Ethical and Regulatory Challenges in Managing AI-Centric Cybersecurity Programs for Critical Infrastructure

(Author Details)
Kumar Saurabh
PMI, USA
Email: ksaurabh.pm@gmail.com

Abstract

Artificial intelligence (AI) has become a prevalent feature of modern cybersecurity programs for critical infrastructure systems such as those used in energy, transportation, healthcare, and industrial control systems. AI-based security solutions can provide more accurate detection, adaptive response, and scalability than traditional rule-based systems. However, the use of these systems in safety-critical and highly regulated sectors of the economy raises significant ethical and regulatory issues which present additional challenges to their effective management. These include risks of adversarial machine learning, explainability limitations of complex models, accountability gaps for autonomous decisions, and the challenge of meeting regulatory requirements for cybersecurity practices and data privacy in AI-driven operations.

This paper presents a thorough, literature-based exploration of the ethical and regulatory challenges of program management for AI-focused cybersecurity programs for critical infrastructure. Building on existing research in areas such as adversarial machine learning, ethical AI, and cybersecurity for critical infrastructure, this work takes a program-level perspective to understand how technical, ethical, and regulatory risks manifest and intersect at various stages of the AI lifecycle. By reviewing existing frameworks, standards, and principles in these areas, this study identifies the structural limitations of traditional cybersecurity program governance frameworks when applied to AI-based solutions.

By conceptually organizing these issues along several dimensions of responsible AI - namely robustness, transparency, accountability, and regulatory compliance - this study aims to provide an analytically coherent foundation to understand responsible AI in the context of critical infrastructure cybersecurity. The insights presented highlight opportunities for holistic governance approaches that balance innovation, ethics, and legal considerations, while supporting sustainable, long-term trust in AI-enabled cybersecurity solutions.

Keywords: AI-centric cybersecurity, critical infrastructure protection, ethical AI governance, regulatory compliance, adversarial machine learning, cybersecurity program management.

DOI: 10.21590/ijtmh.10.04.20

1. Introduction

1.1 Context: AI in Cybersecurity for Critical Infrastructure

Industrial sectors, critical infrastructure, and enterprises increasingly rely on networked computers and automated systems to ensure their reliability, efficiency, and availability to the customers. As these digitalized assets continue to grow in size and complexity, their operational activities generate petabytes of network traffic, sensor data, and execution logs per day, creating an information volume that can overwhelm traditional security monitoring tools. To bridge this gap, cybersecurity programs increasingly adopt artificial intelligence to enhance detection, analysis, automation, and orchestration. Machine learning (ML) models, a primary branch of AI, are in particular ubiquitously leveraged to enable, e.g., intrusion detection, anomaly detection, malware classification, or behavior analysis, as they can learn complex representations that are hard to capture with rule-based techniques (Buczak & Guven, 2015; Khraisat et al., 2019). The integration of AI has been especially prevalent in the field of securing industrial control systems and cyber-physical systems in general. This is due to the fact that protecting systems like supervisory control and data acquisition systems, smart grids, or automated manufacturing from cybersecurity incidents is a prerequisite for ensuring that no cascading effects or physical damages result from those incidents. Furthermore, these applications differ from “standard” enterprise IT systems due to stricter availability requirements, legacy systems, or safety-critical operations. It has been shown in prior work that ML methods are capable of detecting anomalies that are too subtle to be discerned by humans or that do not match common attack scenarios, yet still indicate an attacker having potentially compromised or changed the behavior of an industrial system (Knowles et al., 2015; Humayed et al., 2017).

At the same time, the increasing adoption of AI for cybersecurity also exposes key vulnerabilities in the previous state of affairs, which was mainly based on hard-coded security rules and signatures. The key point is that many traditional detection methods, e.g., signature-based intrusion detection systems, have largely been built around the assumption of a largely stable threat model and well-defined attack patterns. In reality, cyber adversaries are much more dynamic in that their tactics, techniques, and procedures are known to be very heterogeneous and can adapt in near real-time to different contexts, rendering traditional rules highly ineffective. Sommer and Paxson (2010) show, e.g., that learning systems operating in the real world need to address concepts such as concept drift, data imbalance, and adversarial learning, which are not typically seen in “conventional” information security. This has also been expressed by other researchers in the sense that probabilistic and adaptive AI/ML systems replace previous rule-based defenses and deterministic security models (Nguyen et al., 2015).

1.2 Context: Ethical and Regulatory Concerns of AI in Cybersecurity

In addition to the limitations of traditional security systems, AI also opens the door for a wide range of non-technical risks, such as ethical and regulatory risks, which are not captured by purely functional performance measures, e.g., F1-score, precision, or recall. One prominent ethical concern of machine learning is its apparent "black box" nature, i.e., that AI models cannot readily be explained or interpreted by humans. This severely impacts critical infrastructure settings, as it weakens accountability, reduces human operators' trust, and hampers post-breach forensics in safety-critical systems in case of false alarms or, even worse, not raising the alarm at all (Guidotti et al., 2018). Here, the ethical consideration is not just one of reliability and explainability in general, but one of the responsibility in contexts where AI systems are autonomously taking decisions to flag or counter threats, yet the root cause of a false alert cannot be explained in a transparent way.

Another key issue of AI and ML for cybersecurity is that systems built with these technologies are inherently vulnerable to so-called adversarial machine learning (AML) attacks, which are already shown early on in seminal work (Barreno et al., 2006). In brief, these attacks can make use of intentional poisoning of the training data or can evade detection entirely by exploiting unknown weaknesses or prediction patterns of the ML model. Following work has shown this risk to be not only inherent but, also, very actively exploited in security-sensitive applications in the wild, such as software vulnerability detection, antivirus systems, or malware scanners (Papernot et al., 2018). Given the safety-critical nature of critical infrastructure and its services, this too can lead to significant and even potentially existential risks in the form of service disruptions, physical damage, or harm to human lives.

On the regulatory side, the growing use of AI also places cybersecurity programs for critical infrastructure in the field of tension of increasingly stringent data protection requirements. As data protection regulations cover the processing of personal data and, in the EU in particular, special categories of personal data which are in many cases relevant to security-sensitive systems, compliance with data protection obligations, such as lawful processing or purpose limitation, is often highly challenging for AI-centric cybersecurity architectures (Regulation, 2016). The same also applies, to a certain degree, to the cybersecurity regulatory landscape, where standard-setting and specific regulation typically expect critical infrastructure operators to adhere to minimum cybersecurity standards and follow risk management, resilience, and accountability requirements in their operations. While the NIST Cybersecurity Framework, for example, offers a very structured and comprehensive overview of these topics, it remains agnostic on many AI-related aspects and thus does not provide holistic guidance for AI-specific risks and concerns (Cybersecurity, 2018).

