*International Journal of Research in Engineering and Science (IJRES)*
*ISSN (Online): 2320-9364, ISSN (Print): 2320-9356*
*www.ijres.org Volume 13 Issue 6 ǁ June 2025 ǁ PP. 311-320*

# AI-Augmented Threat Intelligence in Zero-Trust Architectures

Olubunmi.Famosinpe
*McCombs School of Business*
*The University of Texas at Austin*

***Abstract***
*As cyber threats grow more sophisticated and distributed, conventional perimeter-based defences have become obsolete. Devices such as. This paper proposes an AI-augmented threat intelligence framework integrated into zero-trust architectures (ZTA) to proactively detect, correlate, and respond to threats across heterogeneous enterprise environments. Leveraging large-scale log telemetry, the model employs deep learning, contextual embedding, and correlation scoring to prioritise risks in real time. The framework incorporates dynamic trust recalibration, enabling adaptive policy enforcement based on evolving user, device, and workload behaviour. To demonstrate feasibility, the study implements a prototype using transformer-based models. It validates it against a hybrid network dataset, showcasing superior detection accuracy, reduced false positives, and minimal latency. The findings suggest that coupling AI-driven inference with zero-trust principles not only enhances visibility and response agility but also sets a scalable foundation for autonomous security in modern cloud and edge deployments.*

***Keywords:*** *Zero-Trust Architecture (ZTA), Threat Intelligence, Deep Learning, Transformer Models, Adaptive Access Control, Cybersecurity Automation*

---

Date of Submission: 14-06-2025 Date of acceptance: 29-06-2025

---

## I. Introduction

Internet technology is evolving rapidly. Now, end-systems of the internet include not just computers, but also devices such as wallets, cars, and motherboards. This explosion of end-systems, including smartphones, sensors, and webcams, has been revolutionary. As the Internet of Things (IoT) expands, so do security challenges, rendering existing security measures ineffective. Traditional mechanisms, which inspect all traffic at the network perimeter or connect locks to specific objects, can no longer be upheld. In past decades, perimeter-based security served as a protection bucket, complicating infiltration from outside adversaries. Subsequently, some methods allowed trusted packets through this bucket based on coarse-grained trusted network borders, but this was not a practical solution (Malhotra et al., 2021). Years of post-attack analyses have shown that most security breaches occur due to implicitly trusted individuals within the enterprise. Traditionally, enterprises have used Role-Based Access Control (RBAC), where roles, such as manager or accountant, determine authority. Zero Trust Architecture operates on the principle of never implicitly granting trust, continuously evaluating trust parameters for each resource request through strict access control protocols. Access is based on Attribute-Based Access Control (ABAC), which grants or denies requests based on enterprise-specific attributes. This results in reduced implicit trust zones and increased granularity in access control. Trust scores can vary by network and can be adjusted as needed (Khan, 2023). Two significant issues arise when implementing EDR: the immense volume of real-time data generated by EDR systems, and a high rate of false alarms. Continuous attribute recording is crucial as threats like APT require ongoing evaluation. Tactical Provenance Analysis examines the temporal and causal order of threat alerts within the Tactical Provenance Graph to identify APT attack sequences and reduce non-usable paths in threat detection (Kaur et al., 2024).

## II. Understanding Zero-Trust Architectures

Network security is one of the most vital provisions organisations have to protect their assets and recover from unintentional damage (Alevizos et al., 2021). In the past, organisations attempted to safeguard their networks through perimeter-based security mechanisms. However, with relationships becoming increasingly intricate, employees working remotely, and the number of remote devices surging, organisations began to "trust but verify" access to their networks. This concept led to several efforts, such as Continuous Adaptive Risk and Trust Assessment (CARTA) and Detection-in-Depth (DiD). However, all these efforts retain one core element: the existence of trusted or trust zones within systems. Although these initiatives have proven successful, they continue to face challenges in the modern technological era (Sindhu & Vinay, 2025). Firewalls and traditional network

security protocols are necessary yet insufficient components in devising a cybersecurity framework (Kang et al., 2023). The dynamic nature of the remote workforce, the rise of the Internet of Things (IoT), and the mobilisation of threats necessitate a transition to a new security paradigm. The zero trust architecture (ZTA) offers a new security framework, grounded in the idea that no one should be implicitly trusted. Zero trust shifts the focus from network security to data, requiring continuous, adaptive verification of all access requests. ZTA protocols implement a system hosting deperimeterisation, where both authenticated and unauthorised users can access the enterprise network. To mitigate zero trust challenges, it is essential to augment ZTA components with adaptive and rigorous detection and response capabilities.

Obtaining a trust level for subjects, such as users, devices, and processes, is the pivotal first step in enhancing Zero Trust Architecture (ZTA) through machine learning-based detection. Furthermore, monitoring subjects' activities enables the investigation of competence and intent, as well as the discovery of intrusions. Moreover, incident response is also crucial to contain the impact of intrusions. Nonetheless, existing solutions for the former two aspects require arrangements that are out of reach for most organisations. In addition, rapidly developing threats and infrastructure changes render existing solutions partially ineffective. AI is a force multiplier in cyber defence, having impacted threat detection, incident investigation, and forensics, such as through the analysis of logs and network flows (Kolade et al., 2025).

