Toward a Hybrid Model for Incident Response: Integrating Traditional Frameworks with AI-Driven Approaches and Regualtions

Srikamkshi Mahesh
Information Security and Policy Management
Heinz College of Heinz College of
Information Systems and Public Policy
line 4: Pittsburgh, United States
srikamam@andrew.cmu.edu

Abstract— **The increasing sophistication and pace of cyber threats present significant challenges for current incident response strategies, whether primarily manual or fully automated. Traditional frameworks often lack the necessary adaptability for modern attacks, while purely AI-driven systems raise concerns about accountability and clarity. This paper introduces a novel Hybrid Incident Response Framework (HIRF) that integrates the strengths of both human-centric and AI-powered approaches. HIRF aims to combine structure and oversight with speed and analytical capabilities across all stages of incident handling. By examining existing methods and proposing this integrated model, this research highlights the potential of HIRF to enhance organizational resilience and improve response effectiveness in the evolving landscape of cyber security. This work suggests a direction for advancing incident response capabilities to address increasingly complex threats.**

Keywords—, Incident Response, Hybrid Incident Response Framework (HIRF), Artificial Intelligence (AI), Traditional Security Frameworks, Organizational Resilience, Automation, Human-Centric Approaches, Response Effectiveness

I. Introduction

In today's rapidly evolving cybersecurity landscape, the capacity to respond swiftly and effectively to security incidents is paramount for minimizing organizational damage, preserving stakeholder trust, and ensuring operational continuity. Historically, organizations have relied on established incident response (IR) frameworks-such as NIST SP 800-61, ISO/IEC 27035, the SANS Incident Handling Model, and the CERT/CC framework-to guide their incident management strategies. These models, emphasizing structured preparation, systematic detection, containment, eradication, recovery, and post-incident analysis, have long served as the cornerstone of organizational resilience.

However, the threat environment has undergone a profound transformation. Modern cyber-attacks are characterized by unprecedented scale, speed, and sophistication, frequently leveraging automation, artificial intelligence (AI), and novel attack vectors that can rapidly overwhelm traditional defenses[1]. In parallel, AI and machine learning (ML) are increasingly being integrated into cybersecurity operations, offering the promise of enhanced detection accuracy, real-time decision-making, predictive analytics, and accelerated containment. Yet, these advancements also introduce new risks, including algorithmic bias, explainability challenges, susceptibility to adversarial manipulation, and heightened regulatory scrutiny.

This thesis addresses the critical need to bridge the gap between traditional and AI-driven approaches to incident response. While classical frameworks provide essential structure, human judgment, and legal accountability, they often lack the agility and scalability required to counter contemporary threats. Conversely, AI-based systems deliver automation and predictive capabilities but raise concerns related to transparency, trust, and governance. Neither paradigm, in isolation, is sufficient to address the multidimensional demands of modern cybersecurity.

A. Focus and Scope

This research is dedicated to the design of a **Hybrid Incident Response Framework (HIRF)** that synthesizes the strengths of both traditional and AI-driven methodologies. The HIRF aims to optimize detection, response, and recovery processes while proactively mitigating the risks inherent to AI technologies. The scope of this study encompasses:

- A critical analysis of the strengths and limitations of established and emerging incident response models.
- Identification of operational gaps where traditional frameworks fail to address the complexities introduced by AI.
- The design of a structured hybrid framework that balances automation with robust human oversight across all phases of incident response.

The research deliberately centers on cybersecurity incident response-distinct from broader risk management or enterprise AI governance-emphasizing operational resilience, explainability, legal compliance, and adaptability

B. Relevance of Research

This work builds upon and extends the existing body of literature on incident response and AI integration. While numerous studies have explored the benefits and challenges of traditional IR frameworks and the promise of AI-driven automation, few have systematically addressed the integration of these paradigms into a unified, operationally viable framework [2-6]. Existing models often focus exclusively on either traditional playbooks or fully automated approaches, frequently overlooking critical issues such as governance, bias mitigation, and explainability. Moreover, the emergence of new regulatory instruments-including the NIST AI Risk Management Framework, ISO/IEC 42001, and the EU AI Act-imposes additional legal and ethical obligations on the use of AI in security operations, further complicating the landscape for purely AI-driven strategies.

By proposing a hybrid model that aligns with these regulatory and operational realities, this thesis seeks to fill a significant gap in both academic research and professional practice.

C. Research Question and Objectives

This study is guided by the following central research question:

How can a hybrid incident response framework effectively integrate traditional models and AI-driven approaches to enhance cybersecurity operations, while mitigating the unique risks associated with AI systems?

To address this question, the research pursues the following objectives:

- Conduct a systematic review of traditional incident response frameworks and AI-driven cybersecurity models.
- Identify operational gaps and emerging risks introduced by AI technologies within incident response processes.
- Design and propose the Hybrid Incident Response Framework (HIRF) that systematically integrates human oversight with AI automation.
- Evaluate the proposed framework against known challenges in incident response to demonstrate its theoretical soundness and practical utility.

Through these objectives, this study aims to provide cybersecurity practitioners and researchers with a structured and adaptable pathway to modernize incident response capabilities, while maintaining trust, transparency, and resilience in an increasingly AI-augmented environment.

II. LITERATURE REVIEW

To propose a comprehensive framework for a hybrid model of incident response, it is essential to first investigate existing frameworks and methodologies. This exploration provides critical insight into the current landscape and allows us to assess whether these frameworks sufficiently address the complexities introduced by AI-driven incident response mechanisms. The review then continues with an examination of traditional incident response models, highlighting their strengths and limitations. Following this, we delve into AI-based techniques, evaluating their potential benefits as well as the novel challenges they introduce. Lastly, we examine techniques and frameworks that aim to mitigate the vulnerabilities and risks specifically associated with AI-enabled incident response systems. Together, these areas of focus contribute to a holistic understanding that underpins the development of a new, hybrid incident response framework

A. Traditional Incident Response Frameworks

1) NIST SP 800-61 Revision 3

The NIST SP 800-61 framework [2] defines six distinct phases in the incident response lifecycle. These phases are designed to manage incidents effectively, reduce their impact, and enhance cybersecurity risk management based on lessons learned. The framework is broadly divided into two categories: Govern, Identify, and Protect, which relate to planning and preparedness; and Detect, Respond, and Recover, which pertain to incident management.

An important feature of this framework is its emphasis on stakeholder roles and responsibilities throughout the incident response process by outlining the development of incident response policies, procedures, and playbooks. A key component is the Community Profile for incident response, which prioritizes cybersecurity outcomes, offers recommendations, and provides supporting information relevant to achieving desired outcomes within the context of incident response.

2) *SANS Incident Handling Framework*

Like the NIST framework, the SANS framework [3] outlines six stages of incident handling: Preparation, Identification, Containment, Eradication, Recovery, and Lessons Learned. Each phase is accompanied by detailed guidance, including example activities, recommended tools, and stakeholder involvement at each stage. This framework effectively serves as a checklist for incident handlers, ensuring that all essential tasks are completed to enable a robust and efficient incident response. Its practical orientation makes it especially useful for operational teams aiming to implement structured and repeatable incident response processes.

