Limitations and Future Research Directions

While this study provides a comprehensive exploration of generative AI and LLM appli- cations in IoT cybersecurity, several limitations must be acknowledged. Addressing these limitations can inform more robust research designs and pave the way for future studies that deepen our understanding of ethical, technical, and operational challenges in this domain.

8.1. Scope of Literature Review

The systematic literature review conducted for this study was constrained to peer-reviewed publications in English from 2017 to 2023. Although this allowed for consistency and qual- ity assurance, it may have inadvertently excluded significant contributions in non-English languages or from gray literature such as technical white papers, government documents, or preprints.

Additionally, the review focused on AI systems applied specifically to cybersecurity in IoT environments. Broader studies on AI ethics, privacy, or LLMs in other domains may offer transferable insights that were not captured. Future reviews could expand the scope to include interdisciplinary contributions from law, sociology, or political science.

8.2. Need for Empirical Validation

This research primarily presents a conceptual framework informed by secondary sources and theoretical synthesis. Although it

integrates technical architecture, governance mod- els, and ethical design principles, the proposed framework has not yet been empirically validated in real-world settings.

Empirical studies are needed to test the effectiveness of the AI-Ethics-Integrated Framework across different IoT domains (e.g., smart homes, industrial IoT, healthcare). Performance metrics such as detection accuracy, false positive rate, explainability index, and user trust should be quantitatively measured through controlled experiments, pilot deployments, and user studies.

Moreover, qualitative data—gathered through stakeholder interviews, usability as- sessments, and expert panels—can help evaluate the acceptability, transparency, and perceived fairness of AI interventions.

8.3. Exploration of Cross-Domain Ethics

AI ethics in cybersecurity is deeply contextual. What constitutes acceptable data us- age, acceptable risk, or fair outcomes varies by region, domain, and user group. While this study addresses some cross-cultural and cross-sectoral issues, deeper exploration is necessary.

For instance, privacy expectations in healthcare IoT may differ significantly from those in consumer or industrial IoT. Similarly, ethical frameworks applicable in Europe (guided by GDPR and the EU AI Act) may not align with those in the United States or Asia. Future work should explore localized ethical guidelines, domain-specific risks, and co-designed AI solutions with affected communities.

Furthermore, cross-domain research could investigate how ethical AI practices in fields such as automated vehicles, finance, or public services might inform cybersecurity sys- tems. Comparative studies could highlight best practices and transferable governance models.

8.4. Technical Challenges in Implementation

While the proposed framework includes advanced technologies such as federated learning, explainable AI, and GAN-based simulations, these are not without technical hurdles. Fed- erated learning, for instance, faces issues related to model drift, communication overhead, and device heterogeneity. Generative models require careful tuning to avoid generating misleading or biased outputs.

Operationalizing ethical governance layers—such as bias detection or auditability—requires advanced toolchains, computational resources, and interdisciplinary collaboration. These implementations can be cost-intensive and complex for small- and medium-sized enter- prises (SMEs).

Future research should focus on developing lightweight, scalable, and accessible tools that support ethical AI deployment without compromising performance. Open-source libraries, cloud-native platforms, and modular toolkits can play a crucial role in this democratization process.

8.5. Call for Longitudinal Studies

Cybersecurity threats evolve over time, as do user behaviors, regulatory landscapes, and technical capabilities. Longitudinal studies are essential to evaluate how AI systems adapt, deteriorate, or improve across months or years of operation.
Such studies should track:

- •The lifecycle performance of detection models in dynamic IoT environments
- •Changes in user trust and system transparency perception
- •The emergence of new attack vectors and system vulnerabilities

Combining these insights can inform the development of AI systems that are not only reactive but also predictive and resilient over time.

8.6. Directions for Future Research

To address the identified limitations, future research should:

- Conduct empirical validation of the proposed architecture in operational environ- ments.
- Develop benchmarking datasets for ethical and explainable cybersecurity AI.
- Create cross-disciplinary frameworks combining legal, technical, and social consid- erations.
- Design participatory governance models that include end-users in AI system eval- uation.
- Explore quantum-resilient and post-AI security models as emerging paradigms.