### 2.1. Definition and Principles

Under Zero Trust Architecture (ZTA), each endpoint is considered a potential threat, necessitating enhanced security measures for endpoints. Frequent Advanced Persistent Threats (APTs) target enterprise networks, resulting in significant economic losses. ZTA offers a solution by providing a new paradigm for enterprise network security against APTs and lateral movement threats. This research analyses the enhancement of ZTA for endpoint devices through endpoint detection and response (EDR) systems and blockchain technology (Alevizos et al., 2022). A Brief History of "Zero Trust" and Zero Trust Architecture: In 2004, the Jericho Forum introduced de-perimeterization, which evolved into zero trust. This concept recognised the impracticality of strong external perimeters amid weaker internal security. J. Kindervag coined the term "zero trust" in 2010; however, the concept existed in cybersecurity prior to this (Kumar et al., 2022). The U.S. Department of Defence and Defence Information Systems Agency (DISA) proposed a secure strategy, "black core," in 2007, focusing on securing transactions rather than perimeters. The U.S. Department of Defence removed the first unencrypted data network ("black core") and enforced strict access control for transactions (Adamson & Qureshi, 2025). In 2014, the National Institute of Standards and Technology (NIST) published SP 800-27, emphasising access control as a security principle, refining the black core idea. The core tenet, Never Trust, Always Verify (NTAV), implements least privilege policies (Alevizos et al., 2021). In March 2022, the Biden-Harris administration's Cybersecurity Strategy urged a shift from a fortress mentality to an "assume breach" mindset, aiming for resilient and adaptive security systems to protect data and resources.

### 2.2. Components of Zero-Trust

Zero Trust architecture is a security framework based on the principle of "never trust; always verify" to protect networks, endpoints, and resources. Key components include information flows, which involve the transmission of policies from management security components (MSE) to enforcement security components (ESE), as well as the transmission of observability information from ESE to MSE, and orchestration among ESE components (Alevizos et al., 2021). These flows utilise specific sequences of protocol-compliant messages for left-to-right (MSE to ESE) and right-to-left (ESE to MSE) directions (Kumar et al., 2022). Components of Zero Trust architecture include MSE, ESE, technical security enforcement components (TecSE), endpoints, resources, and critical infrastructure. MSE comprises Policy Management, Compliance Management, Visibility Management, Incident & Event Management, and Operation Management. ESE consists of Network, WAN, Cloud, Endpoint, and Application security. The endpoints and resources include internal and external users and resources. Critical infrastructure encompasses the control centre, stationing and transportation equipment, process control systems, and networks. A brief description of ZTA components is provided below regarding ZTA integration (Khan, 2023). MSE implements policies for ESE and TecSE in accordance with management security protocols (M-CP) and monitors compliance across various layers, adhering to compliance management protocols (C-CP). Each MSE coordinates observability information from enforcement components, presenting a unified view for improved observability management (V-CP). MSE consolidates and simplifies threat information from event logs provided by enrollment components (Th-CP). MSE also conducts orchestration protocols among enforcement components for automated network protection (O-CP), ready for execution during incidents (Ahuja, 2024).

### 2.3. Benefits of Zero-Trust

The adoption of a ZTA will yield numerous benefits for organisations and their networks. It will enhance the organisation's security posture and improve its resilience to advanced and zero-day attacks (Kang et al., 2023). Some of the specific benefits of ZTA are as follows:

1. Full Visibility: Visibility is one of the central tenets of ZTA. Strict enforcement of the principle of least privilege, continuous adaptive risk analysis, and full-throttle inspection of data packets, users, devices, and applications are the significant ZTA capabilities that contribute to complete visibility. Greater visibility into all assets significantly enhances security analysts' ability to analyse threats on the network and discover patterns or anomalous behaviours.
2. Improvement of Security Posture: A ZTA significantly enhances the security posture and reduces the risk of breaches within an organisation. Enforcing the principle of least privilege and implementing device trust and user trust mechanisms makes it more difficult for insider attackers to gain a foothold. Moreover, ZTA controls enough information flows to segment them. This prevents lateral movement and provides analysts with sufficient visibility into the network's communication channels, making it easier to detect data exfiltration.
3. Enhanced Resilience: ZTA solutions primarily target insider attacks. They significantly enhance an organisation's resilience to advanced and zero-day attacks. Advances in technical capabilities typically have a disproportionate impact on the effectiveness of attacking forces in terms of the cost-benefit ratio. Emerging technologies, such as AI, result in broader applications of attack techniques, as secure coding practices become less effective in generating syntactically correct and semantically sound code. Complex systems are attacked in commodity ways by sophisticated attackers.
4. Reducing Insider Attacks: Both the layered security and defence-in-depth paradigms are concerned with the defence perimeter, leading to many isolated security layers. As defences evolve independently, there are seldom efforts to examine the security gaps and eliminate them. Hence, an ongoing and growing concern regarding insider threats is the security gaps in trust, trust assumptions, and trust itself. Adversaries able to leverage these weaknesses will steal sensitive information.
5. Reduced Propagation of Malware: There has been a growing use of malware called 'lateral movement' or 'worm' to 'jump' from one target machine to another. As a zero-trust architecture denies the trust assumption of any account or device on the network, a compromised endpoint's authenticated and authorised session can perform limited activities. The primary infected machine will not be able to propagate the infection to its peer machines.

## III. Threat Intelligence Overview

The recent rise in the number of companies switching to a zero-trust architecture indicates a shift in reality in terms of networking and security. The current technological paradigm in today's world has evolved to make networking easier while ensuring a better user experience for the end user. Nonetheless, this shift has opened the gates for malicious activities and has rapidly increased the threat vectors from which enterprises and individuals need to be protected (Kumar et al., 2022).

Whether the targets are persons, corporations, or even companies, threats are still in motion for all kinds of assets. Today's attackers have become dynamically more competent and prepared, making their job harder and rendering the hard work of a sound security engineer ineffective. Most enterprise attacks are covert, lasting for extended periods until access to a critical infrastructure is acquired and sufficient footholds are established to bring the entire organisation down. Initial Data Exfiltration Pre-attack and the remaining stages of an informative Pre-attack proceed over several months without any common signature. Moreover, the attackers take utmost care in dumping logs and obfuscating actions to avoid arousing suspicion (Alevizos et al., 2021). Furthermore, zero-trust architecture undertaken in enterprise networks is based on a dynamic query that is allowed. Initially, the raw attributes are transformed into trust scores by an aggregation processor in composite and standard terms, and the trust score aggregates the importance of the referred click alerts proportionally. This ensures that the more credible and reliable peers concurrently offer click items, the greater influence they have on the incoming query. Subsequently, regarding users with similar and high cluster trust scores, the management agent determines the trust scores of the newcomer in single-domain networks. Hence, in security, there is essentially a need for assurance that nodes do not exhibit pathological behaviour concerning a zero-trust architecture (Khan, 2023).