1.3 Scope and Contributions

The purpose of this work is to provide a comprehensive, high-level treatment of ethical and regulatory challenges that are relevant to AI-centric cybersecurity programs for critical infrastructure. This is a program- and governance-level framing that views AI-augmented cybersecurity as a managed organizational capability, which also consists of other organizational elements than the technology alone. In that sense, it will not focus on the deep technical specifics of particular AI/ML approaches for security, such as detection models, but will instead concentrate on ethical, legal, and regulatory topics that impact the adoption, integration, and operations of AI as a whole within the organization. The general objective of this research is then to leverage the rich body of existing literature on adversarial ML, ethical AI, and critical infrastructure cybersecurity to synthesize an integrated analytical perspective on the key challenges that these domains raise for AI governance.

In order to meet this objective, the paper has three main contributions. First, it surveys existing research on ethical and adversarial risks, data protection issues, and cybersecurity regulatory constraints. While much of this research has been conducted independently in each of the three aforementioned fields, this work provides a framework for understanding and categorizing the main issues and challenges in the context of critical infrastructure. Second, this paper also emphasizes the importance of a strong program-level governance for ethically aligned and legally compliant AI-driven cybersecurity operations. Third, it shows how a principled integration of well-established cybersecurity standards and ethical AI guidelines can help organizations practically achieve such responsible and resilient AI adoption in critical infrastructure contexts.

2. AI-Centric Cybersecurity in Critical Infrastructure Systems

Artificial intelligence (AI) has become an integral part of cybersecurity operations, augmenting human analysts in the detection, analysis, and response to cyber threats. In contrast to traditional security mechanisms, which often rely on manually-defined rules and signatures, AI-based solutions are data-driven, learning to detect complex attack patterns and previously unseen threats through exposure to large volumes of network traffic, malware samples, and security logs. In critical infrastructure settings, where resilience and safety are paramount, such capabilities are increasingly seen as necessary. However, their adoption also comes with technical, ethical, and operational risks that need to be managed carefully.

2.1 Machine Learning Techniques for Cyber Defense

Machine learning (ML) techniques are essential for modern cyber defense, underpinning many intrusion detection and network monitoring systems. Supervised learning techniques are based

on labeled training data that inform the classifier how to discriminate between normal and malicious behavior. Supervised methods have been employed extensively in intrusion detection systems, where they are used to detect known attack patterns with high accuracy (Buczak & Guven, 2015). Popular supervised learning techniques include decision trees, support vector machines, and neural networks. They can process large volumes of network traffic data and learn to recognize subtle indicators of malicious activity. However, supervised ML models require representative and labeled training data, which may be difficult to obtain in dynamic threat environments (Khraisat et al., 2019).

Unsupervised learning methods are particularly well-suited to anomaly detection in cybersecurity, as they do not require labeled data to identify suspicious activity. Clustering, density estimation, and autoencoders are some of the common unsupervised ML approaches for anomaly detection in network data. They are used to construct a model of normal behavior and to detect deviations, which are flagged as anomalies for further investigation. Unsupervised learning is valuable in critical infrastructure contexts, where novel or targeted attacks may not have signatures or are difficult to label (Buczak & Guven, 2015). However, since there are no explicit labels, unsupervised methods may have higher false positive rates, which could overwhelm analysts and impede operations.

Predictive analytics is another important machine learning application in cybersecurity. By leveraging historical incident data, system logs, threat intelligence feeds, and other sources of information, predictive models can forecast future attacks, estimate their likelihood, and identify high-risk assets. In network monitoring, anomaly detection algorithms are used to discover deviations from expected traffic patterns, device behavior, and system performance. These predictive capabilities are especially important in critical infrastructure settings, where networks are often large-scale and distributed. In such systems, manual monitoring and analysis are infeasible and delayed detection can lead to catastrophic failures (Khraisat et al., 2019). However, predictive models raise ethical concerns around transparency and accountability, as decisions are increasingly based on automated risk assessments.

2.2 Operational Characteristics of Critical Infrastructure

AI-centric cybersecurity solutions must be evaluated in the context of the infrastructure on which they are deployed. Critical infrastructure environments differ from conventional enterprise IT in several ways, and those differences have a direct impact on the performance of AI algorithms and tools. One of the most important features is that control systems and cyber-physical systems (CPS) often have strict real-time constraints. In industrial control systems (ICS) and CPS, deterministic response times are often required for safe and stable operation (Stouffer et al., 2011). Cybersecurity functions that introduce latency or other forms of unpredictable behavior may interfere with control processes and result in physical damage or safety incidents.

Safety requirements are another important factor that must be considered when applying AI tools for critical infrastructure security. In many critical infrastructure sectors, such as energy, transportation, and healthcare, cybersecurity failures can have direct consequences for human safety and welfare. In these contexts, cybersecurity decisions must be conservative, preferring reliability and fail-safe behavior over more aggressive automated responses. This requirement is often at odds with ML systems, which can change their behavior dynamically in response to new data and may not behave predictably in adversarial settings. Many critical infrastructure environments are also characterized by legacy systems that are not designed with cybersecurity in mind. They may have limited computational capacity, use proprietary or outdated protocols, and have lifecycles measured in decades (Stouffer et al., 2011). These factors constrain the application of complex ML models and complicate processes for updating, retraining, and validating models.

Finally, the cost of failure for cybersecurity in critical infrastructure settings is much higher than for traditional IT systems. Disruptions or malfunctions in power grids, water systems, and transportation networks can have cascading effects, resulting in economic losses, environmental damage, and national security risks. As a result, both the ethical and the security risks of AI-centric cybersecurity systems are magnified. Model errors, biased training data, or adversarial manipulation may lead not only to loss of detection performance but also to loss of confidence in automated decision support. This operational reality means that AI techniques and tools must be carefully aligned with infrastructure constraints and governance.

Table 1: AI Techniques Used in Critical Infrastructure Cybersecurity and Operational Risks

AI Technique	Cybersecurity Function	Infrastructure Context	Ethical and Security Risks	Key References
Supervised Learning	Intrusion detection, malware classification	Industrial control systems, enterprise networks	Dependence on labeled data, susceptibility to adversarial manipulation	Buczak & Guven (2015); Khraisat et al. (2019)
Unsupervised Learning	Anomaly detection, zero-day attack identification	Cyber-physical systems, sensor networks	High false positives, limited explainability	Buczak & Guven (2015)
Predictive Analytics	Threat forecasting, risk prioritization	Large-scale critical infrastructure networks	Opacity in risk scoring, accountability challenges	Khraisat et al. (2019)
Anomaly Detection Models	Continuous network and system monitoring	Real-time operational environments	Latency risks, model drift over time	Stouffer et al. (2011)

3. Adversarial Risks and Security Limitations of AI Models

The development of AI-based cybersecurity models has significantly improved accuracy and automation of AI-supported detection of intrusions, which is particularly complex in large-scale and high-value settings such as critical infrastructure. Prior work, however, has shown that ML-based security models operate in the most fundamental adversarial settings and are thus subject to adversarial risk. Adversaries have a strong incentive to manipulate the learning process and/or outputs, which then leads to a structural limitation on the security, trustworthiness, and longer-term effectiveness of AI-centric cybersecurity programs.

