Computer Fraud and Security
ISSN (online): 1873-7056

Agent-Aware Zero Trust: A Framework for Securing Agentic AI in SASE and Cloud Architectures

Bhavinkumar Jayswal

Independent Researcher, USA

Abstract

Enterprise networking is undergoing a fundamental transition as Secure Access Service Edge (SASE), cloud-native architectures, and software-defined control planes converge with advances in artificial intelligence. A new class of systems, referred to as *Agentic AI*, is emerging within these environments. Unlike traditional automation, agentic systems exhibit goal-directed behavior, adapt to environmental feedback, and execute actions with limited or no human intervention. While such autonomy promises significant gains in efficiency and resilience, it also destabilizes the deterministic assumptions underlying conventional Zero Trust and SASE security models.

This paper introduces Agent-Aware Zero Trust, a security framework designed to govern autonomous, probabilistic agents operating within enterprise SASE and cloud environments. The framework treats autonomous agents as first-class identities subject to continuous behavioral verification, policy-bounded autonomy, and probabilistic trust enforcement. A threat taxonomy specific to agentic systems is presented, including objective drift, delegated privilege escalation, control-plane lateral movement, emergent multi-agent behavior, and decision opacity. To mitigate these risks, the paper proposes architectural mechanisms including cryptographic agent identity, hierarchical policy envelopes, dynamic trust decay models, telemetry-driven supervision, and deterministic kill-switches.

This work presents a conceptual and architectural security framework, grounded in enterprise-scale SASE and cloud operations, rather than a controlled experimental or simulation-based evaluation. The objective is to establish a defensible security model for enterprises seeking to deploy autonomous networking systems while maintaining governance, compliance, and human oversight.

Keywords: Agentic AI, Zero Trust Architecture, SASE, Autonomous Network Security, Trust Decay, Policy-Bounded Autonomy, Enterprise Cloud Security.

1. Introduction

The evolution of enterprise networking over the past decade has been defined by abstraction and decoupling. Software-Defined Networking (SDN) separated control planes from data planes. Network virtualization abstracted logical connectivity from physical infrastructure. Secure Access Service Edge (SASE) unified wide-area networking and security enforcement into cloud-delivered platforms. In parallel, Zero Trust Architecture (ZTA) displaced perimeter-based security models with continuous verification and least-privilege access.

These architectural shifts enabled scalable, cloud-first connectivity across distributed enterprises. They also created environments rich in telemetry, APIs, and programmable control surfaces. Artificial intelligence initially entered this domain as an analytical aid, providing anomaly detection, root cause analysis, and optimization recommendations. Increasingly, however, AI systems are moving beyond recommendation and automation toward autonomy.

Agentic AI systems differ fundamentally from traditional automated controls. They observe system state, infer intermediate goals, adapt strategies over time, and execute actions with minimal human oversight. In enterprise networking, such agents may dynamically reroute traffic, adjust security policies, allocate resources, or respond to perceived threats in real time. These capabilities challenge the core assumptions of existing security architectures, which presume deterministic behavior and human-governed decision logic.

The literature has extensively explored machine learning for networking and security, yet it lacks a unified security framework for governing autonomous, goal-directed agents operating at the control plane of enterprise infrastructure. This paper addresses that gap by proposing Agent-Aware Zero Trust, an architectural extension of Zero Trust principles explicitly designed for agentic systems.

The contributions of this paper are threefold:

1. A precise definition of Agentic AI in the context of enterprise networking and security.
2. A threat taxonomy capturing risks unique to autonomous agents in SASE and cloud environments.
3. A novel security framework that governs agent autonomy through identity, policy constraints, probabilistic trust, and deterministic override mechanisms.