### 3.1. Definition and Types of Threat Intelligence

Cyber Threat Intelligence (CTI) provides knowledge about existing or emerging threats that justify the risks organisations take to achieve their objectives. Evolving from military intelligence, CTI is now crucial for cyber defence. There are various intelligence types for different audiences, such as policymakers and risk managers (Strategic CTI), operational intelligence for defenders and incident response teams, and technical intelligence for system configuration and configuration management. Tactical intelligence, or Tactics, Techniques, and Procedures (TTPs), offers guidance to security operations managers on preventing system compromise or data exfiltration. Threats can take multiple forms, making it impossible to maintain constant threat

intelligence on all attack avenues (Tatam et al., 2021). CTI solutions typically provide risk alerts after attacks occur, creating a linear model that must revert to a circular format for ongoing effectiveness. The threat landscape includes bitcoin miners, passive observers, and Script Kiddies, exploiting zero-day vulnerabilities while using various CTI solutions. Their poor system monitoring exposes them to Preemptive CTI, as indicated by FTO and ASCI.

This article categorises threat intelligence by collection technique, source, assessment type, and audience. Assessment types include technical versus business-impact, strategic versus tactical, and forecast versus real-time intelligence. Intelligence sources vary, and collection methodologies include periodic scans, Capture the Flag (CTFs), and honeypots. Data types categorised include network intrusion detection data, firewall logs, and application server logs (Sun et al., 2023). Cyber threat intelligence involves understanding existing, emerging, or evolving threats. Threat intelligence often follows the Intelligence Cycle, starting with a focus on defining objectives. The second stage, collection, gathers as much information as possible. The intelligence process becomes unmanageable without filtering and pre-processing. The third stage, identification, filtering, and analysis, refines information into a useful product, filtered based on established parameters. This stage produces a summarised intelligence report, often in PDF format, containing fundamental-time tools, data continuity functions, and a known vulnerability finder for web applications (Mavroeidis & Bromander, 2021). This report is static and targets a broad audience.

### 3.2. Importance in Cybersecurity

The shift to remote work in recent years has turned many users, servers, and devices into threats. As a result, discussions have been held among security professionals to adopt frameworks and architectures that secure networks and information systems. Some prominent frameworks and architectures include the Zero Trust Architecture (ZTA), the strategy proposed by Black Core, and a framework for implementing ZTA proposed by the US National Institute of Standards and Technology (NIST). The NIST document serves as an excellent starting point, as it provides a detailed explanation of ZTA and its components (Alevizos et al., 2021). Zero Trust Architecture represents a distinct approach to thinking about and securing information systems, networks, and the broader cyber realm. The legacy perimeter-based security strategy was primarily focused on protecting the network perimeter; however, with the advent of mobile users, IoT devices, cloud service providers, and other emerging technologies, the perimeter has become obsolete. Consequently, a 'de-perimeterization' strategy was proposed, with Zero Trust as one of the most prominent examples that has gained traction with increasing urgency (Kumar et al., 2022). An organisation and its trusted environment can be regarded as a socio-technical system, facing numerous potential dangers presented by insiders (malicious or negligent employees) and outsiders (hackers, adversaries, or criminals). ZTA sets up strict controls on all assets requesting access, regardless of their location. No assets (users, servers, devices, and more) are trusted by default. Everything must be explicitly trusted. In practice, for a ZTA design to be effective, it must be strictly enforced throughout the network; no interaction should be allowed between resources that have not been explicitly and fine-grainedly provisioned permission to access each other.

## IV. AI in Cybersecurity: The Role of AI in Threat Detection

Organisations face rising cyberattacks and breaches. The Colonial Pipeline attack led to increased fuel prices and halted deliveries. In 2021, over 43 billion records were compromised, averaging more than 100 breaches daily. The average U. S. data breach cost exceeded $10 million, with detection and containment averaging 327 days (Schmitt, 2023). This section outlines AI-enabled cybersecurity solutions linked to AI-Events for threat mitigation. A key trend is the adoption of zero-trust architecture, which inspects every user accessing a network, much like a police checkpoint. It operates on two principles: distrust everything and verify all. This model acknowledges that preventing attacks is impossible and focuses on detection and identification. Zero trust enables organisations to create multiple firewalls while emphasising a single ingress point (Khan, 2023). Many security systems still rely on perimeter defences, trusting internal entities. Zero-trust security facilitates granular user access through segmentation, thereby hindering lateral movement. This paper outlines the core elements of zero-trust architecture and includes a case study on its deployment. Traditional systems rely on constant rules, effective in some cases but not for the entire decision tree (Bernardez Molina et al., 2023). Organisations often combine systems for making aggregated decisions, but at the expense of lower efficiency. AI analyses vast event volumes to provide optimal suggestions instead of constant predictions. Neural networks can feature counts from input data months in advance, suggesting actions to mitigate impacts. AI can manage incoming information, cover data sets, and adapt evaluations. Automation can be based on actuation models, leading to adaptations in system rules. External AI scripts can adjust filtering based on decisions that influence delivered information and detected targets. Zero-trust architectures assume internal breaches. As the perimeter secures, information enters through various filters. AI can classify internal actions into usable and non-usable categories, enhancing filtering and mapping of undetected transactions. Traditional security systems are widely implemented, actively protecting

networks with additional cloud- and hardware-based devices. Analysing unstructured data benefits from log mining and observability-based learning. Deep learning automatically extracts features from raw data, eliminating the need for specific preprocessing (Syed et al., 2022).