3.1 The Nature of Adversarial ML in Cybersecurity

Adversarial machine learning is a process of data, model, or decision boundary manipulation with the aim to subvert system performance or to be undetected by a system. Dalvi et al. (2004) were among the first to show that classifiers trained and deployed in adversarial settings are vulnerable to adversarial manipulation, especially when the attacker has partial knowledge of the learning process. The cybersecurity setting is unique in this context, as the attacker can iteratively probe the system and adapt behavior based on system feedback.

Poisoning attacks are some of the most impactful attack types on AI-based cybersecurity models. The focus of poisoning attacks is the training phase of the learning pipeline, during which an attacker injects malicious samples to contaminate the training set and poison the resulting model. Barreno et al. (2006) have shown that even limited poisoning can lead to a significant shift in decision boundaries which result in systematic misclassification of malicious behavior as normal behavior. In a critical infrastructure setting, retraining of a model may be partially or fully automated based on the constant stream of data, meaning a poisoning attack can remain undetected for a long time and gradually degrade the detection performance of the system.

Evasion attacks, in contrast, are performed during the inference phase of model deployment. In these attacks, malicious inputs are modified to avoid detection, without changing the intended goal of the attack. Adversaries craft inputs that are close to the learned decision boundaries of the classifier. Evasion attacks are especially powerful against models that are static or otherwise insufficiently monitored, which allow adversaries to pass intrusions undetected by cybersecurity defenses while maintaining stealth of their operations (Dalvi et al., 2004; Barreno et al., 2006).

Poisoning and evasion attacks, however, are a simplification of the form of adversarial behavior that is found in real-world cybersecurity settings, as attackers are adaptive. Biggio and Roli (2018) point out that attackers are constantly changing their tactics in response to new defensive measures, meaning that there is a continuously changing threat landscape. This adaptive nature of attack behavior stands in contrast to security assumptions that a model operates against a static or stationary distribution of data. In practice, attackers may also use a mix of attack types, use

feedback to maximize impact, and leverage long-term reconnaissance to better position their operations, all of which can expose vulnerabilities in AI-based cybersecurity defenses.

3.2 The Reality of Robustness Failures in AI-Based Security Systems

The adversarial vulnerability of AI models is also closely related to the concept of model robustness. Goodfellow et al. (2014) showed that deep learning models are extremely sensitive to small, imperceptible, but carefully crafted input perturbations. Adversarial perturbations are input modifications that lead a classifier to confidently make the wrong prediction, even when the input data looks legitimate. In the context of cybersecurity systems, this sensitivity can result in false negatives, delayed reaction times, or reduced situational awareness.

Robustness failures can have a high impact in a critical infrastructure context, as a failure to correctly classify an input can have physical and societal impact through the control system(s) on which the cybersecurity system is operating. Deep neural networks, for instance, use high-dimensional representations of input features, which can amplify small changes to input data and make them susceptible to adversarial manipulation. Goodfellow et al. (2014) point out that this vulnerability is not a property of a particular network architecture, but a more general property of many other linear and non-linear models in use.

The ability to evaluate the robustness of AI models is further limited in adversarial settings. Carlini and Wagner (2017) have shown that proposed defenses against robustness attacks have been largely superficial and ineffective under stronger or adaptive attack models. Empirically evaluating model robustness also remains a difficult task, with the lack of standardized threat models and evaluation metrics, and a tendency towards security by obscurity. These challenges limit the generalizability of reported improvements in robustness in the research setting.

AI-based security systems in cybersecurity programs, however, are often only tested under non-adversarial settings, and therefore suffer from a lack of adversary-aware evaluation. This can leave a false sense of security where a model is seemingly robust under limited testing, but has hidden weaknesses that can be exploited in a real-world adversarial context. The lack of widely accepted protocols for robustness evaluation in security-critical domains also complicates the effective and transparent use of AI models in terms of trust in automated defenses, regulatory compliance, and certification.

Figure 1 presents a bar chart illustrating the relative performance degradation of AI-based cybersecurity systems under different types of adversarial attacks.

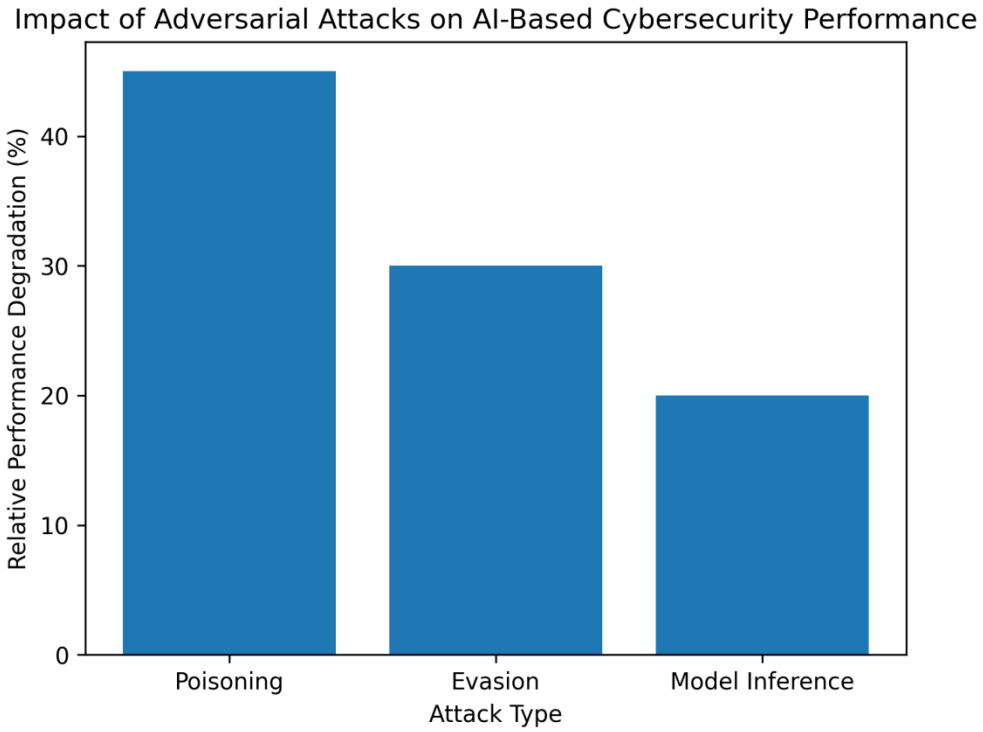


Figure 1: Impact of Adversarial Attacks on AI-Based Cybersecurity Performance

The graph conceptually demonstrates that poisoning attacks typically cause the most severe long-term degradation by corrupting training processes, while evasion attacks lead to immediate but often less persistent performance losses. Model inference attacks show moderate degradation by enabling adversaries to extract system behavior and refine future attacks. This visualization reinforces the cumulative impact of adversarial threats on AI-centric cybersecurity performance.