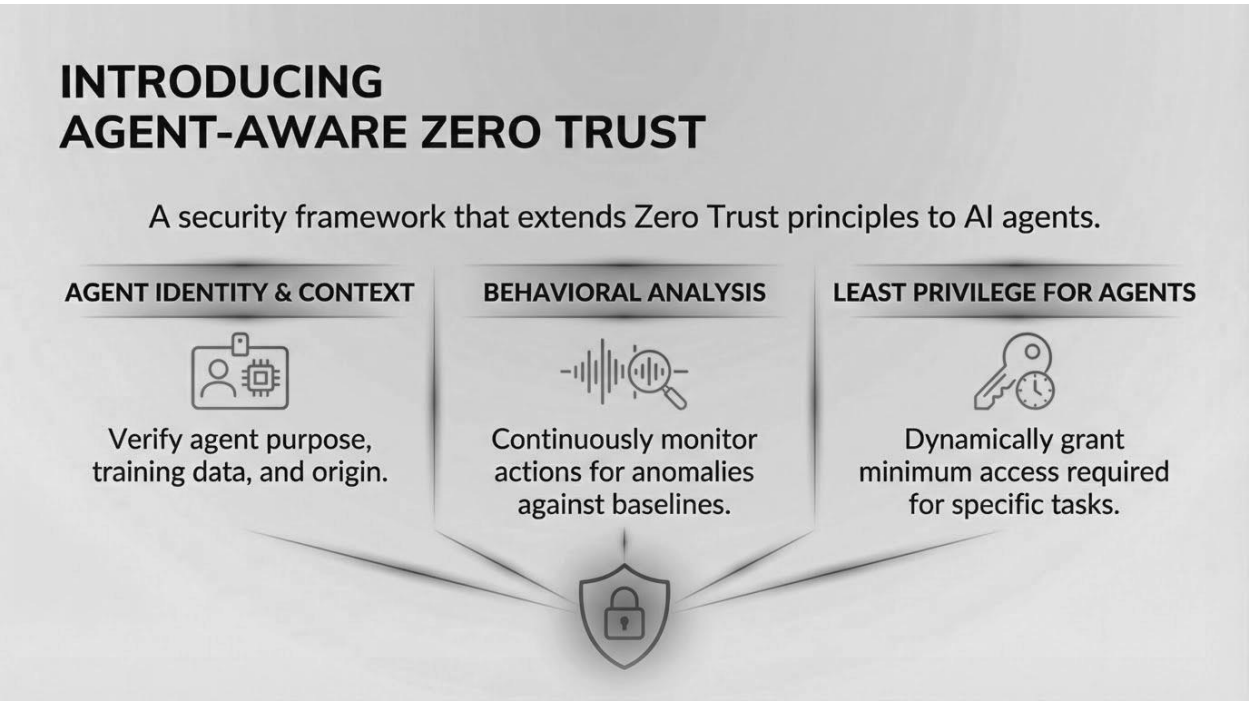

Figure 1: Agent-Aware Zero Trust Architecture Framework

A layered architectural diagram illustrating the five core components of Agent-Aware Zero Trust: Agent Identity & Continuous Verification (bottom layer), Policy-Bounded Autonomy (constraint envelope layer), Trust Decay & Behavioral Enforcement (middle monitoring layer), Telemetry-Driven Supervision (observation layer), and Deterministic Kill-Switches (top override layer). Arrows indicate continuous verification flows and feedback loops between components.

1.1 Scope and Methodology

This work is conceptual and architectural in nature. The framework presented is derived from design synthesis and operational experience in large-scale enterprise SASE, SD-WAN, and cloud networking environments. It does not represent a controlled laboratory experiment or statistically rigorous simulation study.

Threat models and architectural controls are abstracted from observed patterns in enterprise operations, cloud-managed infrastructure, and AI-enabled automation systems. Quantitative metrics referenced are illustrative or operationally observed rather than experimentally validated. The objective is to establish a defensible security architecture that can guide future implementations, standardization efforts, and empirical research.

2. Architectural Foundations and Limitations

2.1 Deterministic Assumptions in Zero Trust and SASE

Zero Trust Architecture is built on continuous verification, least privilege, and explicit trust boundaries. SASE operationalizes these principles by integrating networking and security functions into centralized, cloud-managed platforms. Both models implicitly assume that the entities being governed. Users, devices, workloads. Behave in largely predictable ways.

Policy evaluation within these systems is deterministic. Given a defined identity, posture, and context, enforcement outcomes are expected to be stable and explainable. Decision logic is typically human-authored, centrally governed, and auditable.