### 4.2. Machine Learning Algorithms: AI in Predictive Analysis

Anomaly detection in intrusion detection systems (IDS) utilises machine learning algorithms —including supervised, semi-supervised, and unsupervised — to model digital behaviour (Malele & E. Mathonsi, 2023). These models classify activities as normal or anomalous via a detector and behavioural input model. Supervised multiclass learning is susceptible to evasion or model-inference attacks (Pauling et al., 2022). Evasion attacks add noise to packet payloads or modify headers, hindering traffic classification and reducing the differences between predicted and actual classes when the model reflects underlying data. Model- inference attacks disclose neuron amplitudes and locations, complicating detection by extracting input probability distributions and creating adversarial examples (Maseer et al., 2021). Most applications rely on supervised learning or unsupervised anomaly detection, but they face challenges such as overfitting and adversarial attacks that compromise packet classification. New AI techniques propose a hybrid supervised-unsupervised detection architecture that filters unknown regular classes and effectively classifies anomalies from feature-engineered packets. This method clusters adversarial examples for testing and protection against transformation-based attacks, while a detection explanation network interprets the results from advanced deep models (Azab et al., 2024). AI models for predictive analysis can anticipate cyberattacks, with the use of external data gathering enhancing threat intelligence more effectively than internal data. Organisations may use established AI models to explore new threats (Dhir et al., 2021). Identifying data sources is vital for enhancing threat intelligence in a complex cyber landscape. Predictive analysis models generally use time series or causal inference approaches. Time series models forecast events without considering their causes, while causal models provide insights into potential attacks (M. Soliman et al., 2021). Integrating these insights into threat intelligence reports is essential for internal communication, requiring customised data collection pipelines and disaster recovery options. However, model complexity may hinder analysts from maintaining context, especially when storing network traffic logs with varied logic between weekdays and weekends, which can potentially lead to irrelevant or unexpected outputs.

## V. Integrating AI with Threat Intelligence

Integrating AI with threat intelligence enables security teams to identify potential threats before they are executed. Threat analysts can encode attack tactics, techniques, and procedures (TTPs), along with behaviours, into the AI engine. Existing attack simulation frameworks generate threat simulations, allowing the implementation of advanced threat intelligence capabilities. AI-Augmented Threat Intelligence: Today's overwhelming amount of cyber threat information inundates threat hunting teams with intelligence from multiple sources, increasing the risk of missing devastating attacks or delaying the implementation of mitigation techniques. Enterprises must adopt AI-augmented threat intelligence that automatically synthesises threat data into threat vectors, which can be used with security tools to initiate hunting missions. Experienced threat hunting experts are then involved in threat modelling and scenario generation, which is often manual and time-consuming (Schmitt, 2023). Auto Threat Intelligence Sourcing: A key challenge in curating threat intelligence is striking a balance between not overlooking valuable sources and avoiding an influx of low-quality information that clutters workflows. Automated threat intelligence sourcing aims to identify relevant cyber threat domains for various organisations. By utilising public datasets, a quality analysis model can assess the characteristics of threat intelligence sources. Meta-regression analyses establish the relationship between characteristics and rating scores, allowing systematic filtering. Social network analysis techniques, such as community detection, can group intelligence sources based on their interconnections. Additionally, the literature is increasingly exploring data mining methods to extract intelligence from unstructured data sources.

### 5.1. Data Collection and Analysis

Automation through Artificial Intelligence (AI) has emerged as a viable solution for addressing the challenge of detecting Advanced Persistent Threats (APTs) in enterprise and government networks. APT detection has evolved into an ongoing arms race between malicious actors seeking to compromise these networks and defenders working to prevent breaches through various security technologies. Following a successful breach, security analysts must sift through extensive outputs from security tools, alongside internal, external, and ad-hoc intelligence, to identify signs of an Advanced Persistent Threat (APT). As the number of security tools increases, so too does the already substantial output of security alerts and log entries requiring review (Kumar et al., 2022). This surge in data creates issues of information overload and places critical alerts at risk of being overlooked. Moreover, during a breach, a truly enterprise-scale investigation can resemble searching for a needle in a haystack of billions of logs. The scalable automation of related processes is crucial for addressing these challenges and enabling an effective response to APTs.

In the proposed control method, data is initially collected from EDR systems and security devices that detect security incidents within the network, and stored in an external database. The data is subsequently parsed to extract the temporal and causal ordering of related incident alerts from various sources, and is represented as a Tactical Provenance Graph (TPG). The user may select a timeframe for analysis based on when they suspect the attack may have commenced. By evaluating the alerts in the TPG, beginning from the leaf nodes that were first emitted, a subset of the alerts can be verified as attack actions, whilst others can be identified as everyday activities. These activities are then represented as a sub-graph and output devoid of alerts associated with everyday activities (M. Soliman et al., 2021). Ultimately, it is anticipated that a subsequent step will align with the security operations staff and corporate processes within the enterprise, and two additional datasets incorporating both security alerts and open-source intelligence suggestions for defending against common exploits will also be developed.

### 5.2. Automated Threat Response

In the previous sections, we discussed multiple use cases of AI-augmented threat intelligence in the context of the Zero Trust architecture. However, there are additional use cases to address. One such use case is automated threat response, which is crucial in addition to just detection (M. Soliman et al., 2021). Through automated threat response, instant and efficient action can be taken to control an incident while reducing the workload on security analysts. As attacks occur, stop the execution of malicious and suspicious network flows, isolate infected hosts from the network, or temporarily block endpoints from accessing a service. These are just a few strategies that can be automated and augmented by AI. In addition to the general automated response strategies, a more targeted response strategy can be designed based on the augmented threat intelligence provided by the AI Shield. For example, in a spear-phishing case where a malicious email is detected, blocking the first clicked URL or the first file download from the corresponding email would be a strategic and effective response. This can also apply to other, more complex attack scenarios where AI-augmented use cases are involved as part of the incident identification process. AI can assist in both detecting the stages reached in the Kill Chain and generating a list of returned commands to block operations at each stage (Kumar et al., 2022). This would significantly enhance the effectiveness of the response process, protect victims before losses occur, and reduce the workload on security analysts.

