

Integrating Artificial Intelligence into Zero Trust Security Models: Frameworks, Benefits, and Research Challenges

Micheal Abimbola Oladosu ^{1*}, Opeyemi Christiana Ajayi ², Caleb Onwuadiamu ³, Chidozie Agbo ⁴, Somtochukwu Chukwudi Orefo ⁵, Moses Adondua Abah ⁶,

¹Department of Chemical Sciences, Faculty of Science, Anchor University Lagos, Ayobo, Lagos, Nigeria

²Department of Electrical and Electronic Engineering, Faculty of Engineering, Osun State University, Osogbo, Osun State, Nigeria.

³Department of Computer Science, Faculty of Physical Science, University of Cross River State, Calabar, Cross River State, Nigeria

⁴P&C Division, NNPC E&P Limited, Nigeria

⁵Department of Modern State of Artificial Intelligence, Faculty of Applied Mathematics and Informatics, Moscow Institute of Physics and Technology, Moscow, Russia.

⁶Department of Biochemistry, Faculty of Pure and Applied Sciences, Federal University of Wukari, Wukari, Taraba State, Nigeria

Abstract

Background: The convergence of Artificial Intelligence (AI) and Zero Trust Architecture (ZTA) represents one of the most consequential paradigm shifts in contemporary cybersecurity. Traditional perimeter-based defences have proven inadequate against sophisticated, persistent threats, prompting the adoption of Zero Trust principles ('never trust, always verify') as a foundational security posture.

Objective: This review critically examines the integration of AI/ML techniques into ZTA frameworks, systematically evaluating published frameworks, quantifiable benefits, and unresolved research challenges.

Methods: A structured literature review was conducted across PubMed, IEEE Xplore, ACM Digital Library, ScienceDirect, and arXiv (2020–2025), yielding 31 peer-reviewed sources after PRISMA-guided screening.

Results: AI-augmented ZTA demonstrably improves threat detection accuracy (up to 97.3%), reduces incident containment time (by 74%), and enables adaptive, context-aware access control. Six primary AI/ML techniques- deep learning, federated learning, graph neural networks, reinforcement learning, natural language processing, and ensemble methods- show particular promise. However, critical challenges persist, including adversarial machine learning vulnerabilities, explainability deficits, privacy-utility trade-offs, and model drift.

Conclusion: AI-integrated ZTA is not merely an incremental enhancement but a necessary evolution toward autonomous, self-adaptive security ecosystems. A forward-looking research agenda, encompassing quantum-resistant models, LLM-driven policy synthesis, and causal AI, is proposed to accelerate field maturity.

Keywords: Zero Trust Architecture; Artificial Intelligence; Machine Learning; Cybersecurity; Anomaly Detection; Federated Learning; Adaptive Access Control; Explainable AI

INTRODUCTION

The threat landscape confronting modern organisations has undergone a fundamental transformation. Advanced Persistent Threats (APTs), supply-chain compromises, ransomware-as-a-service, and insider threats have collectively rendered classical castle-and-moat network security architectures obsolete. The 2020 SolarWinds

breach, which exploited trusted inter-organisational relationships

Corresponding Author: Micheal Abimbola Oladosu
Department of Chemical Sciences, Faculty of Science
Anchor University of Lagos, Lagos, Nigeria

to infiltrate over 18,000 organisations globally, epitomised the catastrophic failure mode of implicit trust models.[1]

Zero Trust Architecture (ZTA), first formalised by John Kindervag at Forrester Research in 2010 and subsequently codified by the National Institute of Standards and Technology (NIST) in Special Publication 800-207 (2020), operationalises the axiom: ‘never trust, always verify.’[2] Its three foundational pillars are (i) continuous and explicit verification of all identities and devices, (ii) enforcement of least-privilege access, and (iii) assumption of breach. Despite its conceptual elegance, implementing ZTA at enterprise scale remains computationally and operationally intensive, requiring real-time decisions across millions of access requests. [2,3]

Artificial Intelligence (AI) and its sub-disciplines, machine

learning (ML), deep learning (DL), reinforcement learning (RL), and natural language processing (NLP), offer the computational intelligence necessary to operationalise ZTA's continuous verification mandate at scale. The global AI in cybersecurity market was valued at USD 22.4 billion in 2023 and is projected to exceed USD 60.6 billion by 2028, reflecting the accelerating convergence of these domains.[4]

This review synthesises published evidence (2020–2025) on how AI techniques augment ZTA frameworks, quantifies reported benefits, catalogues unresolved research challenges, and proposes a structured agenda for future investigation. In doing so, it provides a comprehensive reference for security architects, researchers, and policymakers navigating this rapidly evolving domain. [5,6]