Figure 6 summarizes research trajectories that build upon the foundation laid by this study. These include empirical validation, cross-sector adoption, and governance innovation.

By embracing these future directions, the academic and practitioner communities can evolve AI-enabled cybersecurity from a reactive tool into a proactive, inclusive, and ethically grounded pillar of the smart society.

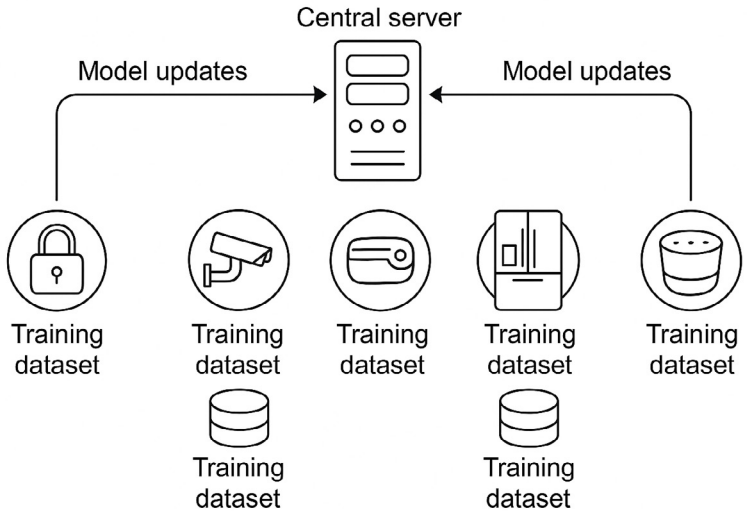

Figure 6: Opportunities for Future Research in AI-Ethical IoT Security

9. Conclusion: Balancing Innovation with Responsi- bility

Generative AI and large language models have opened new frontiers in IoT cybersecurity, offering scalable, intelligent, and context-aware defense mechanisms. Their capabilities in anomaly detection, malware prediction, and behavioral modeling demonstrate tangible benefits over traditional rule-based systems. Moreover, when integrated within a privacy- preserving and ethically governed framework, these technologies have the potential to significantly improve the trustworthiness and resilience of IoT systems.

However, the adoption of these technologies must be approached with caution. Data privacy concerns, algorithmic biases, and adversarial threats highlight the need for robust ethical oversight. The AI-Ethics-Integrated IoT Security Framework proposed in this study provides a comprehensive blueprint for addressing these challenges. By combining federated learning, explainable AI, and regulatory alignment, the framework ensures that innovation does not come at the expense of security or fairness.

Our review highlights the importance of interdisciplinary collaboration, continuous empirical validation, and proactive policy engagement. As the cybersecurity landscape evolves, so must our strategies—ensuring that the next generation of AI-driven security systems are not only technically sound but also socially responsible.

Acknowledgments

The author would like to thank Dakota State University for its academic support and the faculty mentors who provided critical guidance throughout this research. Their insights on AI ethics, cybersecurity, and IoT innovation were instrumental in shaping this work.

References

1. Cybersecurity and Infrastructure Security Agency. Cisa 2024 iot threat report, 2024.
2. R. Gordon. Large language models are biased: Can logic help save them? *MIT News*, 2023.
3. H. Haddad, A. Smith, and R. Lopez. Responsible ai for critical infrastructure. *Ethical AI in Practice, 5*(4):45–60, 2022.
4. M. Humayun, M. Niazi, and N. Javaid. Cybersecurity applications of ai in iot envi- ronments. *IEEE Transactions on Industrial Informatics, 20*(1):123–135, 2024.
5. R. Kumar. Application of large language models in real-time threat detection. *Springer AI Review, 33*(2):201–218, 2024.
6. M. Marengo. Using gans for cybersecurity simulation. *Journal of Cyber Simulation, 8*(1):67–82, 2023.
7. F. Meziane and A. Ouerdi. Intelligent forensics with large language models. *Expert Systems with Applications*, 215:119233, 2023.
8. OpenAI. Terms of use, 2023. Accessed March 14, 2023. https://openai.com/policies/terms-of-use.
9. L. Yosifova. Privacy risks in ai-powered security frameworks. *ACM Journal of Data Ethics, 12*(3):155–172, 2024.