4. Ethical Challenges in AI-Centric Cybersecurity Programs

Integrating AI into cybersecurity programs for critical infrastructure introduces a range of ethical issues that go beyond traditional technical risk management. AI-driven cybersecurity solutions are playing an increasingly decisive role in high-consequence decisions, such as threat classification, response, containment, and even system shutdown. In safety-critical and socio-economically essential critical infrastructure systems, ethical failures can have ripple effects, jeopardizing public safety, economic stability, and trust in digital systems. The three interrelated ethical challenge areas discussed in this essay are transparency and explainability, accountability and human control, and the lack of ethical AI frameworks and documentation.

4.1 Transparency and Explainability

AI-powered cybersecurity systems typically employ sophisticated ML models which function as black boxes, making security decisions without providing meaningful explanations to the human

operators. The resulting "black-box" problem is a critical ethical issue in the context of securing critical infrastructure, as stakeholders such as operators, regulators, and auditors may have the need or even legal obligation to justify and account for security decisions and system behavior. For example, prior work has shown that many state-of-the-art ML models are inherently uninterpretable, and that complex input features like network traffic or system logs cannot be directly connected to a specific classification (Guidotti et al., 2018).

In a cybersecurity setting, a lack of explainability also limits transparency and accountability as well as forensic analysis after an incident. When an AI system autonomously identifies a suspicious event or anomaly or automatically takes a defensive action, the human operators may be unable to verify whether this decision was based on an actual threat signal or an artifact of data noise and model bias. The lack of explainability further hinders detection and correction of errors that occur during adversarial attacks, as the attackers actively probe the system to remain undetected while manipulating the system into making errors.

Approaches to explainable artificial intelligence (XAI) have been developed to address these ethical concerns by providing human-understandable explanations of the ML model decisions after the fact. Explainability methods, such as local surrogate models or feature attributions, seek to make individual model predictions more interpretable to human stakeholders (Ribeiro et al., 2016). In an AI-centric cybersecurity program, explainability can help to address some of the ethical concerns around cybersecurity automation by supporting human validation of the system decisions by the operators, the regulatory evaluation of the AI solutions, and organizational justification of automated security decisions. However, explainability methods also come with tradeoffs between accuracy, fidelity to the original model, computational complexity, and latency, which must be carefully considered in the high-speed setting of cyber operations.

4.2 Accountability and Human Control

AI-based cybersecurity systems which make high-consequence decisions also raise ethical issues around human control and accountability. Autonomous cyber defense mechanisms, such as the automated blocking of network traffic or system isolation in response to a perceived threat, can cause inadvertent harm if not overseen and controlled by humans. In critical infrastructure systems, an incorrect automated decision or containment action can interrupt vital services, introduce safety hazards, or cause cascading failures across infrastructure systems.

Issues around ethical accountability become especially complicated when responsibility is shared across many actors, such as the data providers, model developers, system integrators, and system operators. In the absence of clearly defined accountability and governance, it can become difficult to determine the responsible actor for harmful security actions taken by the AI system. This is further complicated in the presence of adversaries who may intentionally elicit an automated response from the system with the goal of disrupting the infrastructure service.

In the AI ethics literature, many frameworks call for meaningful human control to be a prerequisite for high-risk AI systems, especially in safety-critical applications (Floridi et al., 2018). Human-in-the-loop (HITL) system governance mechanisms are one possible approach to ensure that AI systems are not operating without ethical oversight. HITL approaches typically involve expert review, escalation paths, and judgment from a human supervisor before an automated system response is executed. In a cybersecurity program, such human supervision can provide ethical oversight by allowing the human experts to evaluate the response or containment plan suggested by the AI system and make the final go/no-go decision, combining the efficiency of automated processing with human responsibility to avoid uncontrolled automated harm.

4.3 Ethical AI Frameworks and Documentation

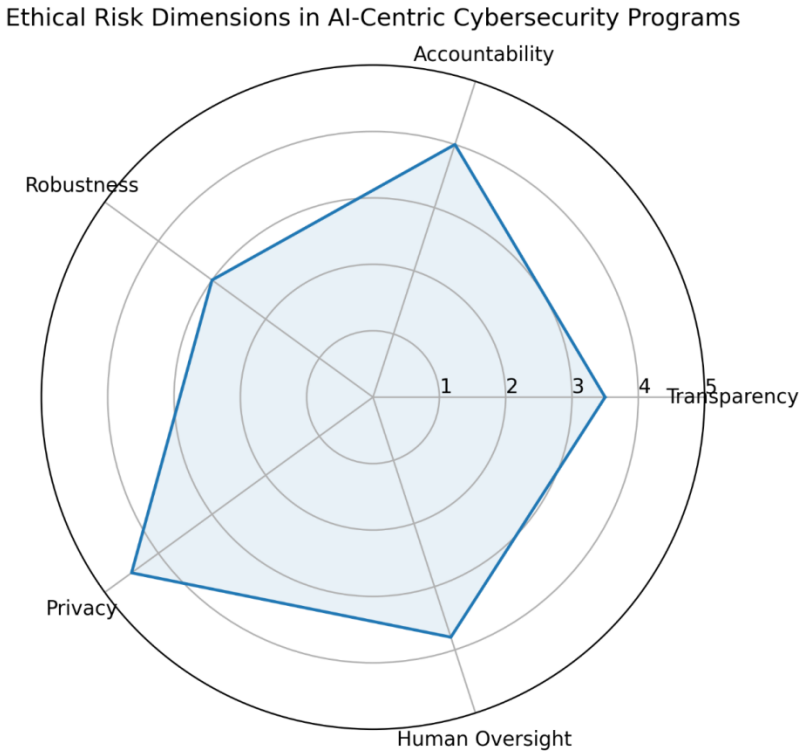
Beyond the need for transparency and human control, ethical management of AI-centric cybersecurity programs also needs to align with ethical AI frameworks and documentation practices. Ethical principles for responsible AI developed in multiple fields call for a range of foundational principles, including fairness, accountability, transparency, and respect for human values (Floridi et al., 2018; Jobin et al., 2019). For example, the IEEE ethical AI principles for example provide a non-exhaustive list of foundational requirements, many of which apply directly to ethical cybersecurity, such as “Avoid Algorithmic Bias” or “Respect Privacy.” These principles are of particular relevance in cybersecurity for critical infrastructure, as security decisions can impact the whole population of a region or essential services for a nation.