3) *ISO/IEC 27035*

The ISO/IEC 27035 framework also presents a structured five-phase model for incident response: Plan and Prepare, Detect and Report, Assess and Decide, Respond, and Learn Lessons. Unlike the previous models, ISO adopts a more holistic and process-driven approach, aligning incident response with broader organizational functions such as risk management, business continuity, and governance.

This framework is part of the larger ISO/IEC 27000 family and places significant emphasis on documented policies, procedures, training, and inter-organizational relationships. It underscores the importance of time-sensitive responses and proper management of incident data by the Incident Response Team (IRT), ensuring actions are in accordance with predefined procedures [4].

4) *CERT/CC Incident Response Model*

Although it is an older framework compared to some modern models, the CERT/CC Incident Response Model remains highly relevant and is often used in conjunction with others, such as the SANS model, to effectively address and mitigate security threats. It outlines five key processes—Prepare, Protect, Detect, Triage, and Respond—that guide incident handling. What sets this model apart is its broader perspective: it views incident response not only as a means of securing systems through risk management, but also to ensure organizational resilience in the face of incidents. This holistic approach not only emphasizes technological controls but also the critical role of human actions and decision-making during crisis [5].

The incident response models mentioned above are widely used by organizations to build robust incident response strategies. These frameworks provide a well-defined and structured pipeline that helps organizations take control of incidents and minimize their impact. In addition to guiding technical actions, these models also define the roles and responsibilities of the incident response team members, ensuring that each person knows their part in managing a crisis.

B. *Traditional Frameworks and Industry Adoption*

Over the years, frameworks such as ISO/IEC 27035 and NIST SP 800-61 have been instrumental in helping organizations design response strategies aligned with their business needs. In the paper Information Security Incident Management: Current Practice as Reported in the Literature [6], Inger Anne Tøndel evaluates how various organizations apply these frameworks in practice. Her research sheds light on the commonly followed stages of incident response and how they align with frameworks discussed earlier

1) *Planning and Preparation*

The first stage involves planning and preparation. Organizations begin by defining what constitutes a relevant incident in their context—such as data breaches, system misconfigurations, or service interruptions. This phase goes beyond definitions, requiring contingency planning, role assignments, and the implementation of proactive practices like vulnerability assessments and penetration tests. Training staff is also essential at this stage, equipping them with the knowledge needed to recognize and respond to threats effectively.

2) Detection and Reporting

The next stage is detection and reporting. Organizations typically use both manual and automated approaches. Manual detection includes alerts from internal teams, reports from external entities, and direct communication via email or phone. On the other hand, automated systems like IDS and IPS continuously monitor networks and systems to detect anomalies. However, before AI became mainstream, these tools often struggled with high false positive rates. Compatibility issues between tools or unexpected system behavior also introduced complexity. Even with automated ticketing systems in place, only 17% of surveyed IT security managers stated that all incidents were consistently logged, with up to 50% reporting that some alerts came via informal channels like email or phone [7].

3) Assessment and Decision-Making

Once an alert is received, the incident enters the assessment and decision-making phase. Incidents are evaluated and prioritized—typically as low, medium, or high—based on various criteria, such as the number of systems affected, the type of systems impacted, and whether the affected assets are internal or customer-facing. While security tools assist in analysis, much of the assessment relies on personnel who understand the organization's systems intimately. Their experience, intuition, and situational awareness are often essential in accurately gauging severity and determining the appropriate course of action.

4) Incident Response Execution

After the assessment, organizations shift into the response phase. This stage involves taking immediate action to resolve the incident or, in some cases, applying a temporary solution to maintain operational continuity while the root issue is addressed. Throughout the response, detailed documentation is maintained to track every step taken, facilitating transparency and coordination across departments. When incidents are caused by human error, this documentation helps identify gaps and ensures that stakeholders are informed. Low-priority or recurring issues are often handled automatically to reduce response times and minimize workload.

5) Post-Incident Review and Lessons Learned

The final stage involves a structured review once the incident has been resolved. The timing and depth of the review usually depend on the severity of the incident. Critical incidents are typically reviewed within 24 hours, while lower-impact events are addressed during regular review meetings [8]. These reviews provide opportunities to fine-tune systems, refine processes, and prevent recurrence. However, one of the biggest challenges in this phase is the lack of information sharing. Internally, this may stem from concerns over accountability. Externally, especially when third parties are involved, issues related to trust, liability, and data sensitivity often hinder open communication.

The practices observed in organizations reflect a time when AI was not yet a central player in incident response. Although automation was used in areas like detection and ticketing, the overall approach remained heavily reliant on traditional tools and human expertise. These limitations highlighted the growing need for more intelligent and adaptive response mechanisms.

C. Transitioning into AI-Powered Incident Response

With the ever-increasing volume and sophistication of cyber threats, organizations are turning to AI to revolutionize how incidents are detected, analyzed, and mitigated. From real-time threat detection to intelligent decision-making and response automation, AI has begun to address many of the shortcomings of traditional approaches. In this section, we’ll explore how modern incident response strategies are evolving through AI integration.

One of the most impactful areas of artificial intelligence (AI) integration can be found in the incident response strategies employed by information security professionals. As cyber-attacks become increasingly sophisticated and frequent, organizations have turned to AI and machine learning (ML) solutions to meet the growing demands of cybersecurity. AI has proven especially valuable in areas where human capabilities are limited—particularly in the detection and analysis of large volumes of data, where isolating patterns or anomalies is crucial to reducing vulnerabilities stemming from human error [9][10].

A significant number of studies have focused on AI's role in IT and computer-based incident response due to its capacity to cross-analyze diverse data sources such as system logs, network traffic, and user activity logs. AI systems excel at identifying deviations from normal activity and, to some extent, can even anticipate cyber-attacks before they occur. Research efforts [11][12] into AI integration have led to advancements in technologies such as Intrusion Detection Systems (IDS), Intrusion Prevention Systems (IPS), Security Information and Event Management (SIEM) systems, malware detection tools, and Identity and Access Management (IAM) systems. These systems offer high accuracy and low false positive rates, significantly improving the speed and efficiency of incident resolution.

Beyond IT environments, AI has also shown substantial benefits in industries where digital technologies intersect with the physical world [13][14]. Devices like sensors, processors, and Internet of Things (IoT) systems collect real-time data from the environment to inform and control physical processes. AI analyzes this data—along with historical trends—to enhance hardware performance and optimize infrastructure efficiency. In incident response, this capability is particularly valuable for predictive maintenance: AI can forecast the potential failure of hardware components, allowing organizations to replace them proactively and avoid costly downtime. This approach has saved companies considerable resources by preventing production halts and reducing operational risks.