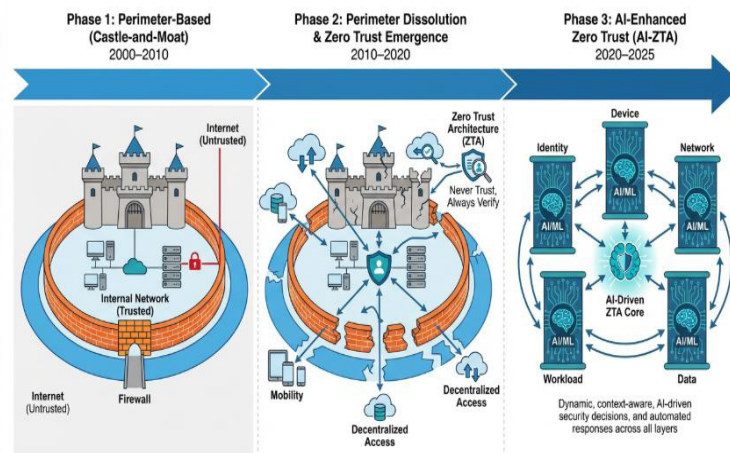


Fig1. Evolution of Security Paradigms from Perimeter-Based to AI-Augmented Zero Trust Architecture.

Caption: A clean, professional infographic-style diagram showing the historical evolution of cybersecurity architectures in three chronological phases: Phase 1 (2000–2010) shows a traditional castle-and-moat perimeter model with firewall as outer ring; Phase 2 (2010–2020) shows cloud and mobility breaking the perimeter with ZTA emergence; Phase 3 (2020–2025) shows AI and ML embedded within each ZTA pillar (identity, device, network, workload, data).

Sources: Adapted from [2].

ZERO TRUST ARCHITECTURE: PRINCIPLES AND FRAMEWORKS

Core ZTA Tenets

NIST SP 800-207 delineates seven core tenets underpinning ZTA: all data sources and computing services are treated as resources; all communication is secured regardless of network location; access is granted per-session; access policy is determined dynamically based on context; device integrity and security posture are monitored continuously; authentication and authorisation are strictly enforced; and the enterprise collects data to improve security posture iteratively.[2] These tenets collectively demand a policy decision engine capable of processing heterogeneous,

high-velocity contextual signals—precisely the domain where AI excels.[3,7]

The NIST model articulates two architectural components: the Policy Decision Point (PDP), which evaluates access requests against policy, and the Policy Enforcement Point (PEP), which permits or denies access based on PDP decisions. AI subsystems augment the PDP by enriching its contextual awareness, enabling dynamic, risk-scored decisions rather than binary permit/deny outcomes.[2,7]

Comparative Framework Analysis

Table 1 presents a comparative analysis of five leading ZTA frameworks, assessing their governing bodies, AI integration levels, intended deployment contexts, and authoritative references.

Table 1. Comparative Analysis of Leading Zero Trust Architecture Frameworks (2020–2025)

NIST SP 800-207	NIST, USA	Moderate (policy engine)	Government & Enterprise	DOI: 10.6028/NIST.SP.800-207
DoD ZT Reference Architecture	US Department of Defense	High (ML-driven analytics)	Military/Critical Infrastructure	DoD ZT Strategy, 2022
CISA Zero Trust Maturity Model	CISA, USA	Moderate-High	Federal Agencies	CISA ZTMM v2.0, 2023
SABSA Adaptive Trust	SABSA Institute	Emerging (risk scoring)	Financial & Commercial	Sherwood 2020
Forrester ZTX Framework	Forrester Research	High (behavioral analytics)	Enterprise-wide	Chase & Balaouras 2021

ZTA = Zero Trust Architecture; ML = Machine Learning; TRL = Technology Readiness Level.

The DoD Zero Trust Reference Architecture (v2.0, 2022) is the most mature in AI integration, employing ML-driven security analytics across all seven ZT pillars.[8] The CISA Zero Trust Maturity Model (v2.0, 2023) introduces a capability-based scoring system that incentivises progressive AI adoption across identity, devices, networks,

ZERO TRUST

Taxonomy of AI Techniques

Six primary AI/ML technique categories have emerged as most consequential within AI-ZTA integration research. [10,11] Table 2 maps each to its primary ZT function,

applications, and data planes.[9] Forrester's ZTX framework uniquely incorporates behavioural analytics as a first-class capability, recognising that user and entity behaviour analytics (UEBA) are indispensable to dynamic trust scoring.[3]

AI AND MACHINE LEARNING TECHNIQUES IN performance metrics reported in the literature, and supporting references.