One practical challenge in operationalizing ethical AI principles in cybersecurity programs is the lack of standardized model documentation practices for AI models. With a lack of standard or even ad hoc documentation, it is not possible to fully understand and account for model assumptions, intended use cases, limitations, and known ethical risks of the AI model. A number of model documentation practices have been suggested to address this need, ranging from structured reporting templates to good documentation practices (Mitchell et al., 2019). In the context of cybersecurity programs, model documentation supports ethical due diligence and risk management by providing operators and stakeholders with essential information about the model design, training data, performance characteristics, and intended operational use case.

Incorporating ethical AI frameworks and documentation practices into cybersecurity programs can help ensure that the cybersecurity systems are aligned with ethical AI practices beyond ad hoc risk management. This also ties ethical responsibility to established program-level management and regulatory practices.

Figure 2. Ethical Risk Dimensions in AI-Centric Cybersecurity Programs

Figure 2 (Graph): Radar Chart of Ethical Risk Dimensions in AI-Centric Cybersecurity Programs



The radar chart illustrates the relative prominence of ethical risk dimensions in AI-centric cybersecurity programs for critical infrastructure. Each axis represents a core ethical dimension identified in prior literature. Higher values indicate greater ethical risk exposure. The visualization supports comparative analysis of ethical vulnerabilities, highlighting imbalances between technical robustness and governance-oriented controls such as transparency and human oversight.

5. Regulatory and Legal Challenges

Regulatory and legal considerations are important to discuss in the context of this paper. Cybersecurity programs for critical infrastructure have long been subject to a variety of standards and laws, but AI introduces new dimensions to compliance. Most security tools use data collection and preconfigured rules, while AI-centric systems often involve real-time data streams, automated decision-making, and dynamic learning. These elements can interact with data protection, accountability, transparency, and cybersecurity governance requirements in complex ways. Ensuring regulatory compliance is therefore an ongoing activity integrated into AI systems life cycle management.

5.1 Data Protection and Privacy Regulations

AI-powered cybersecurity systems function through pervasive monitoring, data aggregation, and real-time analytics. In the context of critical infrastructure, this typically means processing large volumes of network traffic, system logs, user behavior, and operation telemetry. The scale and invasiveness of this data processing can lead to privacy concerns and potential violations of data protection laws, especially when sensitive or personal data is involved. Intrusion detection and anomaly detection based on machine learning often need access to sensitive data that may include personally identifiable information, metadata, or even indirectly derivable attributes that may fall under regulatory purview.

Automated decision-making represents another key regulatory issue. Security systems are increasingly expected to autonomously classify and respond to perceived threats, often without human intervention. This could involve automatically blocking access, denying services, or revoking user privileges. Automated actions like this have to be proportional and limited in scope, especially when data is processed in ways that users may not have explicitly consented to. The lack of transparency in many AI models further exacerbates these concerns, as it may be difficult to understand how a specific data point contributed to an alert or other automated security decision.

Regulatory requirements based on data protection laws revolve around transparency, data minimization, accountability, and user rights. Privacy considerations in AI-based cybersecurity therefore require clear limitations on what data is collected and how it is retained, used, and shared with other organizations. Data should not be used for secondary purposes that have not been previously authorized or regulated, and collection and retention policies should be clearly defined and enforced. The challenge is amplified by the dynamic nature of AI models, where a learning system may change its behavior over time in ways not initially accounted for during a compliance assessment.

Auditability and governance requirements are also linked to data protection regulations. It is necessary to have detailed records that allow security operators to document how data is used, how models are trained and updated, and how automated decisions are justified. This can be challenging in critical infrastructure environments where requirements around operational continuity, system availability, and real-time processing often take priority over other organizational concerns. Managing privacy and data protection in AI-powered cybersecurity programs for critical infrastructure therefore requires a coordinated approach that balances legal, technical, and organizational factors (Regulation, 2016).

5.2 Cybersecurity Standards and Frameworks

Cybersecurity standards and frameworks, which are often a critical part of regulatory requirements, also play a role in shaping AI adoption. These frameworks and standards often

provide structured approaches for implementing risk management in critical infrastructure, but they were not originally designed with adaptive AI models in mind. This creates a gap in which AI-specific risks and issues must be understood and then mapped onto existing compliance frameworks.

The National Institute of Standards and Technology (NIST) Cybersecurity Framework provides a common risk-based approach to improving cybersecurity for critical infrastructure. It is based on core functions – identify, protect, detect, respond, and recover – that are relevant to AI-enabled security operations. However, incorporating AI systems in a NIST-compliant manner requires additional governance considerations, including model validation, adversarial robustness, and life cycle considerations. AI-driven detection, for example, may improve the detect function of the framework, but its use also introduces false positive rates, model drift, and explainability risks that are not explicitly considered in current frameworks (Cybersecurity, 2018).

Industrial control systems (ICS) have additional regulatory considerations. ICS are often safety-critical, have longer system life cycles, and have limited resilience to disruption. Regulations and guidance documents specific to ICS security focus heavily on system availability, deterministic behavior, and risk containment. As a result, deploying AI-enabled cybersecurity tools in an ICS environment can raise questions about predictability and control, as well as traditional certification and compliance expectations. Adaptive AI models may have variable behavior over time that is at odds with regulatory expectations about system stability and verification in safety-critical environments (Stouffer et al., 2011).

The main governance challenge in the context of AI and regulation is in reconciling the adaptive behavior of AI systems with an existing compliance regime that does not necessarily account for such behavior. This implies extending cybersecurity governance practices to encompass AI-specific controls, such as model performance monitoring, periodic risk re-evaluation, and formalized documentation of training datasets and decision logic. Compliance with existing cybersecurity standards and frameworks will not be meaningful if the additional considerations around AI systems are not captured.

Table 2. Regulatory and Standards Landscape for AI-Centric Cybersecurity

Regulation or Standard	**Scope**	**AI Relevance**	**Key Compliance Challenges**
Data Protection Regulation (EU) 2016/679	Personal data protection and privacy	Governs data collection, processing, and automated decision-making in AI-based security systems	Ensuring transparency, lawful processing, auditability, and explainability of AI-driven security

			decisions
NIST Cybersecurity Framework	Critical infrastructure cybersecurity risk management	Provides structure for AI-enabled detection, response, and recovery functions	Integrating AI lifecycle governance, managing model drift, and addressing adversarial risks
ICS Security Guidelines	Industrial control system protection	Regulates cybersecurity in safety-critical and operational technology environments	Ensuring predictability, stability, and certification of adaptive AI security mechanisms
Organizational Cybersecurity Policies	Enterprise-level governance and compliance	Defines internal oversight for AI deployment and monitoring	Aligning technical AI controls with legal and regulatory accountability requirements

6. Program-Level Governance and Management Challenges

The technical aspects of AI implementation in cybersecurity programs for critical infrastructure must be complemented with program-level governance considerations. AI-driven systems are not just another technical control; their learning, adaptive, and evolving nature leads to dynamic and potentially unpredictable behaviors. This evolution presents additional complexity to governance, risk management, and regulatory compliance efforts, especially in safety-critical or heavily regulated infrastructure sectors (Knowles et al., 2015).