In Adebayo D'Costa's review of AI-driven triage systems [15], the author explores how AI technologies are being used to assess, categorize, and prioritize patients more objectively and efficiently than traditional methods. These systems leverage data from electronic health records (EHRs), wearable sensors, and real-time patient inputs to quickly analyze symptoms, vital signs, and historical health trends.

AI-based triage tools can support clinicians by suggesting likely diagnoses, flagging critical conditions, and recommending prioritization based on the severity and urgency of cases. This is especially valuable in high-demand scenarios—such as natural disasters, pandemics, or mass casualty incidents—where healthcare providers are overwhelmed, and manual triage is too slow or inconsistent. For example, during the COVID-19 pandemic, AI was used to remotely screen patients, allocate ventilators, and optimize ICU bed usage. In such contexts, AI not only accelerates response times but also helps reduce bias and human error, ensuring that resources are allocated where they are most needed.

The overarching theme across all these examples is that AI significantly enhances incident response by increasing accuracy, reducing false positives, and automating repetitive or complex tasks—thereby minimizing the influence of human error and emotional bias. By streamlining decision-making and enabling faster, data-driven actions, AI supports more efficient and consistent responses across various domains.

However, these advantages come with notable challenges. One of the key drawbacks is the need for vast and diverse datasets to effectively train AI models—data that must be accurate, representative, and ethically sourced. Additionally, AI systems can inherit biases from their training data, leading to skewed or unfair outcomes if not properly monitored. This underscores the importance of developing robust ethical and legal frameworks to guide AI deployment, ensuring that these technologies operate within the bounds of privacy laws, accountability standards, and fairness principles while maintaining their effectiveness in real-world applications.

D. Policies and Laws Governing AI

The widespread use of Artificial Intelligence (AI) and its integration into critical sectors such as healthcare, finance, and security has led to an increased emphasis on the governance of AI systems and their responsible use. Given the tremendous benefits and potential drawbacks of AI, several frameworks have been developed to promote the ethical, transparent, and secure deployment of AI technologies. Among the most prominent are the NIST AI Risk Management Framework (AI RMF) and the ISO/IEC 42001 standard.

1) NIST AI Risk Management Framework (AI RMF)

The NIST AI Risk Management Framework [16] (AI RMF) provides organizations with guidance on understanding and managing the risks associated with AI systems across their lifecycle. It addresses concerns such as security vulnerabilities, the introduction of bias, ethical issues, and potential points of failure. The framework outlines four core actions — Map, Measure, Manage, and Govern — that organizations should take to systematically assess and mitigate AI risks:

- Map: Identify and contextualize the risks associated with AI systems by engaging relevant roles and stakeholders early in the lifecycle.
- Measure: Continuously monitor, assess, and benchmark AI systems to understand the nature and magnitude of associated risks.
- Manage: Implement strategies to mitigate identified risks through proper configuration, operational controls, and resource allocation.
- Govern: Establish governance structures, policies, accountability mechanisms, and roles to ensure responsible AI deployment.

2) ISO/IEC 42001: AI Management Systems Standard

ISO/IEC 42001[17] recommends the establishment of an AI Management System that continuously monitors and governs AI issues related to risk management, transparency, accountability, robustness, fairness, and data handling. Unlike the AI RMF, ISO/IEC 42001 places particular emphasis on context-specific adaptation and stakeholder engagement. It requires organizations to:

- Clearly define AI use cases
- Understand internal and external threats
- Identify and involve relevant stakeholders.

While ISO/IEC 42001 offers the option of formal certification, its implementation remains voluntary — similar to the AI RMF. Together, these frameworks provide a foundational approach to fostering trustworthy and resilient AI systems.

3) EU AI Act

While previous models such as the NIST AI RMF and ISO 42001 are voluntary, the EU AI Act [18] takes a pivotal step by legalizing AI regulation based on principles of privacy, human rights, and safety. This Act governs the development, deployment, and use of AI across the European Union, setting a binding legal framework for all AI systems operating within its jurisdiction.

The EU AI Act classifies AI systems into four tiers based on their risk levels:

- Unacceptable Risk: AI systems that pose clear threats to people's safety, livelihoods, or fundamental rights are strictly prohibited (e.g., social scoring by governments, manipulative AI applications).
- High Risk: AI systems used in critical sectors such as biometric identification, critical infrastructure, education, healthcare, employment, and law enforcement must comply with strict requirements related to transparency, data quality, human oversight, and robustness.
- Limited Risk: AI systems like chatbots must meet basic transparency obligations, such as informing users that they are interacting with an AI system.
- Minimal or No Risk: Most AI applications, such as spam filters or simple recommendation engines, are largely unregulated but encouraged to follow voluntary codes of conduct.

For high-risk AI systems, the Act emphasizes the need for strong practices in risk management, data governance, technical documentation, human oversight, and robust testing to ensure the system's safety and accountability and failure to do so can result in be up to €35 Million or 7% of global revenue for Unacceptable risk violations and €15 Million or 3% for high-risk violations.

III. METHODOLOGY

To develop a comprehensive Hybrid Incident Response Framework (HIRF) that integrates both traditional and AI-driven approaches, this study adopts a structured, multi-phase methodology. The process begins with an extensive literature review to establish a thorough understanding of the current incident response landscape. Traditional frameworks, including NIST SP 800-61, the SANS Incident Handling Framework, ISO/IEC 27035, and the CERT/CC Incident Response Model, are critically analyzed to assess their strengths, limitations, and adaptability to evolving cybersecurity threats.

Following the review of traditional frameworks, the study examines contemporary AI-driven incident response techniques. This includes an evaluation of machine learning (ML) and artificial intelligence (AI) models applied to different phases of the incident response lifecycle, such as anomaly detection, automated triage, dynamic containment, and predictive recovery. Emphasis is placed on identifying both the benefits these techniques offer, such as enhanced speed and scalability, and the challenges they introduce, including issues of bias, explainability, and operational complexity.

In parallel, the research explores emerging regulatory, ethical, and risk management frameworks specifically aimed at governing AI use within critical systems, such as the NIST AI Risk Management Framework and the European Union AI Act. This examination ensures that the proposed hybrid framework aligns not only with technical best practices but also with evolving compliance and governance standards.

The insights gathered from these analyses are synthesized to identify commonalities, gaps, and potential points of integration between traditional and AI-driven approaches. Based on this synthesis, the study proposes a novel hybrid framework that strategically combines the structure, accountability, and resilience of traditional models with the speed, adaptability, and analytical capabilities of AI-based systems.

IV. ANALYSIS

The integration of AI into incident response frameworks has introduced a range of new challenges, vulnerabilities, and complexities that are often overlooked. Traditional incident response strategies have typically focused on the incident itself — identifying, containing, and remediating threats — without giving equal attention to optimizing the strategies and processes used to manage those incidents. This oversight can lead to gaps in preparedness, especially as the nature of threats evolves alongside technological advancements. AI-driven systems bring powerful capabilities, such as real-time threat detection, automated decision-making, and predictive analysis. However, they also introduce risks that can create new attack surfaces that traditional frameworks are not equipped to handle. The conventional incident response methods, built primarily around manual processes and human expertise, offer strengths like adaptability, contextual judgment, and creativity in complex or novel situations. However, they are often slower, less scalable, and prone to human error, especially under pressure.