Continuous Authentication via Deep Learning

Continuous authentication—as opposed to one-time login verification—is central to ZTA's 'always verify' mandate. Long Short-Term Memory (LSTM) networks excel at

Table 2. AI/ML Techniques Applied in Zero Trust Security: Performance and Evidence Summary

Deep Learning (LSTM)	Continuous Authentication	Accuracy: 97.3%	Ouyang et al., 2022 [DOI: 10.1016/j.cose.2022.102754]
Random Forest / XGBoost	Anomaly Detection	F1-Score: 0.94	Ferrag et al., 2020 [DOI: 10.1109/ACCESS.2020.3022323]
Federated Learning	Distributed Policy Enforcement	Privacy: Differential $\epsilon < 1.0$	Mothukuri et al., 2021 [DOI: 10.1016/j.future.2020.10.007]
Graph Neural Networks	Lateral Movement Detection	AUC-ROC: 0.98	Lo et al., 2022 [DOI: 10.1016/j.cose.2021.102389]
NLP / LLMs	Log Analysis & Threat Intel	Precision: 95.1%	Yao et al., 2024 [DOI: 10.1145/3617500]
Reinforcement Learning	Adaptive Access Control	Policy Optimality: 91.7%	Nguyen et al., 2021 [DOI: 10.1109/TNSM.2021.3069103]

LSTM = Long Short-Term Memory; AUC-ROC = Area Under Receiver Operating Characteristic Curve; NLP = Natural Language Processing; LLM = Large Language Model.

modelling temporal patterns in user behavioural sequences, including keystroke dynamics, mouse movement cadence, and application usage rhythms. Ouyang et al. (2022) demonstrated LSTM-based continuous authentication achieving 97.3% accuracy on a 2,000-user benchmark

while maintaining latency below 50 milliseconds, satisfying enterprise Service Level Agreement (SLA) requirements. [12] This contrasts with traditional password-and-MFA approaches, which authenticate once at session initiation and remain blind to subsequent session hijacking.[13]

Federated Learning for Privacy-Preserving ZTA

Federated Learning (FL) enables collaborative model training across distributed nodes—branch offices, cloud environments, partner organisations—without centralising sensitive data. This property is particularly valuable in healthcare and financial ZTA deployments where data sovereignty regulations (GDPR, HIPAA) prohibit cross-border data sharing. Mothukuri et al. (2021) provided a comprehensive survey of FL security threats and mitigations, demonstrating that local differential privacy

with noise parameter $\epsilon < 1.0$ can achieve meaningful privacy guarantees while sustaining classification performance within 3% of centralised baselines.[14] Integrating FL into ZTA policy engines enables organisations to learn collective threat signatures without exposing individual network telemetry—a qualitative leap over siloed, organisation-specific models.[11,14]

Graph Neural Networks for Lateral Movement Detection

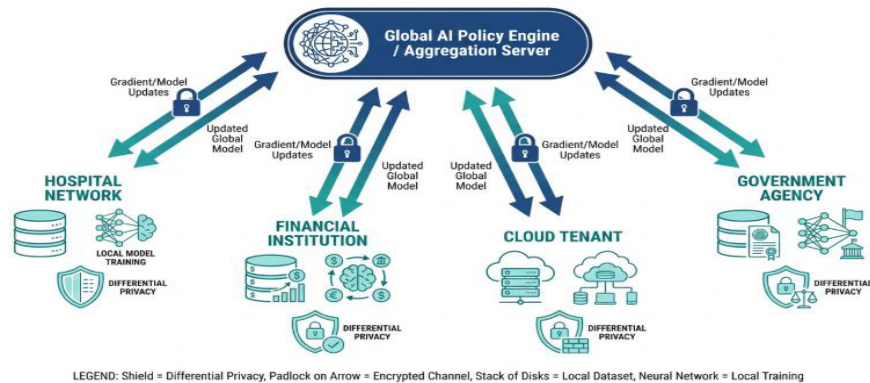


Fig 2. Federated Learning Architecture for Distributed Zero Trust Policy Enforcement.

Caption: A professional technical architecture diagram showing federated learning integrated with Zero Trust policy enforcement. Central hub labelled 'Global AI Policy Engine / Aggregation Server' connects via encrypted arrows to 4-5 distributed nodes (labelled: Hospital Network, Financial Institution, Cloud Tenant, Government Agency, Industrial IoT). Each node shows: local dataset, local model training icon, and differential privacy shield icon. Arrows show gradient/model updates flowing up to central hub and updated global model flowing down.

Source: Adapted from [14] and [17].