Program-level governance of AI in cybersecurity encompasses aspects related to the management of AI systems across their operational lifecycle, from development and deployment to decommissioning, as well as the identification, assessment, and mitigation of risks specific to AI technologies. It also includes the ethical and regulatory compliance aspects within the broader cybersecurity management framework. Neglecting these program-level considerations can lead to unanticipated and uncontrolled model behaviors, potential non-compliance with regulations, and increased exposure to systemic risks.

6.1 AI Lifecycle Management

AI-driven cybersecurity solutions have their own lifecycle, which involves data acquisition, model training, deployment, monitoring, updating, and eventual decommissioning. Many AI

models, in contrast to conventional security systems, are designed to learn continuously from new data, which introduces the concept of model drift, where the model's performance and decision boundaries change over time. In the adversarial context of cybersecurity, this drift could also be a result of strategic alteration of input data by an attacker (Barreno et al., 2006; Papernot et al., 2018).

Continuous model learning poses challenges to governance, as the behavior of the system may deviate from its originally validated state over time. In the absence of structured lifecycle management and oversight, AI-based threat detection or response systems may, over time, start to violate internal security policies, ethical standards, or even legal regulations. This risk is magnified in critical infrastructure domains, where unforeseen actions by the system could have serious operational and safety repercussions (Stouffer et al., 2011).

Lifecycle management, thus, becomes critical and must include governance mechanisms for ongoing model oversight. This goes beyond the initial model testing and validation to include regular performance audits, controlled update and tuning procedures, and clear ownership for model changes and updates. Industrial control systems cybersecurity management emphasizes structured change management, thorough documentation, and clear delineation of responsibilities and roles, principles that are equally relevant to AI lifecycle governance (Knowles et al., 2015). Integrating AI system oversight into existing cybersecurity management processes ensures that AI solutions are subject to the same rigor in terms of risk analysis, approval, and change management as other critical security controls.

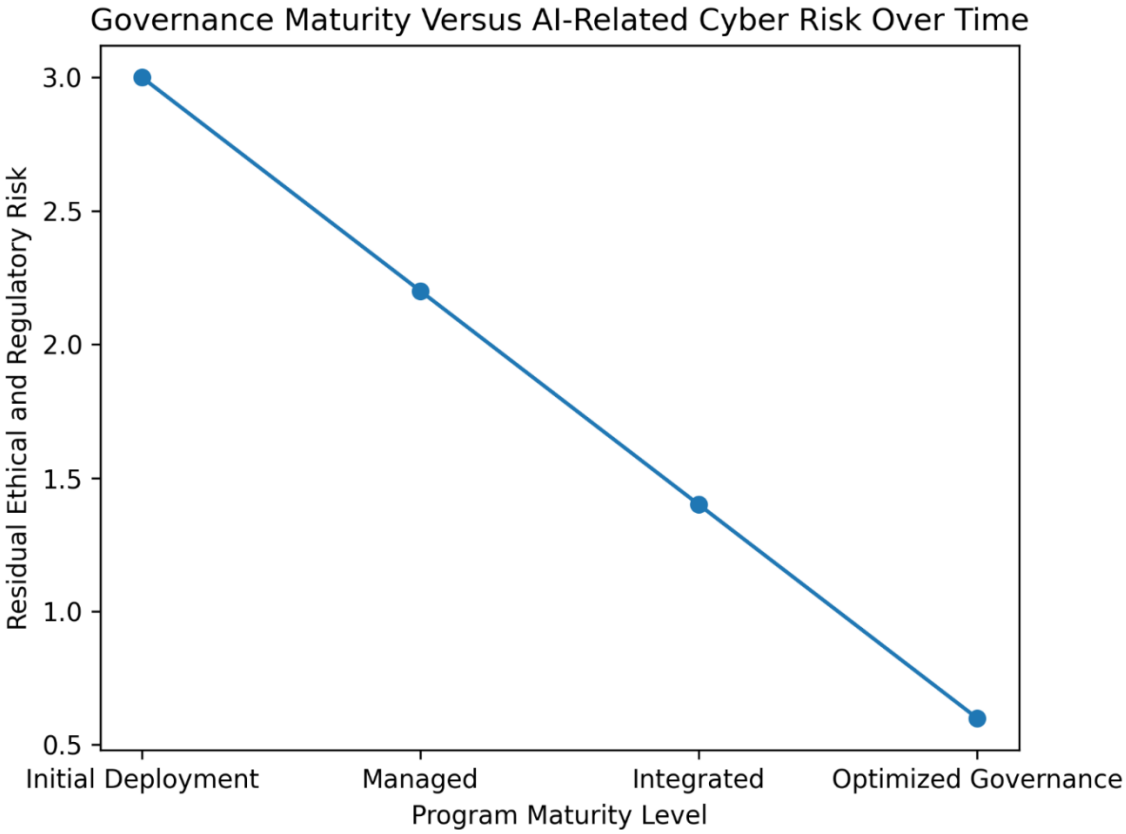
6.2 Risk Management and Regulatory Compliance

AI-enabled cybersecurity programs also have to balance the considerations arising from operating in adversarial and highly regulated settings. Research in adversarial machine learning highlights the potential for attackers to deliberately exploit the learning algorithms through various techniques, including poisoning, evasion, and model inversion, thereby directly challenging the system's integrity (Dalvi et al., 2004; Goodfellow et al., 2014; Carlini & Wagner, 2017). Parallelly, regulatory and compliance mandates dictate that cybersecurity systems operate in a transparent, lawful, and safe manner, especially when handling personally identifiable or sensitive data (Regulation, 2016).

Mitigating risks from adversarial attacks and ensuring regulatory compliance are interrelated governance challenges. Defensive strategies like overly aggressive automated response mechanisms or non-transparent detection processes can, on the one hand, enhance security performance against model-attack techniques but may also run counter to principles of accountability and explainability (Floridi et al., 2018; Guidotti et al., 2018). Papernot et al. (2018) specifically stress that robust machine learning systems should not be optimized for security in isolation but must also respect privacy, transparency, and governance considerations.

Program-level risk management of AI systems, thus, needs to be integrative, combining the technical security measures with compliance and ethical oversight. Ethical review, legal compliance verification, and human-in-the-loop processes should be embedded within AI-driven cybersecurity operations. Automation should not be a replacement for human oversight, more so in the critical infrastructure domain where the demands for accountability and traceability are higher. The human-in-the-loop concept needs to balance automated decision-making with safety and ethical responsibilities, with governance structures explicitly defining acceptable risk levels, escalation procedures, and mechanisms for regulatory reporting.