The following sections will outline some of the specific risks introduced by AI systems, highlight gaps in traditional approaches which help in not only accommodating the realities of AI-augmented operations but also strengthens the overall resilience and agility of the incident response strategy.

A. Resource Requirements and Usage

Developing an AI-based system that achieves high accuracy requires substantial resources, including access to high-quality training data, the selection of suitable algorithms, effective training strategies, and significant computational power. Research into this area revealed several key factors:

- Data Quality and Suitability:
  The training data must closely align with the system’s detection objectives. It must also be clean, diverse, balanced, and properly labeled. Achieving such data quality is particularly challenging in domains involving confidential or sensitive information—such as cybersecurity incident response—where access to real-world datasets is restricted. According to the studies conducted on data requirements for model training [19][20], organizations often must rely on synthetic or internally generated datasets to meet these requirements.
- Model Development and Expertise:
  It is not sufficient to have personnel capable of developing machine learning models; the involvement of Machine Learning Systems Engineers (MSEs) is crucial. MSEs guide the development process by helping to identify key features, prioritize tasks, and refine models to optimize performance and efficiency.
- Computational Infrastructure:
  AI models require robust computational resources to operate effectively. While many models presented in the reviewed studies were developed as prototypes using curated datasets, deploying models that perform well in real-world environments demands significantly greater

infrastructure. According to survey conducted in "Green AI" [21], the total cost of an AI solution increases linearly with three factors: the cost of processing each example, the size of the training dataset, and the number of hyperparameter optimization experiments conducted.

These escalating resource requirements present substantial challenges, especially in industries such as healthcare, where technological investments are often limited by strict budgetary constraints. As highlighted in the paper "The importance of resource awareness in artificial intelligence for healthcare" [22], the sustainability of AI applications in healthcare is increasingly under scrutiny due to the high energy consumption, computational demands, networking requirements, and storage needs associated with maintaining these systems.

In contrast, traditional detection techniques—such as signature-based detection, rule-based systems, and heuristic methods—do not impose the same resource demands. These methods rely on predefined rules and known attack vectors, enabling effective detection with minimal data requirements and limited computational overhead. Traditional systems also do not require continuous access to high-quality datasets, as no ongoing training is necessary.

However, traditional techniques have notable limitations. Their performance is inherently restricted to the completeness and currency of the rules or signatures they employ. As a result, they may fail to detect novel or sophisticated threats that do not match known patterns. Nevertheless, traditional methods remain suitable for environments with limited budgets and infrastructure, where resource optimization is critical.

A promising approach to overcoming the shortcomings of both AI-driven and traditional methods is the development of hybrid systems [23]. A hybrid solution integrates traditional techniques with AI capabilities, thereby addressing the challenges associated with each approach when used independently. In a hybrid model, traditional systems handle routine threats, reducing AI's computational load, while AI focuses on complex, ambiguous cases. This approach optimizes resources, improves data utilization by feeding traditional outputs into AI training, and enhances accuracy and resilience by combining rule-based and adaptive methods.

B. *Explainability and Transparency*

A fundamental requirement for effective incident response is the ability to understand how and why an incident is occurring. Only with this knowledge can responders validate alerts, distinguish between true and false positives, and select appropriate courses of action. Explainability empowers analysts to make informed decisions, while transparency fosters trust and accountability by providing clear insight into the rationale behind specific actions.

In contemporary security environments, the need to explain why decisions are made has become increasingly important. Stakeholders-including those responsible for incident response-require not only notification of what actions were taken, but also a clear justification for those actions.

Traditional incident response frameworks have addressed these needs through rigorous documentation, auditing, and accountability measures. These frameworks emphasize the importance of documenting every step from initial detection to final remediation, both for post-incident review and organizational learning. Auditing ensures regulatory and legal compliance, while effective communication guarantees that all stakeholders are informed and that decisions are timely and well-accepted. However, these strategies primarily capture the "who," "when," and "what" of incident response, rather than providing substantive explanations of "why" particular decisions were made. This is largely because traditional systems rely on predetermined logic and signatures, which are explicitly defined by human experts.

The advent of artificial intelligence (AI) in incident response introduces new challenges. Although AI models are initially trained on specific parameters to guide their decision-making, they often evolve in complexity, resulting in systems that function as "black boxes" [23]. Such models can become highly opaque, making it difficult-even for their creators-to understand the reasoning behind specific predictions or actions. This opacity can lead to critical miscommunications or unintended consequences. For example, if one team disables a device in response to an unexpected event, but another team's machine learning model subsequently flags this action as anomalous and automatically initiates remediation, the lack of explainability and coordination could result in significant operational disruptions.

Addressing these challenges within the broader context of incident response necessitates the integration of explainable artificial intelligence (XAI) [23] and the continued involvement of human oversight [24]. Most existing incident response frameworks do not explicitly address the need for explainability when AI is incorporated. While traditional frameworks implicitly trust human judgment and prioritize accountability through documentation, the introduction of AI systems requires a deliberate focus on explainability requirements. Only by explicitly incorporating explainability into incident response frameworks can organizations ensure that AI-driven decisions remain transparent, justifiable, and aligned with broader organizational goals and regulatory standards.

C. Biases and Adversarial Attacks

The research literature consistently highlights bias and adversarial attacks as primary concerns in AI-based systems. This extends earlier discussions about the importance of explanation and trust, emphasizing that these vulnerabilities can fundamentally undermine the reliability, security, and long-term stability of AI-driven incident response frameworks. As organizations increasingly rely on AI to automate detection and response tasks, the potential impact of these risks grows more severe.

Bias in AI typically arises when the training data does not comprehensively represent the diversity of the environment the system is meant to protect. If the data is skewed, incomplete, or outdated, the model may inadvertently favor certain behaviors, protocols, or network patterns while overlooking others. This imbalance can lead to two major types of errors: the AI may misinterpret benign activities as threats (resulting in false positives), or worse, fail to detect genuine attacks (false negatives) that deviate from its learned patterns [25]. It is important to note that traditional systems are not immune to bias either; however, in their case, bias generally stems from human assumptions and design choices embedded in static rules and procedures. Because traditional logic is hand-crafted and visible, identifying and correcting such bias tends to be more straightforward.

Adversarial attacks introduce another serious dimension of risk. These attacks involve malicious actors intentionally crafting poisoned or manipulated data to deceive AI models. By subtly altering inputs during training or live operation, attackers can corrupt the model's internal representations, causing it to misclassify malicious activities as safe or trigger a flood of false alarms [27]. This technique can be used both to bypass security defenses and to exhaust incident response teams by overwhelming them with noise. Although traditional rule-based systems can also be deceived—usually by attackers deviating from known patterns—the adaptability of AI systems makes them particularly vulnerable. The same learning mechanisms that allow AI to evolve and improve over time can, if manipulated, lead to unintended behaviors without immediate visibility.