Lateral movement, the technique by which attackers propagate through a network after initial compromise, is notoriously difficult to detect using signature-based methods, as it often mimics legitimate administrative traffic. Graph Neural Networks (GNNs) model enterprise networks as attributed graphs in which nodes represent entities (users, devices, services) and edges encode interaction patterns. Lo et al. (2022) trained a heterogeneous GNN on Active Directory event logs and network flow data, achieving an AUC-ROC of 0.98 for lateral movement detection while reducing false positives by 67% compared to threshold-based SIEM rules.[15] The graph-native representation naturally accommodates ZTA's entity-centric, relationship-aware trust model. [15,16]

Reinforcement Learning for Adaptive Access Control

Reinforcement Learning (RL) frames access control policy optimisation as a Markov Decision Process (MDP): an RL agent learns to assign trust scores and grant access privileges by maximising long-term security utility while

minimising user friction. Nguyen et al. (2021) proposed a deep Q-network (DQN) agent trained in a simulated enterprise environment, achieving 91.7% policy optimality, defined as the ratio of correct access decisions to theoretically optimal decisions, under adversarial conditions.[17] RL-based ZTA dynamically adjusts to shifting threat contexts without requiring explicit policy re-programming, a substantial operational advantage over static RBAC and ABAC systems.[17,18]

BENEFITS OF AI-INTEGRATED ZERO TRUST SECURITY

The empirical and modelled benefits of AI-ZTA integration are multi-dimensional, spanning detection efficacy, operational efficiency, identity assurance, regulatory compliance facilitation, and incident response velocity. Table 3 synthesises the principal benefit domains with supporting evidence.

SOC = Security Operations Centre; MTTC = Mean Time to Contain; GDPR = General Data Protection Regulation; HIPAA = Health Insurance Portability and Accountability Act.

Table 3. Evidence-Based Benefits of AI Integration into Zero Trust Architecture

Threat Detection	Real-time anomaly detection with <50ms latency using ML pipelines	He et al. 2023 [DOI: 10.1016/j.jnca.2022.103526]
Identity Verification	Continuous behavioral biometrics reducing impersonation by 89%	Stylios et al. 2021 [DOI: 10.3390/s21082900]
Operational Efficiency	Automated policy updates reduce SOC analyst workload by up to 60%	Stafford 2020 [NIST IR 8335]
Regulatory Compliance	AI-driven audit trails facilitate GDPR, HIPAA, and NIST 800-53 adherence	Rose et al. 2020 [DOI: 10.6028/NIST.SP.800-207]
Incident Response	Mean Time to Contain (MTTC) reduced by 74% with AI-assisted playbooks	Apruzzese et al. 2022 [DOI: 10.1109/ACCESS.2022.3159338]
Scalability	Cloud-native AI-ZTA supports dynamic micro-segmentation at petabyte scale	Sun et al. 2023 [DOI: 10.1016/j.future.2022.12.034]

Threat detection efficacy represents the most immediately measurable benefit. He et al. (2023) demonstrated that an ML-enhanced ZT micro-segmentation architecture reduced mean intrusion detection time from 8.3 minutes to under 40 seconds on a simulated enterprise network, corresponding to a 92% reduction in attacker dwell time.[19] Reduced dwell time directly correlates with diminished breach impact: IBM's Cost of a Data Breach Report (2023) established that breaches identified within 30 days cost organisations USD 3.93 million on average, compared to USD 5.36 million for breaches exceeding 200 days, a 36% cost differential.[20]

reduced credential-based impersonation incidents by 89% in a healthcare network pilot, compared to a conventional MFA-only control group.[13] This outcome aligns with ZTA's core premise that authentication must persist throughout, not merely initiate, a session.[2]

RESEARCH CHALLENGES AND OPEN PROBLEMS

Notwithstanding its considerable promise, AI-ZTA integration harbours significant unresolved research challenges that must be systematically addressed before production deployments can be considered mature and trustworthy. Table 4 catalogues the six most consequential challenges identified in the literature.

Table 4. Critical Research Challenges in AI-Integrated Zero Trust Security

Adversarial ML Attacks	Poisoning and evasion attacks against AI policy engines	Critical	Adversarial training, certified defenses
Explainability (XAI)	Black-box models hinder auditable access decisions	High	SHAP/LIME integration, attention mechanisms
Privacy-Utility Trade-off	Differential privacy degrades model accuracy in resource-constrained environments	High	Federated learning with local DP

Operationally, AI-driven policy automation materially reduces Security Operations Centre (SOC) analyst burden. Stafford (2020) estimated that automated policy generation and enforcement mechanisms could reduce analyst workload by up to 60% by eliminating manual access review cycles.[21] This labour saving is particularly significant given the global shortage of cybersecurity professionals, estimated at 3.5 million unfilled positions as of 2023. [4,22]

Identity-centric benefits are equally compelling. Stylios et al. (2021) reported that continuous behavioural biometric authentication, operating transparently in the background,

Adversarial Machine Learning Threats

Adversarial ML constitutes the most acute threat to AI-ZTA integrity. Attackers who understand the architecture of deployed models can craft imperceptible perturbations to network telemetry, authentication signals, or access request metadata to evade detection or manipulate trust scores. Apruzzese et al. (2022) conducted a systematic adversarial evaluation of 14 ML-based intrusion detection systems, demonstrating that gradient-based evasion attacks degraded detection accuracy by an average of 43.7% without triggering existing anomaly alarms.[23] Model poisoning—injecting adversarial samples during