Figure 3. Governance Maturity Versus AI-Related Cyber Risk Over Time



The line graph illustrates an inverse relationship between cybersecurity program governance maturity and residual ethical and regulatory risk associated with AI-centric systems. At early maturity stages, limited oversight, insufficient lifecycle controls, and fragmented compliance integration result in elevated risk exposure. As governance structures mature through formal lifecycle management, integrated risk oversight, and regulatory alignment, residual ethical and regulatory risk declines. The graph emphasizes that technical robustness alone is insufficient and that sustained risk reduction depends on program-level governance evolution.

7. Discussion

7.1 How adversarial robustness, ethical governance and regulatory compliance fit together

The literature reviewed in this project has been synthesized in to a number of key insights. The primary insight is that adversarial robustness, ethical governance and regulatory compliance are interwoven, fundamental aspects of AI-centric cybersecurity programs, rather than separate or siloed areas of concern. Technical evidence from the adversarial machine learning literature has shown that AI-based cybersecurity systems are subject to evasion, poisoning, and manipulation attacks that can degrade their detection accuracy and reliability (Dalvi et al., 2004; Barreno et al., 2006; Biggio & Roli, 2018). These technical vulnerabilities have direct ethical and regulatory implications when these systems are used in critical infrastructure protection, as false positives or manipulated decisions can have real-world consequences for safety and society (Stouffer et al., 2011; Humayed et al., 2017).

Ethical AI governance principles, on the other hand, explicitly call for robustness, explainability, accountability, and human oversight, among others, as key requirements for trustworthy AI deployment (Floridi et al., 2018; Jobin et al., 2019). However, the problem of adversarial fragility in AI models raises challenges to these ethical principles, by demonstrating that technically insecure models can easily produce misleading, opaque, and unaccountable outputs (Moramarco et al., 2019; Zisselman, 2019). In other words, adversarial robustness is not just a technical criterion, but also an ethical imperative for AI systems, because weakly robust systems could be considered irresponsible or negligent in their protection of critical infrastructures.

The regulatory dimension of cybersecurity further complicates this relationship. The existing data protection and cybersecurity regulations and standards implicitly assume that the systems in question function reliably, transparently, and within an acceptable level of risk tolerance (Regulation, 2016; Cybersecurity, 2018). When AI models used in security settings are susceptible to adversarial attacks, or simply lack explainability and transparency in their operations, it becomes very difficult to prove compliance with many accountability, auditability, and risk management requirements. In this way, the technical robustness, ethical alignment, and regulatory compliance of AI-based cybersecurity systems are mutually dependent on one another.

7.2 Identified structural gaps in AI-specific cybersecurity management programs

The literature analysis conducted for this project also enabled identification of certain structural gaps in existing cybersecurity management frameworks for AI-driven systems. These frameworks were developed based on deterministic systems, static threat models, and rule-based

controls (Sommer & Paxson, 2010; Knowles et al., 2015). AI-centric cybersecurity, by contrast, is more adaptive, probabilistic, and subject to continuous model updates and learning, which means that its security governance needs and challenges may not be fully addressed or understood by current standards.

The first gap relates to the lifecycle of AI-driven cybersecurity systems. Traditional cybersecurity frameworks focus on system deployment and runtime monitoring, but leave much of the continuous learning, model drift, and model retraining under-governed (Papernot et al., 2018). While the training and operational phases of AI model lifecycle may be well defined, the processes that occur in between, such as retraining and validation of models, are not as explicitly governed. This gap in lifecycle management, in turn, leaves room for ethical and regulatory risks to materialize and accumulate unobserved over time.

The second gap is that of explainability and accountability of decisions. Ethical AI research is by now familiar with the notions of interpretable and explainable AI, which lay emphasis on making automated decisions understandable and justifiable (Ribeiro et al., 2016; Guidotti et al., 2018). However, the current landscape of cybersecurity management practice has few mandates or expectations of explainability for cybersecurity decisions, be they detection or response in nature. As such, the gap between research on the ethics of AI explainability and current cybersecurity governance practice complicates the verification of compliance, incident investigation and reporting, and stakeholder trust, especially in regulated critical infrastructure settings.

Finally, and most critically, the issue of adversarial risk management is not sufficiently integrated at the program level. The existing cybersecurity management frameworks often only address adversarial risks in passing, or at most as a program-level risk, and not as a coordinated risk to be integrated with technical, ethical, and regulatory dimensions of a cybersecurity program. The adversarial machine learning literature abounds with examples of AI models being susceptible to various attacks or subversion (Goodfellow et al., 2014; Carlini & Wagner, 2017). However, this is still seen as a kind of technical edge case in cybersecurity management literature, despite such risks being practically relevant and material. This structural gap in the management literature precludes development of coordinated mitigation approaches that can holistically integrate all program-level risks (technical, ethical, and regulatory) into a unified threat intelligence and control matrix.

7.3 Discussion of the study's relevance for critical infrastructure operators and policymakers

The critical infrastructure cybersecurity operators, in light of the above analysis, should bear in mind that AI-centric cybersecurity is much more than just technical systems in use. It is also a software development project (training) phase, a risk management practice (operations), and an

ethical and regulatory decision (deployment). The practical implications for critical infrastructure operators is therefore that they should pay special attention to the interdependencies between robustness, ethics and compliance and address them explicitly across the AI system lifecycle.

The same issue, but in reverse, is faced by policymakers. The available regulation and standards provide general, high-level guidance on data security and privacy, ethical risk, and risk management practice, but few operational details are given on managing adversarial resilience, explainability, and even lifecycle risks associated with AI models used in critical infrastructure protection (Regulation, 2016; Cybersecurity, 2018). As such, the standards and regulations may need to evolve to take AI-specific risks into consideration, while still leaving room for innovation and not constraining operations or commercial value.

Overall, the argument developed here is that a comprehensive understanding and management of AI-centric cybersecurity programs for critical infrastructure needs to work across technical, ethical and regulatory dimensions, and across all stages of an AI system's lifecycle. Addressing the above-identified gaps and their implications will be critical for continuing to build trust, resilience, and compliance in the future as AI continues to play an increasingly important role in critical infrastructure protection.

8. Conclusion and Future Research Directions

8.1 Main Results

The study explores ethical and regulatory implications of program management of AI-centric cybersecurity systems in critical infrastructure. The results corroborate the premise that AI can enhance cybersecurity program effectiveness by providing advanced intrusion detection, anomaly identification, and adaptive response capabilities. Existing research has shown that ML-based approaches can outperform rule-based systems in managing complex and high-volume data streams, especially in cyber-physical and industrial control system (ICS) environments (Buczak & Guven, 2015; Khraisat et al., 2019; Humayed et al., 2017). These capabilities are particularly relevant to critical infrastructure systems, which often have high availability, safety, and reliability requirements (Knowles et al., 2015; Stouffer et al., 2011).