While both AI and traditional techniques face similar types of vulnerabilities, the magnitude of the risk is amplified in AI-driven systems because their behaviors are learned rather than explicitly programmed. In traditional systems, the decision-making process is transparent, auditable, and modifiable: if a flaw is discovered, security teams can update or patch the rule set directly. In contrast, AI models, especially those with opaque, complex, or "black box" architectures—can develop biases or fall victim to adversarial influences without easy detection [28]. The lack of transparency makes tracing the source of an error extremely challenging, often requiring extensive investigation into the layers of model training, feature selection, and learning history.

One particularly severe threat unique to AI is "catastrophic forgetting", a phenomenon where undetected bias or adversarial influence does not just cause new mistakes but actively erodes the model's ability to correctly perform tasks it had previously mastered. Over time, an AI model could degrade in performance to the point where it no longer detects attacks it was once highly accurate at identifying, without clear warning signs. In traditional systems, the worst-case scenario is usually a missed event or false alert that can be corrected through periodic ruleset reviews. In contrast, AI systems demand continuous oversight, periodic retraining, and rigorous validation to ensure that they do not silently degrade or shift their detection priorities.

These factors make bias and adversarial attacks significantly more critical concerns in AI-driven incident response compared to traditional systems. While traditional approaches remain relatively stable and manageable through regular updates and human reviews, the dynamic, evolving, and opaque nature of AI requires organizations to implement stronger safeguards. Measures such as explainable AI (XAI), regular auditing, adversarial robustness testing, and hybrid models [29] that combine traditional logic with AI intelligence are essential to ensuring that incident response systems remain trustworthy, resilient, and aligned with organizational risk management strategies over time. Without these precautions, there is a real danger that AI systems could fail silently, or worse, be subverted in ways that traditional defenses would not allow.

D. Bridging Gaps in Traditional Incident Response

The analysis comparing traditional techniques and AI-focused incident response highlights recurring challenges such as resource constraints, transparency and explainability issues, biases, and evolving threat surfaces — concerns also emphasized in AI risk management literature. Most current incident response strategies prioritize detection but overlook the complexities that arise after AI-related incidents, leading to organizational vulnerabilities and legal risks.

Organizations must address AI-specific risks at the earliest stages of system design and incident planning, as retrofitting AI considerations later is difficult and ineffective. Traditional frameworks, designed for deterministic IT systems and reactive processes, offer limited accountability and do not account for the unique behaviors of AI systems.

Emerging AI governance standards recognize that AI models are non-deterministic, evolve over time, and can produce harmful outcomes independently of external attacks. These frameworks emphasize continuous lifecycle management, fairness, transparency, and broader accountability across
developers, deployers, and users.

Additionally, the analysis of different models shows that hybrid solutions — combining traditional methods with AI capabilities — improve incident management outcomes. Aligning with the EU AI Act’s emphasis on human oversight, hybrid approaches are especially critical when managing high-risk resources. Sole reliance on AI increases risks, whereas hybrid systems offer a more resilient and balanced approach.

Based on these findings, it is recommended that traditional incident response frameworks be systematically adapted to address the unique risks and complexities introduced by AI technologies and it is advisable to prioritize hybrid methodologies over fully autonomous AI-driven processes at this stage.

TABLE I.

Aspect	Traditional Incident Response	AI-Focused Incident Response	Gaps Identified
Resource Constraints	Human-driven, labor-intensive, often slow due to manual workflows.	Automation improves speed but demands specialized skills and computing resources.	Need for scalable resource models that balance human and AI workloads.
Transparency and Explainability	High (actions can be traced easily).	Low (AI systems often behave like "black boxes").	Lack of explainability mechanisms increases audit and compliance risks in AI-based responses.
Bias and Fairness	Minimal algorithmic bias; dependent on human judgment.	Potential algorithmic bias inherent in AI models.	Need for bias detection, fairness checks, and responsible AI practices during incident management.
Adaptability to Evolving Threats	Reactive and static; slower to adapt.	More adaptive but risks overfitting or missing novel threats.	Require dynamic threat intelligence integration and continuous learning mechanisms.
Post-Detection Response	Well-structured but assumes deterministic system behavior.	Struggles with non-deterministic AI behaviors post-incident.	Incident playbooks must account for AI model drift, autonomous decision-making errors, and unpredictability.
Accountability and Governance	Clear human accountability structures.	Ambiguous; responsibility split among developers, deployers, and users.	Need for governance models assigning accountability across the AI

			system lifecycle.

V. RECOMMENDATIONS

The Hybrid Incident Response Framework (HIRF) is proposed to bridge the gap between traditional incident response methodologies and the dynamic, automated capabilities offered by AI. This framework recognizes that neither a purely traditional nor a purely AI-driven system is sufficient to address modern information security threats. Instead, a combination of both is necessary to build an accurate, resilient, and adaptable system.

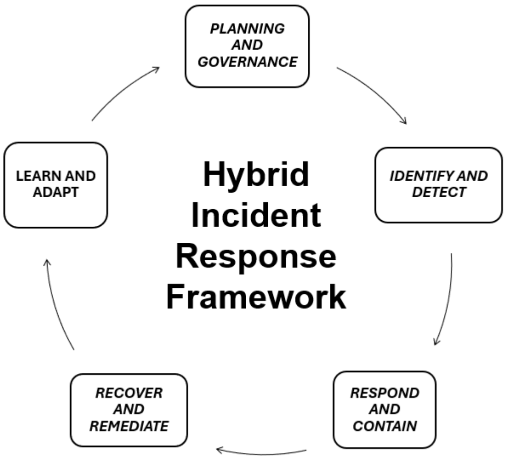

Fig 1 Hybrid Incident Response Framework Cycle

A. PLANNING AND GOVERNANCE

The first stage, Planning and Governance, establishes the foundation for an effective hybrid response. This stage involves several critical activities to ensure the system is robust, compliant, and operationally ready.

1) Asset and Resource Mapping

Organizations must first identify and catalog their critical assets, systems, and datasets. This process involves assessing the sensitivity, criticality, and business impact of each asset to prioritize protection efforts appropriately. Prioritization enables the organization to allocate resources efficiently and design incident response strategies aligned with the relative importance of each asset.

2) Policy and Legal Review

A thorough review of existing legal requirements, regulatory mandates, and internal organizational policies is essential. This ensures that incident response procedures comply with applicable laws and standards. Aligning incident response strategies with risk management objectives and legal obligations minimizes the risk of non-compliance and enhances the organization's overall security posture.