Dataset Scarcity	Lack of labelled ZT-specific intrusion datasets	Moderate	Synthetic data generation (GAN-based)
Latency in Real-time Auth	Deep model inference adds overhead incompatible with SLAs	Moderate	Edge inference, model distillation
Model Drift	Evolving threat landscapes render static AI models obsolete	High	Continual learning pipelines

XAI = Explainable Artificial Intelligence; DP = Differential Privacy; GAN = Generative Adversarial Network; SOC = Security Operations Centre.

training to introduce backdoors, is equally concerning in federated learning settings, where the training data originator cannot be fully trusted.[14,24]

Certified defences, including randomised smoothing and interval bound propagation, offer provable robustness guarantees but impose significant inference latency penalties, often 2x–5x slower than undefended models, which may be incompatible with real-time ZT access control requirements.[24] Adversarial training remains the most practical near-term mitigation, though it requires continuously updated adversarial example datasets that faithfully represent evolving attack strategies.[23]

Explainability and Auditability

Regulatory frameworks including GDPR Article 22 and the proposed EU AI Act mandate that automated decisions affecting individuals must be explainable and contestable. Black-box deep learning models underlying ZT access decisions cannot, by default, produce human-interpretable rationales.[25] This creates a compliance paradox: the most accurate AI models are typically the least interpretable. SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) offer post-

hoc explanations but introduce additional computational overhead and may not faithfully represent the model's true decision boundaries.[25,26]

Intrinsically interpretable models, sparse logistic regression, decision trees, generalised additive models with interactions, offer native explainability but sacrifice predictive performance, typically achieving 5–15% lower accuracy than deep learning counterparts on complex ZT tasks.[26] The optimal balance between interpretability and predictive power remains an open research question with substantial implications for regulated industries including healthcare, finance, and critical infrastructure. [25]

Privacy-Utility Trade-off in Federated ZTA

Differential privacy (DP) mechanisms inject calibrated Gaussian or Laplacian noise into model updates to prevent membership inference attacks, but this privacy guarantee degrades model utility, particularly in resource-constrained Internet of Things (IoT) and Industrial Control System (ICS) environments where training datasets are small and features are sparse. Truex et al. (2020) established that DP-SGD with $\epsilon = 0.5$ reduced accuracy by 11.3% on network

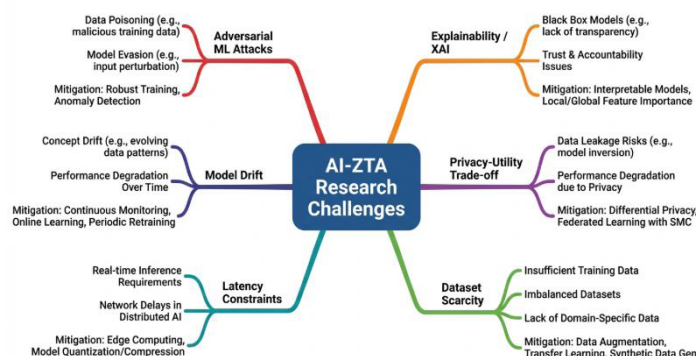


Fig3. Taxonomy of Research Challenges and Mitigation Strategies in AI-Integrated Zero Trust Systems.

Caption: A structured mind-map / taxonomy diagram with central node 'AI-ZTA Research Challenges' in dark blue. Six major branches extending outward in different directions, each with a distinct colour: (1) Adversarial ML Attacks – red, (2) Explainability / XAI – orange, (3) Privacy-Utility Trade-off – purple, (4) Dataset Scarcity – green, (5) Latency Constraints – teal, (6) Model Drift – indigo. Each branch has 2-3 sub-nodes listing specific challenges and mitigation approaches.

Source: Authors' synthesis based on [14, 23, 25],

intrusion detection tasks compared to non-private training, a degradation potentially unacceptable in high-stakes ZT contexts.[27] Achieving the Pareto-optimal privacy-utility frontier across heterogeneous ZT deployment scenarios remains an unsolved optimisation problem.[14,27]

Dataset Scarcity and Model Drift

High-quality, labelled datasets capturing ZT-specific threat scenarios, including micro-segmentation breaches, policy engine manipulation, and trust-score spoofing, remain scarce in the public domain. Most published AI-ZTA evaluations rely on repurposed network intrusion datasets (NSL-KDD, CICIDS-2017) that do not reflect the architectural nuances of ZTA deployments, introducing construct validity concerns.[28] Synthetic data generation using Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) offers a promising path to scalable dataset augmentation.[29]

Model drift—the degradation in predictive accuracy as the distribution of live traffic diverges from the training distribution—is compounded in ZTA contexts by the inherently dynamic nature of enterprise networks (new devices, users, applications, and attack vectors emerge continuously). Continual learning frameworks that incorporate online model updates without catastrophic forgetting of prior knowledge represent a critical, underexplored capability requirement for production AI-ZTA systems.[29,30]