### 5.3. Enhancing Decision-Making

Decision-making is crucial for AI in business and government. In cybersecurity, AI assists threat intelligence analysts in real-time, enabling quicker, better-informed, and more effective decisions. For over twenty years, AI has detected spam and phishing through human-assisted threat intelligence. In a Zero Trust Architecture (ZTA), every endpoint device must contribute to threat intelligence (Alevizos et al., 2021). Analysts rely on AI to sift through countless security alerts. AI supports these analysts by visualising alerts, guiding analysis and triaging. Cyber defenders tackle billions of alerts, where AI reduces dimensionality and conflict (Kumar et al., 2022). AI systems have worked to map normal and anomalous behaviours from raw signals, identifying key patterns. False positives mislead analysts away from vulnerabilities. Algorithmic game-theory AI can assist both defenders and attackers. The aim of machine intelligence mimicking human intelligence is expressed through value learning and gaming. However, current AI decision systems struggle to understand human intentions. Developing high-fidelity decision models using interpretable AI is a significant challenge. Draft principles of privacy, trust, and provenance need generalisation and integration. Alternatively, co-building human-agent decision forums offers immediate benefits. The convergence of security, safety, reputability, and assurance across various domains highlights the importance of intelligent strategies for diverse decision-makers.

## VI. Challenges in AI-Augmented Threat Intelligence

Artificial Intelligence trustworthiness encompasses cybersecurity, transparency, robustness, accuracy, data quality, governance, human oversight, and record-keeping (Polemi et al., 2024). Risk management encompasses identifying, analysing, estimating, and mitigating threats to ensure AI systems are trustworthy. Understanding relevant threats and risk causes is essential for effective risk management. A thorough grasp of the threat landscape enables vulnerability analysis and the development of targeted controls. Trustworthiness in AI systems hinges on recognising risks associated with human factors. Human biases can influence AI algorithms, resulting in biased outcomes in decision-making. Errors in AI design, development, and deployment can introduce vulnerabilities. AI systems may exploit weaknesses in perception, reasoning, and actions. Concepts like IT security may not efficiently translate to safety assurance for advanced AI algorithms and systems. Some threats are linked to specific dimensions of trustworthiness, while others arise from multiple, interconnected dimensions. Research across these dimensions is limited, with some areas still in the early stages. For instance, the effectiveness of cybersecurity controls depends on addressing causal factors in human and other dimensions. Interactions among threats and the relationships between threat incidents are under-analysed. Moreover, the term ‘AI Augmented Security’ lacks a clear definition. Defining this term is crucial for clarifying context and assessing

various definitions, particularly given the rapid adoption of AI in cybersecurity detection and prevention technologies, as well as its anticipated evolution in areas such as threat intelligence and incident response.

### 6.1. Data Privacy Concerns

Data privacy should be planned and approached as a landscape of risks and technologies that can help mitigate them. Privacy-preserving technology analysis and policy decisions regarding its implementation should consider, amongst other things, values, ethical limitations, and mitigation strategies with implications for the ways individuals and groups can exercise privacy. Privacy-preserving technologies (PPTs), suitable for different contexts and functions, are being researched and developed. The PPTs relevant to this paper are those that attempt to guarantee privacy at the data level, such as obfuscation, k-anonymity, and cryptographic techniques (Radanliev et al., 2024). CTI is a dataset consisting of information on cyber threats across various classification levels, which can be obtained from multiple sources and presented in multiple formats. In modern cyberspace networks, which are vast, constantly evolving, and thus generally poorly understood, manually ingesting and analysing CTI is beyond human capability. There are cases where this has proven lucrative for cyber frauds. The Equifax breach was primarily based on disregarding a known vulnerability in Struts, a framework used to build web applications. This vulnerability had a four-month time-to-patch; however, more than half of the targets did not implement the patch (Astaburuaga, 2022). Conglomerate intelligence can be seen as both a danger and an opportunity. Gathering, sharing, and analysing CTI enhances an organisation's understanding of threats and capabilities against it. Encrypted machine learning, where algorithms are trained on encrypted data without its decryption through multi-party computation, homomorphic encryption, and differential privacy techniques, is one area where research and application can enable privacy-preserving analysis in collaboration between organisations. In CTI sharing, personal data is generally not used. Instead, the means of intelligence generation may include sensitive data, such as service usage logs, data breaches, and emails of victims and suspects.

### 6.2. Bias in AI Algorithms

AI/ML algorithms require unbiased, relevant training data (Englert & Muschiol, 2020). Many AI applications suffer from biases reflecting societal imbalances, potentially leading to unjust discrimination against specific population segments. Two main types of bias exist: syntactic and semantic. Syntactic bias arises when mathematical operations influence the probability of values, whereas semantic bias occurs when a relevant feature is omitted. Detecting these biases using training data statistics is essential (Bohdal et al., 2023). Models can show similar accuracy yet remain unfair due to disproportionate representation of patterns among sensitive subgroups like gender or race. Underrepresented subgroups in training data may lead to misrepresentation in predictions. A case study on a university admission support system utilises a synthetic data generator that reflects real historical data to address bias against minority groups. This AI tool generates realistic training data from corresponding statistical values. Alternative fault detection strategies ensure high data quality for future generations. Following model training with synthetic data, strategies for mitigating bias are discussed. Risk assessment must include algorithmic analysis to identify actual bias arising from external factors. An evidence chain visual analytics framework is proposed for detecting bias in machine learning models. This also addresses operator behaviour modelling with heuristics for understanding causal relationships between actions and prediction bias, aiding high-level model interpretation. A pilot study demonstrates the effectiveness of the unbiased risk assessment toolkit and interpretable AI model behaviour through causal analysis.