3) Role Definition

Clearly delineated roles and responsibilities are fundamental for managing incidents effectively within a hybrid model. Organizations should define specialized positions, such as AI Incident Response Leads and Human Oversight Officers, to oversee AI-driven processes. Additionally, clear escalation paths must be established to address scenarios where AI-generated decisions require human intervention, correction, or override.

4) Simulation and Training Strategy

Organizations must evaluate their computational resources and human expertise to design appropriate response models. Based on the asset mappings and threat scenarios, incident response playbooks should be developed for traditional AI-driven, or hybrid approaches The This decision matrix (Refer to Table II) ensures incident handling is tailored to the incident's criticality, data availability, and resource constraints, thereby enhancing organizational resilience. Conducting regular simulations and exercises ensures that teams are trained to operate effectively across these different models and refines organizational readiness**.**

5) Data Governance Framework

Effective data governance is crucial for the success of AI components within the incident response framework. Organizations must implement policies to maintain the integrity, quality, and security of their data. Particular emphasis must be placed on ensuring that training datasets used for AI models are clean, unbiased, current, and securely maintained. At this stage, decisions should also be made regarding the extent to which AI will be integrated into subsequent incident response stages, depending on data quality and operational needs.

TABLE II. INCIDENT RESPONSE AUTOMATION AND HUMAN OVERSIGHT BY PRIORITY LEVEL

Scenario	**Automation Level**	**Human Involvement**	**Example Actions**
Low-Priority Incidents	Full automation	Minimal to none	AI autonomously isolates threats, blocks malicious IPs, updates detection models.
Medium-Priority Incidents	Initial automation + rapid human review	Moderate — post-action validation and tuning	AI initiates immediate containment: humans review, adjust, and ensure no overreach.
High-Priority Incidents	Partial automation with continuous oversight	Significant — strategic decision-making and escalation	Human responders approve or adjust AI actions, manage high-risk containment, and coordinate business communications.

B. IDENTIFY AND DETECT

Effective planning and policy mapping empower organizations to transition seamlessly into subsequent phases of the Hybrid Incident Response Framework. This ensures both agility and control when responding to security incidents. In this phase, the identification and detection of incidents become more streamlined, proactive, and resilient. The hybrid model is particularly beneficial here, as it leverages the strengths of both traditional and AI-driven methods to maximize reliability, accuracy, and coverage.

While adhering to traditional frameworks to establish mechanisms for detecting information security events, key resources can be enhanced through hybridization in the following ways:

1) Threat Landscape Mapping

Organizations must maintain an up-to-date understanding of emerging threats, attack vectors, and tactics by integrating external threat intelligence feeds, vulnerability advisories, and internal incident records to identify patterns and trends. In a hybrid approach, it is recommended to deploy AI systems capable of providing visualizations of attack chains, mapping detected anomalies to established frameworks such as MITRE ATT&CK. This combination allows organizations to benefit from traditional threat indicators while enabling advanced contextualization through AI-driven analysis.

2) Detection and Analysis

As previously discussed, the most powerful application of AI occurs in this stage, particularly in identifying deviations from established behavioral baselines to detect potential threats. However, to maintain unbiased and explainable results, it is critical to retain traditional detection mechanisms alongside AI systems. Traditional tools can detect simple breaches without heavy reliance on AI, serving as baseline monitors to ensure that AI models do not deviate excessively or introduce unforeseen biases.

The hybrid model ensures that organizations maintain comprehensive visibility, robust detection capabilities, and trustworthy outputs, ultimately strengthening the overall security posture.

C. RESPOND AND CONTAIN

The Respond and Contain phase represent one of the most time-sensitive elements of incident handling and benefits significantly from automation. However, as discussed earlier, there are critical scenarios where human oversight must be integrated. A properly designed hybrid system should carefully balance automation with human intervention based on incident severity.

1) Hybrid Approach to Response and Containment (Refer Table III)

- Low-Priority Incidents:

For incidents categorized as low priority, it is ideal to automate the response and containment process entirely, minimizing the need for human intervention. AI systems are particularly effective in tailoring containment strategies based on the specifics of an incident and its predicted propagation model. Full automation in these cases not only speeds up response times but also reduces operational costs over the long term.

- High-Priority Incidents:
  For incidents classified as high priority, it is necessary to involve human responders to validate and adjust the containment strategies. This step ensures that critical business operations are not inadvertently disrupted by overly aggressive automated actions. Additionally, human responders can act as points of contact for escalation, providing critical judgment in serious or ambiguous situations where context beyond algorithmic output is required.

TABLE III. AUTOMATION AND HUMAN INVOLVEMENT STRATEGY ACROSS INCIDENT SEVERITY LEVELS

Scenario	**Automation Level**	**Human Involvement**	**Example Actions**
Low-Priority Incidents	Full automation	Minimal to none	AI autonomously isolates threats, blocks malicious IPs, updates detection models.
Medium-Priority Incidents	Initial automation + rapid human review	Moderate — post-action validation and tuning	AI initiates immediate containment; humans review, adjust, and ensure no overreach.
High-Priority Incidents	Partial automation with continuous oversight	Significant — strategic decision-making and escalation	Human responders approve or adjust AI actions, manage high-risk containment, and coordinate business communications.

2) Visibility and Governance

To ensure transparency and accountability in the hybrid response process, it is recommended to implement centralized dashboards. These dashboards should be provided:

- Full visibility into the entire incident responds to lifecycle.
- Real-time updates on both automated and manual interventions.
- Audit trails that are aligned with organizational policies and regulatory requirements.

Such visibility ensures that all actions, whether automated or human-driven, are monitored, documented, and can be reviewed for continuous improvement and compliance.

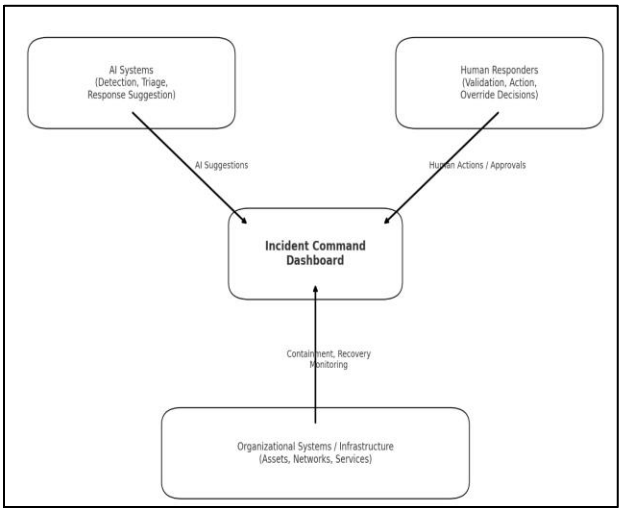

Fig 2 Incident Command Dashboard Integration

D. RECOVER AND REMEDIATE

Swift and secure recovery is essential following containment and eradication efforts. Within the Hybrid Incident Response Framework (HIRF), the integration of artificial intelligence (AI) and machine learning (ML) significantly enhances the efficiency, accuracy, and compliance of recovery operations, particularly in complex and hybridized environments. However, while AI-driven systems can optimize many aspects of recovery, human oversight remains indispensable in critical areas to ensure contextual accuracy, business alignment, and legal compliance.