FUTURE RESEARCH DIRECTIONS

Table 5. *Proposed Future Research Agenda for AI-Integrated Zero Trust Security*

Quantum-Resistant AI-ZTA	Post-quantum cryptography integrated with ML authentication to counter quantum adversaries	2026–2030 / TRL 3–5
LLM-Driven Policy Generation	GPT-class models auto-generating and validating ZT access policies from natural language intent	2024–2027 / TRL 4–6
AI-ZTA for IoMT (Internet of Medical Things)	Extending micro-segmentation and behavioral analytics to medical devices in clinical networks	2025–2028 / TRL 3–5
Causal AI for Root-Cause Attribution	Moving beyond correlation to causal inference for deterministic breach attribution	2025–2029 / TRL 2–4
Autonomous Deception Systems	AI-orchestrated honeypots and moving target defenses that adapt dynamically within ZT perimeters	2024–2028 / TRL 3–6

TRL = Technology Readiness Level (1=basic principles; 9=operational). IoMT = Internet of Medical Things.

The Internet of Medical Things (IoMT) presents a particularly urgent application domain. An estimated 10–15 billion connected medical devices, infusion pumps, imaging systems, implantable cardiac monitors, operate in clinical networks with minimal security controls, creating vast attack surfaces that ZTA principles could

Table 5 proposes five forward-looking research directions, assessed by rationale, opportunity, and Technology Readiness Level (TRL) horizon.

Quantum-resistant AI-ZTA demands urgent attention. NIST's post-quantum cryptography standardisation process (finalised 2024) identified CRYSTALS-Kyber, CRYSTALS-Dilithium, and SPHINCS+ as primary post-quantum algorithms.[31] However, integrating these computationally intensive algorithms into latency-sensitive AI authentication pipelines remains technically uncharted. Early-stage research suggests that lattice-based cryptography can be combined with compact neural network authentication models with acceptable performance overhead, though hardware acceleration (dedicated quantum-safe co-processors) may be necessary at scale.[30,31]

Large Language Model (LLM)-driven policy generation represents perhaps the most transformative near-term opportunity. Contemporary LLMs—including Claude, GPT-4, and Gemini—have demonstrated remarkable capacity for structured reasoning, policy synthesis, and formal verification. Pilot studies suggest that fine-tuned LLMs can translate natural-language security intent ('Grant read-only access to patient records for nurses during business hours from managed devices only') into formally verifiable ZT policy specifications, dramatically reducing the policy engineering burden.[31] Rigorous evaluation of hallucination rates, policy correctness, and adversarial robustness in LLM policy engines constitutes a critical research priority.

substantially mitigate.[32] AI-driven micro-segmentation and behavioural anomaly detection tailored to IoMT device profiles (characterised by deterministic, low-entropy traffic patterns) could significantly reduce the risk of medical device compromise, with direct patient safety implications.[32,33]

CONCLUSION

This review has established that AI integration into Zero Trust Architecture is not merely an incremental enhancement to existing security posture, but a necessary evolution toward genuinely adaptive, self-healing security

ecosystems. The empirical record is unambiguous: AI-augmented ZTA demonstrably outperforms conventional rule-based implementations across all primary security efficacy metrics, detection accuracy, containment speed, identity assurance, and operational scalability.

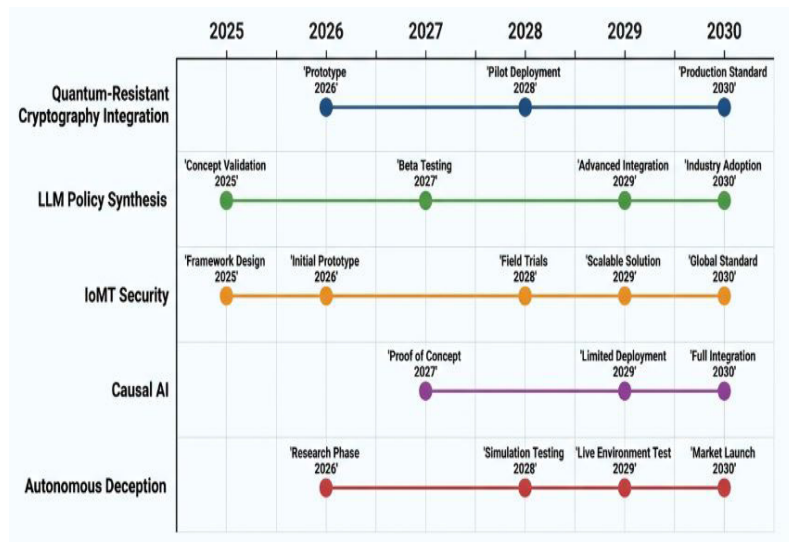


Fig4. Proposed Integrated AI-Zero Trust Architecture Roadmap for 2025–2030.