On the other hand, the research indicates that incorporating AI into cybersecurity programs can significantly increase ethical and regulatory risk. AI-based security systems can be subject to various forms of adversarial attacks, such as data poisoning, evasion, and model extraction, which can compromise detection accuracy and system reliability (Dalvi et al., 2004; Barreno et al., 2006; Biggio & Roli, 2018). In addition, adversarial examples can persist and evade defenses over time, posing a challenge to the robustness of learning-based security mechanisms in adversarial settings (Goodfellow et al., 2014; Carlini & Wagner, 2017).

Model opacity is another fundamental issue that remains unresolved. The lack of transparency and interpretability of black-box decision-making processes can limit accountability and explainability and make it difficult to meet regulatory expectations in safety-critical domains (Sommer & Paxson, 2010; Guidotti et al., 2018). The ethical concerns of trust, oversight, and responsibility associated with AI-based cybersecurity systems are exacerbated in cases where automated responses can have a direct impact on the delivery of critical services or public safety (Floridi et al., 2018; Jobin et al., 2019).

8.2 Recommendations for Practice and Policy

The study points to the need for holistic, program-level governance approaches to manage the ethical and regulatory dimensions of AI-driven cybersecurity programs in critical infrastructure. AI should not be viewed as an isolated technical asset but as a long-term program component integrated into an organization's risk management, compliance, and oversight processes (Knowles et al., 2015). Cybersecurity managers and operators should consider lifecycle governance of AI models, including training, deployment, monitoring, and retirement, with a focus on adversarial risk management and regulatory compliance (Papernot et al., 2018).

From a policy perspective, the research suggests that AI-driven cybersecurity program operations should be aligned with existing regulatory and standards-based frameworks. While cybersecurity guidance such as the NIST Cybersecurity Framework and industrial control system security recommendations provide a foundation for risk management, additional provisions may be needed to address the unique challenges and risks of AI-based systems (Cybersecurity, 2018; Stouffer et al., 2011). Regulators and policymakers can play an important role in incentivizing governance approaches that effectively integrate ethical AI principles with cybersecurity-specific compliance needs, especially in highly regulated domains with strict data protection and safety requirements (Regulation, 2016).

In particular, explainability and documentation should be leveraged to enhance ethical and regulatory compliance of AI-centric cybersecurity systems. Explainability techniques can help increase user and operator trust and provide a basis for accountability in regulated environments (Ribeiro et al., 2016; Guidotti et al., 2018). Documentation and model cards provide a useful mechanism for communicating model purpose, capabilities, limitations, and ethical considerations and supporting the responsible deployment and auditability of AI in cybersecurity programs (Mitchell et al., 2019).

8.3 Directions for Future Research

Future work should focus on the empirical validation of responsible AI governance frameworks in operational cybersecurity programs in critical infrastructure environments. While much of the current discourse on ethical and responsible AI is conceptual and normative, there is limited evidence of how such principles and frameworks apply to and perform in real-world

cybersecurity settings, which are often characterized by complex adversarial dynamics and regulatory constraints (Floridi et al., 2018; Jobin et al., 2019). Experimental studies, as well as case-based evaluations, can help provide insight into the operational impact of governance mechanisms on security performance, ethical compliance, and organizational decision-making.

Longitudinal studies are another important area for future research. AI-centric cybersecurity programs are not static; they evolve as models adapt to new data and as threat actors and defenders modify their behaviors and strategies. Long-term studies can shed light on how ethical and regulatory risks develop and change over time, especially in AI systems that continuously learn and update their models (Papernot et al., 2018; Biggio & Roli, 2018). This can inform the design of governance strategies that can adapt to shifting risk landscapes.

Finally, more work is needed to establish standardized evaluation metrics that can jointly assess AI-centric cybersecurity programs on multiple dimensions, including cybersecurity effectiveness, adversarial robustness, explainability, and regulatory compliance. The development and validation of unified evaluation frameworks will be critical to enabling consistent benchmarking of different AI-based cybersecurity programs and supporting evidence-based decisions by policymakers, regulators, and cybersecurity program managers.

References

1. Barreno, M., Nelson, B., Sears, R., Joseph, A. D., & Tygar, J. D. (2006, March). Can machine learning be secure?. In Proceedings of the 2006 ACM Symposium on Information, computer and communications security (pp. 16-25).
2. Biggio, B., & Roli, F. (2018, October). Wild patterns: Ten years after the rise of adversarial machine learning. In Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security (pp. 2154-2156).
3. Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., ... & Amodei, D. (2018). The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. arXiv preprint arXiv:1802.07228.
4. Buczak, A. L., & Guven, E. (2015). A survey of data mining and machine learning methods for cyber security intrusion detection. IEEE Communications surveys & tutorials, 18(2), 1153-1176.
5. Carlini, N., & Wagner, D. (2017, May). Towards evaluating the robustness of neural networks. In 2017 ieee symposium on security and privacy (sp) (pp. 39-57). Ieee.
6. Dalvi, N., Domingos, P., Mausam, Sanghai, S., & Verma, D. (2004, August). Adversarial classification. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 99-108).
7. Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., ... & Vayena, E. (2018). AI4People—An ethical framework for a good AI society:

Opportunities, risks, principles, and recommendations. Minds and machines, 28(4), 689-707.

8. Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572.
9. 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 (CSUR), 51(5), 1-42.
10. Humayed, A., Lin, J., Li, F., & Luo, B. (2017). Cyber-physical systems security—A survey. IEEE Internet of Things Journal, 4(6), 1802-1831.
11. Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature machine intelligence, 1(9), 389-399.
12. Khraisat, A., Gondal, I., Vamplew, P., & Kamruzzaman, J. (2019). Survey of intrusion detection systems: techniques, datasets and challenges. Cybersecurity, 2(1), 1-22.
13. Knowles, W., Prince, D., Hutchison, D., Disso, J. F. P., & Jones, K. (2015). A survey of cyber security management in industrial control systems. International journal of critical infrastructure protection, 9, 52-80.
14. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... & Gebru, T. (2019, January). Model cards for model reporting. In Proceedings of the conference on fairness, accountability, and transparency (pp. 220-229).
15. Cybersecurity, C. I. (2018). Framework for improving critical infrastructure cybersecurity. URL: https://nvlpubs. nist. gov/nistpubs/CSWP/NIST. CSWP, 4162018(7).
16. Papernot, N., McDaniel, P., Sinha, A., & Wellman, M. P. (2018, April). Sok: Security and privacy in machine learning. In 2018 IEEE European symposium on security and privacy (EuroS&P) (pp. 399-414). IEEE.
17. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). " Why should i trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135-1144).
18. Sommer, R., & Paxson, V. (2010, May). Outside the closed world: On using machine learning for network intrusion detection. In 2010 IEEE symposium on security and privacy (pp. 305-316). IEEE.
19. Stouffer, K., Falco, J., & Scarfone, K. (2011). Guide to industrial control systems (ICS) security. NIST special publication, 800(82), 16-16.
20. Regulation, P. (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council. Regulation (eu), 679(2016), 10-13.