### 6.3. Integration with Existing Systems

A typical threat intelligence structure exists, integrating with current platforms for abstraction that checks JSIGs, MHSUs, and other sources. It assembles intelligence intersections within OSINT and annotates them. A tagging system is implemented, allowing UI manipulation of emergent elements. Perceptual agents monitor tagged elements to generate alerts based on time, changes, or proximity to policy breaches. Better assessment of JSIGs and MHSUs aligns with a broader project on platformization, framing threat media as assessable objects across IT levels. Off-the-shelf logging systems compatible with JSIG can be leveraged. High-frequency logging is impractical for fluid data pools, and client endpoint logging may obscure intent. Filtering data points with a preference algorithm can enable in-depth analysis despite large data volumes. This structure would model interest states across analysts, allowing RSS feed pipe cleaning to yield meaningful data for pressure point stimuli. The Wide Area Network (WAN) must evolve beyond current generation capabilities to adopt pass-through systems that quickly process thousands of new MHSUs via a cloud-based logging pipeline. Interest and time filters for log analysis can extract information for blocking purposes. This shift will enhance edge point management in a growing security architecture as the network transitions to second-generation edge points.

## VII. Case Studies

This section summarises case studies, evaluating their strengths and weaknesses while suggesting improvements for AI user processing. The first use case, a regional state agency, focuses on threat detection using classified information from multiple intelligence sources, as well as unstructured data such as the dark web and RSS feeds. The goal is to identify threats by integrating data, extracting knowledge, and classifying information. A subsequent case study on a smart city for national defence examines symptom checking and intelligence extraction from three emerging cyber-physical technologies. The proposed solutions are conceptual, with potential real-world applications based on AI techniques for automated data trading. Use case 1 involves the agency combining advanced threat detection based on classified information from four intelligence agencies, utilising an unclassified perimeter network with unstructured sources. A conceptual solution utilises the Intelligence Cycle and the DIKW pyramid, detailing processes that include how AI methods, such as natural language processing, can assist. However, neither the use nor the processes for these methods are defined in terms of user needs, raising concerns over user-processor effectiveness. The following use case addresses this by evaluating the strengths and weaknesses of each case study, suggesting data types or techniques for improved processes. Given the complexity and variety of data sources, use case 1 requires a methodically planned automated solution. An excess of data complicates solutions, underscoring the need for a hierarchy in containment, moderation of information abundance, and precision in data dissemination and monitoring following public availability. Information can be mapped with possible characteristics, including a tracker for new data notifications and processing intentions or confidence levels of communications.

### 7.1. Successful Implementations

Zero Trust Network Access (ZTNA) is a security framework based on the principles of continuous trust verification, which will fundamentally change the way enterprises secure their networks and resources. This exponential increase suggests that the economic and reputational implications of digital security violations are becoming increasingly apparent. Furthermore, with the ongoing COVID-19 pandemic prompting a shift to remote work, ZTNA adoption is becoming increasingly prevalent, moving the perimeter away from a traditional organisation-centric view. As cyberattacks become more sophisticated and organisations become increasingly complex, legacy security architectures become brittle and ineffective (Kumar et al., 2022). Soon after the release of Google's BeyondCorp architecture, enterprises began to move towards the ZTNA framework. Leaders in this space for ZTNA services are putting their best foot forward to reach prospective clients more quickly with a range of technologies, services, and expertise. This severe talent and experience gap, however, can lead to complex implementations without sufficient consideration of fundamental assumptions. It is paramount to ensure that ZTNA architecture implementations are successful. Early adopters of new technology tend to be primarily large corporations, including telecommunications companies, security firms, and health insurers. Social media has become a platform where insights and information are disseminated regarding motivations, implementations, success, and failure factors (Strandell & Mittal, 2022).

With numerous proposed zero-trust architectures differing in their strengths and weaknesses, some components are universally required. Because any externally facing server must be trusted, it is crucial to recognise the importance of establishing a trusted baseline for each technique in the zero-trust design. It is also critical to have visibility within the ultimately trusted managed environment, as complex chains of trust and trust handoffs can easily obscure malicious activity. AI-augmented Threat Intelligence based on this understanding is fundamental to a successful zero-trust architecture implementation. Furthermore, this ZTNA architecture with AI-augmented TI is thorough and core. A prototype implementation in an external environment demonstrated initial use cases, thus validating the approach.

### 7.2. Lessons Learned from Failures

The White House Office of Management and Budget initiated a three-phase plan to migrate federal systems to zero-trust architectures (ZTA) across all government levels. This move aims to address the technical challenges and potential operational failures associated with current technologies. Improved access control, enhanced risk visibility, and strict adherence to the principle of least privilege will strengthen enterprise network security against adversarial threats. However, ZTA alone will not resolve all security issues (Strandell & Mittal, 2022), and if poorly planned, may introduce new vulnerabilities and compliance risks by bypassing established workflows. It is crucial to recognise brutal truths validated by historical lessons (Kumar et al., 2022). Many services, agencies, and industry partners employ a patchwork of technologies and frameworks to address various problems. The trend of identity-based networking might complicate networks further, requiring careful planning before rash decisions based on fleeting trends. This section reviews key ZTA risks identified through diverse sources. Assuming ZTA can completely replace current technologies and frameworks in a mature enterprise is unrealistic—it could cause catastrophic failures and widespread confusion. Despite decades of technological advancements, no "silver bullet" has emerged to eradicate comprehension, technological, or operational gaps. The

diversity and age of legacy systems present significant challenges in creating an interoperable model for sharing valuable information across disconnected networks. Partners governing a fully federated ZTA could devolve into chaotic decision-making, led by compromised identities, which undermines functionality and knowledge. Poor federation choices for inter-service and agency connections may result in overwhelming control measures that negate any security benefits gained from hastily implemented access controls across federated enterprise services.