The following key activities define the recovery and remediation phase:

1) *AI-Assisted Recovery Guidance*

AI-driven analytics are utilized to predict residual risks and recommend optimal recovery strategies. These strategies include identifying appropriate system restoration points, selecting backup reversion tactics, and prioritizing patching sequences. This AI-guided approach enables rapid, targeted remediation, reducing operational downtime and minimizing business disruption.

However, when incidents are prioritized as high-severity or involve critical business operations, human oversight becomes necessary. Human responders must validate AI-driven recovery decisions to ensure that broader business impacts and operational nuances are properly accounted for.

2) *Compliance-Driven Restoration*

Recovery activities must align closely with regulatory requirements (e.g., GDPR, HIPAA) and internal organizational policies. Although automated compliance checks greatly assist in speeding up restoration, human experts, including legal and compliance teams, must interpret complex and nuanced regulatory obligations. Their oversight ensures that all remediation actions meet not only technical requirements but also legal, contractual, and ethical standards, thereby minimizing the risk of regulatory violations.

3) *Business Contextual Decision-Making*

While AI systems may recommend system restorations based solely on technical prioritization (e.g., threat containment, vulnerability remediation), business context must guide final decision-making. Human judgment is necessary to prioritize restoration activities based on business criticality. For instance, critical systems such as payment gateways, client portals, and operational databases may require restoration ahead of technically prioritized systems to prevent revenue loss, preserve client trust, and ensure operational continuity.

4) *Human-Led Communication and Coordination*

Recovery from a cybersecurity incident extends beyond technical restoration; it necessitates deliberate and strategic communication management. Effective engagement with stakeholders—including executive leadership, clients, partners, regulatory authorities, and the public—must be led by human actors to ensure accuracy, credibility, and sensitivity to the broader organizational and societal context. While automation can assist in disseminating updates rapidly, it lacks the nuance, empathy, and situational awareness required for high-stakes communication.

Human-led coordination ensures that messaging is appropriately tailored to diverse audiences, accounts for reputational risk, addresses legal and regulatory considerations, and adapts dynamically as the incident evolves. Consequently, integrating human oversight into communication and coordination processes is essential for preserving organizational trust, maintaining transparency, and supporting long-term recovery efforts.

E. LEARN AND ADAPT

Continuous improvement should be foundational to building cyber resilience. Within the Hybrid Incident Response Framework (HIRF), adaptive learning and structured feedback mechanisms should be embedded to ensure that the organization evolves in response to emerging threats, shifting risk landscapes, and lessons learned from past incidents. This approach should blend automation with human oversight to drive sustained enhancements in detection, response, and recovery capabilities.

1) *Automated Learning*

Machine learning (ML) models should be periodically retrained using data collected from recent incidents. This retraining should refine detection, triage, and response strategies, allowing the organization to adapt dynamically to evolving attack techniques and internal changes in infrastructure and business processes. Automated learning should enable faster updates to threat models without requiring full manual reengineering.

2) *Lessons Learned Integration*

Structured, human-led post-incident reviews ("post-mortems") should be conducted to analyze the root causes, successes, and failures of incident response efforts. Insights from these reviews should inform targeted updates to AI models, incident response playbooks, security policies, and workforce training programs. This process should ensure that both technological tools and organizational processes evolve based on real-world operational experience, rather than theoretical assumptions.

3) *Resilience Metrics*

Key performance indicators (KPIs)—such as detection-to-containment time, false positive and false negative rates, and system recovery speed—should be systematically tracked using both AI-driven analytics and traditional monitoring techniques. These metrics should guide continuous strategic improvements in incident response readiness and provide empirical evidence to demonstrate the effectiveness and maturity of the organization’s cybersecurity program.

4) *Feedback-Driven AI Training*

AI model retraining should not be conducted on a fixed schedule alone but should also be dynamically triggered when specific thresholds are breached. For example, significant increases in false positive/negative rates or the identification of novel attack patterns should initiate a retraining process. This feedback loop should ensure that AI-driven detection and analysis systems remain closely aligned with the current threat landscape and organizational priorities.

5) *Cross-Organizational Learning*

To enhance the collective resilience of the cybersecurity community while preserving confidentiality, anonymized incident data should be shared through federated learning frameworks. This collaborative approach should enable multiple organizations to improve their threat detection models without exposing sensitive internal data, thereby balancing collective intelligence with data privacy obligations.

6) *Hybrid Team Optimization*

AI-based co-pilots should assist human analysts by drafting incident reports, suggesting meeting agendas for post-incident reviews, and proposing preliminary remediation plans. Concurrently, ongoing AI literacy programs should be instituted to ensure that analysts can effectively interpret, audit, and, when necessary, override automated decision outputs. This dual-pronged optimization should foster symbiotic collaboration between human expertise and AI-driven augmentation, strengthening both tactical operations and strategic decision-making.

VI. CONCLUSION & FUTURE WORK

While artificial intelligence (AI) and machine learning (ML) have made significant advancements and offer numerous advantages in the field of cybersecurity, they are not yet at a stage where they can fully replace traditional methodologies. AI-driven systems, although powerful, still face challenges such as bias, explainability, unpredictability in novel situations, and regulatory compliance gaps. Meanwhile, traditional incident response techniques continue to offer strengths in human judgment, contextual understanding, and legal accountability.

This paper proposes the Hybrid Incident Response Framework (HIRF) as a comprehensive reference model that organizations can use to make informed decisions about integrating automation and AI/ML technologies into their incident response processes. By systematically evaluating the trade-offs between AI-driven and traditional methods at each phase of the incident response lifecycle, the framework empowers organizations to design strategies that maximize the advantages of both while mitigating their respective limitations.

The goal of HIRF is to enable the creation of robust, resilient, and adaptive incident response systems. Through the thoughtful blending of human oversight with machine efficiency, organizations can better address the complex and dynamic threat landscape of modern cybersecurity.

While this paper provides a foundational framework for hybrid incident response, several avenues remain open for future exploration and development:

- Empirical Validation: Future research should involve empirical testing of the Hybrid Incident Response Framework across diverse organizational environments to validate its effectiveness, adaptability, and return on investment.
- AI Explainability Mechanisms: There is a critical need for further development of explainable AI (XAI) tools specifically tailored for incident response. Research could focus on enhancing transparency in AI-driven detection, triage, and containment processes.
- Dynamic Role Adaptation: Future work could explore frameworks where human-AI role distributions dynamically shift based on severity, complexity, or type of incident, optimizing resources in real-time rather than using static decision trees.

REFERENCES

[1] SentinelOne, “10 Cyber security trends for 2025,” *SentinelOne*, Apr. 01, 2025. https://www.sentinelone.com/cybersecurity-101/cybersecurity/cyber-security-trends/

[2] A. Nelson, “Incident Response Recommendations and Considerations for cybersecurity Risk Management:,” Jan. 2024. doi: 10.6028/nist.sp.800-61r3.ipd.