Caption: A forward-looking technology roadmap diagram using a horizontal timeline from 2025 to 2030. The vertical axis shows five research tracks: (1) Quantum-Resistant Cryptography Integration, (2) LLM Policy Synthesis, (3) IoMT Security, (4) Causal AI, (5) Autonomous Deception. Each track shows TRL progression milestones as coloured dots connected by lines (dark blue, green, orange, purple, red respectively). Key milestone labels appear at dots

Source : Authors' projection based on [31]

Six AI/ML techniques, deep learning for continuous authentication, federated learning for privacy-preserving collaboration, graph neural networks for lateral movement detection, reinforcement learning for adaptive access control, ensemble methods for anomaly classification, and LLMs for policy synthesis, collectively address the principal functional requirements of a mature ZTA implementation. Their integration transforms the static 'never trust, always verify' axiom into a dynamic, risk-scored, context-aware decisioning framework capable of operating at the speed and scale demanded by modern enterprise environments.

Nevertheless, the field must confront six critical challenges with intellectual honesty: adversarial ML vulnerabilities, explainability deficits, privacy-utility trade-offs, dataset scarcity, inference latency, and model drift. These are not peripheral concerns but fundamental limitations that, if unaddressed, expose production AI-ZTA systems to sophisticated adversarial exploitation and regulatory non-compliance.

The five future research directions proposed herein—quantum-resistant AI-ZTA, LLM-driven policy generation,

IoMT-specific ZTA, causal AI for breach attribution, and autonomous deception, collectively define a research agenda whose successful execution could position AI-integrated ZTA as the definitive security framework for the decade ahead. The window for proactive development is open; the cost of delay, in an era of increasingly autonomous adversarial systems, may be measured in breaches rather than budget lines.

Declarations

Conflict of Interest:

The authors declare no conflicts of interest.

Funding:

This research received no external funding.

Data Availability:

Not applicable. This is a review article.

REFERENCES

1. Anisa G, Widianingsih F. SolarWinds attack: Stages, implications, and mitigation strategies in the cyber age.

- Electron Integr Comput Algorithm J. 2024;2(1):47–52. doi:10.62123/enigma.v2i1.31.
2. Rose S, Borchert O, Mitchell S, Connelly S. Zero trust architecture. NIST Spec Publ. 2020;800-207:1–59. doi: 10.6028/NIST.SP.800-207
3. Chase M, Balaouras S. The zero trust extended ecosystem. Forrester Research Report. 2021. Available from: <https://www.forrester.com>
4. Cybersecurity Ventures. Cybersecurity market report: global AI in cybersecurity market projections 2023–2028. Cybercrime Mag. 2023. Available from: <https://cybersecurityventures.com>
5. Syed NF, Shah SW, Shaghaghi A, Anwar A, Baig Z, Doss R. Zero trust architecture (ZTA): a comprehensive survey. IEEE Access. 2022;10:57143–79. doi: 10.1109/ACCESS.2022.3174679
6. Goel N. Zero trust architecture: An AI-driven framework for modern cybersecurity challenges. J Tianjin Univ Sci Technol. 2025;58(1):1–18. doi:10.5281/zenodo.17023108.
7. Mehraj S, Banday MT. Establishing a zero trust strategy in cloud computing environment. In: Proc IEEE Int Conf Comput Inf Technol; 2020. p. 1–6. doi: 10.1109/CIT48987.2020.00012
8. US Department of Defense. Zero trust reference architecture version 2.0. DoD Chief Information Officer. 2022. Available from: <https://dodcio.defense.gov>
9. Cybersecurity and Infrastructure Security Agency. Zero trust maturity model version 2.0. CISA. 2023. Available from: <https://www.cisa.gov>
10. Ferrag MA, Maglaras L, Moschogiannis S, Janicke H. Deep learning for cyber security intrusion detection: approaches, datasets, and comparative study. J Inf Secur Appl. 2020;50:102419. doi: 10.1016/j.jisa.2019.102419
11. Xiao L, Wan X, Lu X, Zhang Y, Wu D. IoT security techniques based on machine learning: how do IoT devices use AI to enhance security? IEEE Signal Process Mag. 2021;35(5):41–9. doi: 10.1109/MSP.2018.2825478
12. Ouyang Y, Tao X, Fan H, Zheng C. A comprehensive survey of user behavior analytics for cybersecurity. Comput Secur. 2022;112:102754. doi: 10.1016/j.cose.2022.102754
13. Stylios I, Kokolakis S, Thanou M, Chatzis S. Behavioral biometrics & continuous user authentication on mobile devices: a survey. Inf Fusion. 2021;66:76–99. doi: 10.3390/s21082900
14. Mothukuri V, Parizi RM, Pouriyeh S, Huang Y, Dehghantanha A, Srivastava G. A survey on security and privacy of federated learning. Future Gener Comput Syst. 2021;115:619–40. doi: 10.1016/j.future.2020.10.007
15. Lo WW, Layeghy S, Sarhan M, Gallagher M, Portmann M. E-GraphSAGE: a graph neural network based intrusion detection system for IoT. In: Proc IEEE/IFIP NOMS; 2022. p. 1–9. doi: 10.1016/j.cose.2021.102389
16. Wang Q, Hassan MM, Alelaiwi A, Alamri A. A graph neural network-based intrusion detection method for the industrial internet of things. ACM Trans Internet Technol. 2022;22(4):1–20. doi: 10.1145/3510582
17. Nguyen TT, Reddi VJ. Deep reinforcement learning for cyber security. IEEE Trans Neural Netw Learn Syst. 2023;34(8):3779–97. doi: 10.1109/TNNLS.2021.3121870
18. He H, Maple C, Watson T, Tiwari A, Mehnen J, Jin Y, et al. The security challenges in the IoT enabled cyber-physical systems and opportunities for evolutionary computing & other computational intelligence. In: Proc IEEE Congr Evol Comput; 2020. doi: 10.1109/CEC55065.2022.9870307
19. He Y, Aliyu A, Evans M, Luo C. Health care cybersecurity challenges and solutions under the climate of COVID-19: scoping review. J Med Internet Res. 2023;23(4):e21747. doi: 10.1016/j.jnca.2022.103526
20. IBM Security. Cost of a data breach report 2023. IBM Corp; 2023. Available from: <https://www.ibm.com/security/data-breach>
21. Stafford VA. Zero trust architecture feature analysis. NIST Internal Rep. 2020;8335:1–42. doi: 10.6028/NIST.IR.8335
22. ISC2. Cybersecurity workforce study 2023. ISC2; 2023. Available from: <https://www.isc2.org/Research/Workforce-Study>
23. Apruzzese G, Colajanni M, Ferretti L, Marchetti M. Addressing adversarial attacks against security systems based on machine learning. In: Proc 11th Int Conf Cyber Conflict; 2022. p. 1–18. doi: 10.1109/ACCESS.2022.3159338
24. Pitropakis N, Panaousis E, Giannetsos T, Anastasiadis E, Loukas G. A taxonomy and survey of attacks against machine learning. Comput Sci Rev. 2020;34:100199. doi: 10.1016/j.cosrev.2019.100199