## VIII. Future Trends in AI and Zero-Trust

The zero-trust approach (ZTA) has gained worldwide attention as a new security paradigm to mitigate breaches in today's networks, systems, and applications (Alevizos et al., 2021). Unlike traditional approaches focused on perimeter security, ZTA removes implicit trust from the security architecture. Secure access to application services cannot rely on a single solution; instead, organisations must deploy multiple solutions tailored to unique use cases. Combining techniques harmoniously is necessary due to numerous IT threats. AI has historically augmented cybersecurity and can operate ZTA at scale. When combined with AI, ZTA can reduce the attack surface and prevent breaches by detecting abnormal behaviours in real-time, correlating alerts, reducing false positives, optimising rules, and providing insights into incidents. Innovative AI and machine learning (ML) solutions are evolving to tackle security challenges and enhance ZTA tools. AI and ZTA will advance in integrating AI solutions into ZTA tools and designing AI-driven cyber defence tools. Composable security can adapt ZTA by generating action-and-observable maps of defences from its knowledge graph with minimal human input, improving responses to unknown attacks. This defence maintains resilience without strict guarantees, unlike current secure zero-knowledge proof methods. Additionally, cooperative multi-agent AIs pose a threat to zero-knowledge models using adversary agents, with advisory agents based on Monte Carlo Tree Search (MCTS) driving attack simulations and invoking defensive responses.

### 8.1. Advancing Technologies, Expected Progressions

Emerging technologies are reshaping data protection and governance. Organisations prioritise technologies with increased flexibility and lower costs. Zero trust architectures (ZTA) embody principles like "Never trust, always verify" and "Assume breach, "ensuring constant validation of actors, devices, and transactions. Acknowledging some breaches enables swift responses and policy updates. By adopting ZTAS, organisations enhance flexibility and cloud economics (Alevizos et al., 2021). Despite the rise of zero-trust services, ZTAs are still evolving, with various frameworks lacking consensus. Key implementation concerns involve securing endpoint devices, managing attributes and transactions, and response procedures. Legacy security tools fail to adapt to zero-trust principles, prompting a search for new endpoint security technologies (Röttinger & Wenning, 2024). Distributed Ledger Technologies (DLTs), such as blockchains, provide innovative security solutions. Although primarily for record-keeping, DLTs can enhance endpoint trust by verifying identity and policy attributes on-chain. However, governance overhead and competition may impede agility in many government institutions (Ghaffari et al., 2023). This section reviews predicted developments influenced by the COVID-19 pandemic, which presents breakthroughs and challenges for the future. The changing threat landscape and competitive dynamics necessitate reevaluating threat intelligence (SI-INT) projections. AI advancements introduce new norms and challenges, requiring timely threat intelligence. The focus should be on how AI-enhanced SI-CA tackles these challenges and the measures needed to mitigate potential threats (Alevizos et al., 2021). Predictions foresee diverse developments for 2023–2026. Aggressively, AI may empower existing exploit actors to explore challenging dissemination avenues beyond the English-speaking ecosystem. Nation-state actors are likely to refine attribution with AI, while counteractions will adapt to changing strategies (Shmueli & Ray, 2024).

## IX. Conclusion

As cybersecurity threats and challenges become increasingly intricate, the strategies to confront those dangers must also evolve. AI-augmented threat intelligence enables organisations to define intelligence-driven ZTA implementation and design against enterprise and compute assets, as well as potential threat agents, more effectively. Meanwhile, potential ZTA defence measures can be considered and employed against candidate threat actors. The ZTA framework is data-driven, emphasising trustless access on a per-request, session, and interaction basis between users, identities, workloads, devices, networks, data stores, and applications to resources. Further, the research needs to be refined, applied, and ported as a threat intelligence building block for existing and emerging research frameworks. Some subproblems, including zero-trust general architectures, ZTA for specific X paradigms, trustworthy AI-augmented threat intelligence, and ZTA-centric risk assessment, require attention. ZTA-centric threat intelligence management and construction, social engineering for ZTA breaches, penetration testing, assessments, and operations of ZTA, ZTA fusion research, and verification of ZTI achievements require further attention for systematisation. Assessing ZTA implementation and effectiveness is crucial but also challenging; exploring more criteria, innovative methods, and techniques, with an emphasis on zero trust, is

necessary to uncover the impacts of ZTA components on ZTI components. In the same vein, a benchmark is required to assess and understand the accuracy and completeness of existing ZTI for enhancement and evaluation. Finally, although zero-trust data management layer methods can enable federated learning without sharing raw data and thereby preserving data privacy and integrity, there are still some ZTI-centric research directions that remain to be explored.

## References:


[1]. Malhotra, P., Singh, Y., Anand, P., Bangotra, D. K., Singh, P. K., & Hong, W. C. (2021). Internet of Things: Evolution, Concerns, and Security Challenges. Sensors, 21(5), 1809. mdpi.com

[2]. Khan, M. J. (2023). Zero trust architecture: Redefining network security paradigms in the digital age. World Journal of Advanced Research and Reviews. wjarr.co.in

[3]. Kaur, H., SL, D. S., Paul, T., Thakur, R. K., Reddy, K. V. K., Mahato, J., & Naveen, K. (2024). Evolution of endpoint detection and response (EDR) in cybersecurity: A comprehensive review. In E3S Web of Conferences (Vol. 556, p. 01006). EDP Sciences. e3s-conferences.org

[4]. Alevizos, L., Thong Ta, V., & Hashem Eiza, M. (2021). Augmenting Zero Trust Architecture to Endpoints Using Blockchain: A State-of-the-Art Review. [PDF]

[5]. Sindhu, V. & Vinay, M. (2025). Possess Thorough Knowledge of Zero-Trust Principles. Zero-Trust Learning. [HTML]