[3] “33901.pdf on Egnyte,” *Egnyte*. https://sansorg.egnyte.com/dl/SzUc95nE0x

[4] “ISO/IEC 27035-2:2023,” *ISO*. https://www.iso.org/standard/78974.html

[5] Software Engineering Institute and G. Killcrece, “Incident management,” Dec. 2005. [Online]. Available: https://insights.sei.cmu.edu/documents/406/2005_019_001_295923.pdf

[6] C. M. Patterson, J. R. C. Nurse, and V. N. L. Franqueira, “Learning from cyber security incidents: A systematic review and future research agenda,” *Computers & Security*, vol. 132, p. I. A. Tøndel, M. B. Line, and M. G. Jaatun, “Information security incident management: Current practice as reported in the literature,” Computers & Security, vol. 45, pp. 42–57, May 2014, doi: 10.1016/j.cose.2014.05.003.103309, May 2023, doi: 10.1016/j.cose.2023.103309.

[7] R. Werlinger, K. Muldner, K. Hawkey, K. Beznosov Preparation, detection, and analysis: the diagnostic work of IT security incident response

[8] S. Metzger, W. Hommel, H. Reiser Integrated security incident management – concepts and real-world experiences

[9] C. M. Patterson, J. R. C. Nurse, and V. N. L. Franqueira, “Learning from cyber security incidents: A systematic review and future research agenda,” *Computers & Security*, vol. 132, p. 103309, May 2023, doi: 10.1016/j.cose.2023.103309.

[10] I. Jada and T. O. Mayayise, “The impact of artificial intelligence on organisational cyber security: An outcome of a systematic literature review,” *Data and Information Management*, vol. 8, no. 2, p. 100063, Dec. 2023, doi: 10.1016/j.dim.2023.100063.

[11] M. A. M. Farzaan, M. C. Ghanem, A. El-Hajjar, and D. N. Ratnayake, “AI-Enabled system for efficient and effective cyber incident detection and response in cloud environments,” *arXiv (Cornell University)*, Apr. 2024, doi: 10.48550/arxiv.2404.05602.

[12] I. Ghafir, M. Hammoudeh, V. Prenosil, L. Han, R. Hegarty, K. Rabie, F.J. Aparicio-Navarro Detection of advanced persistent threat using machine-learning correlation analysis Future Gener. Comput. Syst., 89 (2018), pp. 349-359

[13] A. J. G. De Azambuja, C. Plesker, K. Schützer, R. Anderl, B. Schleich, and V. R. Almeida, “Artificial Intelligence-Based Cyber Security in the context of Industry 4.0—A survey,” *Electronics*, vol. 12, no. 8, p. 1920, Apr. 2023, doi: 10.3390/electronics12081920.

[14] L. Qudus, “Resilient Systems: Building secure Cyber-Physical infrastructure for critical industries against emerging threats,” *International Journal of Research Publication and Reviews*, vol. 6, no. 1, pp. 3330–3346, Jan. 2025, doi: 10.55248/gengpi.6.0125.0514.

[15] A. Da’Costa, J. Teke, J. E. Origbo, A. Osonuga, E. Egbon, and D. B. Olawade, “Ai-driven triage in emergency departments: A review of benefits, challenges, and future directions,” *International Journal of Medical Informatics*, vol. 197, p. 105838, Feb. 2025, doi: 10.1016/j.ijmedinf.2025.105838.

[16] National Institute of Standards and Technology, "AI Risk Management Framework (AI RMF) 1.0," NIST, Gaithersburg, MD, USA, Mar. 2024. [Online]. Available: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf

[17] International Organization for Standardization, "ISO/IEC 42001: Artificial intelligence - Management system," ISO, Geneva, Switzerland, 2023. [Online]. Available: https://www.iso.org/standard/83210.html

[18] European Commission, "Proposal for a Regulation of the European Parliament and of the Council laying down harmonized rules on artificial intelligence (Artificial Intelligence Act)," COM(2021) 206 final, Brussels, Belgium, Apr. 2021. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206

[19] I. Mbona and J. Eloff, “Data Sets for Cyber Security Machine Learning Models: A Methodological Approach,” *International Conference on Internet of Things, Big Data and Security*, pp. 149–156, Jan. 2024, doi: 10.5220/0012598400003705.

[20] F. Cremer *et al.*, “Cyber risk and cybersecurity: a systematic review of data availability,” *The Geneva Papers on Risk and Insurance Issues and Practice*, vol. 47, no. 3, pp. 698–736, Feb. 2022, doi: 10.1057/s41288-022-00266-6.

[21] R. Schwartz, J. Dodge, N. A. Smith, and O. Etzioni, “Green AI,” *arXiv (Cornell University)*, Jan. 2019, doi: 10.48550/arxiv.1907.10597.

[22] Jia, Z., Chen, J., Xu, X. et al. The importance of resource awareness in artificial intelligence for healthcare. Nat Mach Intell 5, 687–698 (2023). https://doi.org/10.1038/s42256-023-00670-0

[23] M. Soni, A. M. Pawar, A. Godbole, A. Sharma, S. A. Tiwaskar, and C. D. Kokane, “AI-Driven Incident Response Systems for Cybersecurity: A Hybrid

approach," in Smart innovation, systems and technologies, 2025, pp. 537–547. doi: 10.1007/978-981-96-0143-1_43.

[24] M. Mersha, K. Lam, J. Wood, A. K. AlShami, and J. Kalita, "Explainable artificial intelligence: A survey of needs, techniques, applications, and future direction," *Neurocomputing*, vol. 599, p. 128111, Sep. 2024, doi: 10.1016/j.neucom.2024.128111.

[25] A. K. Reddy, "AI vs. Traditional IDS: Comparative Analysis of Real-World Detection Capabilities," Dec. 31, 2024. https://ijritcc.org/index.php/ijritcc/article/view/11463

[26] B. Xi, "Adversarial machine learning for cybersecurity and computer vision: Current developments and challenges," *Wiley Interdisciplinary Reviews Computational Statistics*, vol. 12, no. 5, Apr. 2020, doi: 10.1002/wics.1511.

[27] E. Ferrara, "Fairness and Bias in Artificial intelligence: A brief survey of sources, impacts, and mitigation strategies," *Sci*, vol. 6, no. 1, p. 3, Dec. 2023, doi: 10.3390/sci6010003.

[28] "View of AI vs. Traditional IDS: Comparative Analysis of Real-World Detection Capabilities." https://ijritcc.org/index.php/ijritcc/article/view/11463/8800

V. P. S, "How can we manage biases in artificial intelligence systems – A systematic literature review," *International Journal of Information Management Data Insights*, vol. 3, no. 1, p. 100165, Mar. 2023, doi: 10.1016/j.jjimei.2023.100165