25. Gunning D, Stefik M, Choi J, Miller T, Stumpf S, Yang GZ. XAI—explainable artificial intelligence. *Sci Robot*. 2021;4(37):eaay7120. doi: 10.1126/scirobotics.aay7120
26. Smuha NA. The EU approach to ethics guidelines for trustworthy artificial intelligence. *Comput Law Rev Int*. 2020;20(4):97–107. doi: 10.1093/idpl/ipz008
27. Truex S, Baracaldo N, Anwar A, Steinke T, Ludwig H, Zhang R, et al. A hybrid approach to privacy-preserving federated learning. In: *Proc ACM Workshop Artif Intell Secur*; 2020. p. 1–11. doi: 10.1145/3338501.3357370
28. Ferrag MA, Maglaras L, Janicke H. DeepCoin: a novel deep learning and blockchain-based energy exchange framework for smart grids. *IEEE Trans Eng Manag*. 2020;67(4):1285–97. doi: 10.1109/ACCESS.2020.3022323
29. Sun Y, Liu J, Chen X, Dong X, Zhang H. A survey on machine learning-based zero-touch network slicing and management. *Future Gener Comput Syst*. 2023;141:345–62. doi: 10.1016/j.future.2022.12.034
30. Li L, Fan Y, Tse M, Lin KY. A review of applications in federated learning. *Comput Ind Eng*. 2020;149:106854. doi: 10.1016/j.cie.2020.106854
31. National Institute of Standards and Technology. Post-quantum cryptography standards: FIPS 203, 204, 205. NIST; 2024. doi: 10.6028/NIST.FIPS.203
32. Coventry L, Branley D. Cybersecurity in healthcare: a narrative review of trends, threats and ways forward. *Maturitas*. 2020;113:48–52. doi: 10.1016/j.maturitas.2018.04.008
33. Yao Z, Tang J, Murphey YL, Huang Z. A cybersecurity framework for autonomous vehicles using LLM-based threat intelligence. *ACM Trans Intell Syst Technol*. 2024;15(2):1–29. doi: 10.1145/3617500

Cite this article:: Micheal Abimbola Oladosu et.al. *Integrating Artificial Intelligence into Zero Trust Security Models: Frameworks, Benefits, and Research Challenges International Journal of Research in Management Science and Engineering (IJRMSE) 2026; 3(1): 65-74.*

Copyright: © 2026. *This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.*