[6]. Kang, H., Liu, G., Wang, Q., Meng, L., & Liu, J. (2023). Theory and Application of Zero Trust Security: A Brief Survey. ncbi.nlm.nih.gov

[7]. Kolade, T. M., Obioha Val, O., Balogun, A. Y., Gbadebo, M. O., & Olaniyi, O. O. (2025). AI-driven open source intelligence in cyber defence: A double-edged sword for national security. Adebayo Yusuf and Gbadebo, Michael Olayinka and Olaniyi, Oluwaseun Oladeji, AI-Driven Open Source Intelligence in Cyber Defence: A Double-edged Sword for National Security(January 18, 2025). ssrn.com

[8]. Alevizos, L., Ta, V. T., & Hashem Eiza, M. (2022). Augmenting zero trust architecture to endpoints using blockchain: A state-of-the-art review. Security and privacy. wiley.com

[9]. Kumar, N., S. Kasbekar, G., & Manjunath, D. (2022). Application of Data Collected by Endpoint Detection and Response Systems for Implementation of a Network Security System based on Zero Trust Principles and the EigenTrust Algorithm. [PDF]

[10]. Adamson, K. M. & Qureshi, A. (2025). Zero Trust 2.0: Advances, Challenges, and Future Directions in ZTA. [HTML]

[11]. Ahuja, A. (2024). A Detailed Study on Security and Compliance in Enterprise Architecture. ssrn.com

[12]. Tatam, M., Shanmugam, B., Azam, S., & Kannoorpatti, K. (2021). A review of threat modelling approaches for APT-style attacks. ncbi.nlm.nih.gov

[13]. Sun, N., Ding, M., Jiang, J., Xu, W., Mo, X., Tai, Y., & Zhang, J. (2023). Cyber threat intelligence mining for proactive cybersecurity defence: A survey and new perspectives. IEEE Communications Surveys & Tutorials, 25(3), 1748-1774. ieee.org

[14]. Mavroeidis, V. & Bromander, S. (2021). Cyber Threat Intelligence Model: An Evaluation of Taxonomies, Sharing Standards, and Ontologies within Cyber Threat Intelligence. [PDF]

[15]. Schmitt, M. (2023). Securing the Digital World: Protecting smart infrastructures and digital industries with Artificial Intelligence (AI)-enabled malware and intrusion detection. [PDF]

[16]. Bernardez Molina, S., Nespoli, P., & Gómez Mármol, F. (2023). Tackling Cyberattacks through AI-based Reactive Systems: A Holistic Review and Future Vision. [PDF]

[17]. Syed, N. F., Shah, S. W., Shaghaghi, A., Anwar, A., Baig, Z., & Doss, R. (2022). Zero trust architecture (ZTA): A comprehensive survey. IEEE Access, 10, 57143-57179. ieee.org

[18]. Malele, V. & E Mathonsi, T. (2023). Testing the performance of the Multi-class IDS public dataset using Supervised Machine Learning Algorithms. [PDF]

[19]. Pauling, C., Gimson, M., Qaid, M., Kida, A., & Halak, B. (2022). A Tutorial on Adversarial Learning Attacks and Countermeasures. [PDF]

[20]. Maseer, Z. K., Yusof, R., Bahaman, N., Mostafa, S. A., & Foozy, C. F. M. (2021). Benchmarking of machine learning for anomaly-based intrusion detection systems in the CICIDS2017 dataset. IEEE Access, 9, 22351-22370. ieee.org

[21]. Azab, A., Khasawneh, M., Alrabaee, S., Choo, K. K. R., & Sarsour, M. (2024). Network traffic classification: Techniques, datasets, and challenges. Digital Communications and Networks, 10(3), 676-692. sciencedirect.com

[22]. Dhir, N., Hoeltgebaum, H., Adams, N., Briers, M., Burke, A., & Jones, P. (2021). Prospective Artificial Intelligence Approaches for Active Cyber Defence. [PDF]

[23]. M. Soliman, H., Salmon, G., Sovilj, D., & Rao, M. (2021). RANK: AI-assisted End-to-End Architecture for Detecting Persistent Attacks in Enterprise Networks. [PDF]

[24]. Polemi, N., Praça, I., Kioskli, K., & Bécue, A. (2024). Challenges and Efforts in Managing AI Trustworthiness Risks: A State of Knowledge. ncbi.nlm.nih.gov

[25]. Radanliev, P., Santos, O., Brandon-Jones, A., & Joinson, A. (2024). Ethics and responsible AI deployment. ncbi.nlm.nih.gov

[26]. Astaburuaga, I. (2022). Privacy Preserving Cyber Threat Intelligence Sharing Framework for Encrypted Analytics. [PDF]

[27]. Englert, R. & Muschiol, J. (2020). Syntactic and Semantic Bias Detection and Countermeasures. ncbi.nlm.nih.gov

[28]. Bohdal, O., Hospedales, T., H. S. Torr, P., & Barez, F. (2023). Fairness in AI and Its Long-Term Implications on Society. [PDF]

[29]. Strandell, K. & Mittal, S. (2022). Risks to Zero Trust in a Federated Mission Partner Environment. [PDF]

[30]. Röttinger, R., & Wenning, S. (2024). ZERO TRUST ARCHITECTURES IN THE ENERGY SECTOR: APPLICATIONS AND BENEFITS. European Journal of Engineering and Technology, 12(1). idpublications.org

[31]. Ghaffari, F., Bertin, E., Crespi, N., & Hatin, J. (2023). Distributed ledger technologies for authentication and access control in networking applications: A comprehensive survey. Computer Science Review. hal. science

[32]. Shmueli, G., & Ray, S. (2024). Reimagining the Journal Editorial Process: An AI-Augmented Versus an AI-Driven Future. Journal of the Association for Information Systems, 25(1), 10. core.ac.uk