Autonomous agents violate these assumptions. Their behavior evolves over time as internal models update. Identical environmental inputs may yield different actions depending on learning state, reward structures, or prior outcomes. This probabilistic behavior introduces uncertainty into security enforcement that existing architectures are not designed to manage.

2.2 Network Virtualization and Control-Plane Risk

Network virtualization and cloud orchestration allow rapid reconfiguration of topology, routing, and security boundaries. While this flexibility enables scalability, it also concentrates risk in programmable control planes. Autonomous agents operating at this layer can modify network state at machine speed.

In multi-cloud and SASE environments, an agent with control-plane access may inadvertently or maliciously collapse isolation boundaries, bypass inspection paths, or propagate misconfigurations globally. Traditional controls assume that such changes are human-initiated and reviewed. Autonomous agents remove that assumption.

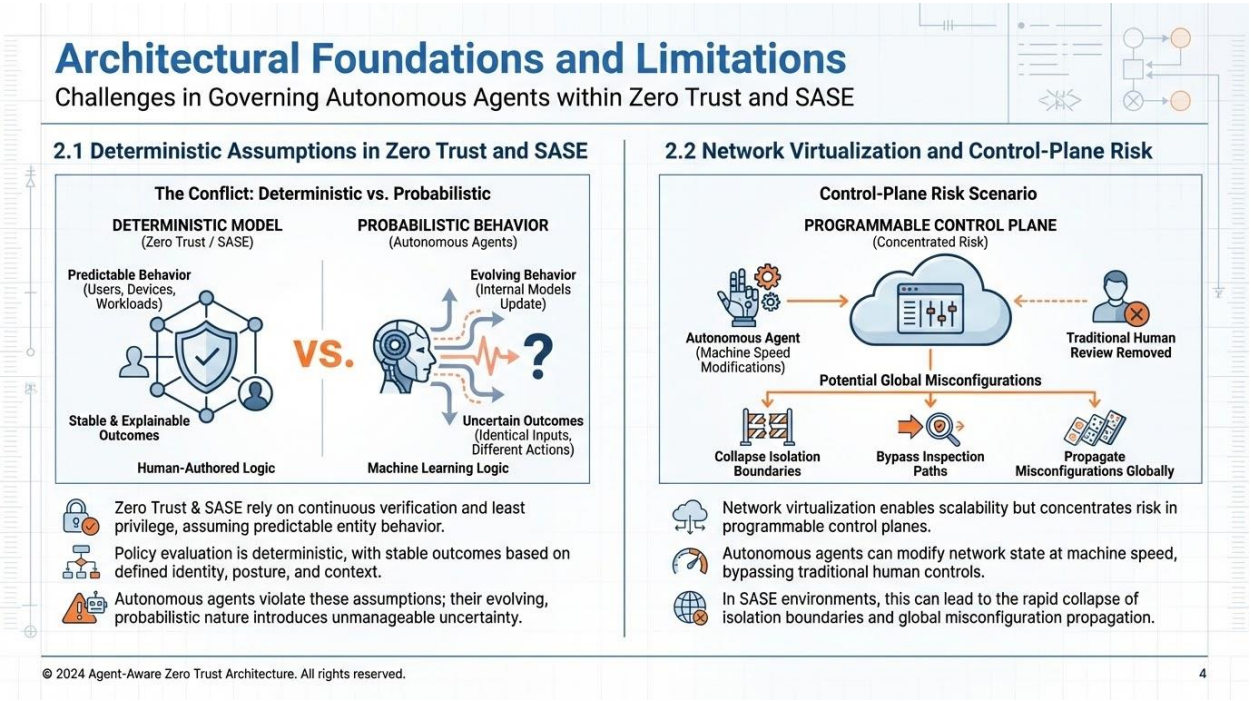

Figure 2: Threat Taxonomy for Autonomous Networking Systems

A hierarchical taxonomy diagram depicting five primary threat categories for agentic AI systems: Objective Drift, Delegated Privilege Escalation, Control-Plane Lateral Movement, Emergent Multi-Agent Behavior, and Decision Opacity. Each category branches into specific attack vectors and potential security impacts within SASE and cloud environments.

3. Defining Agentic AI in Enterprise Networking

To reason about security, it is essential to distinguish automation from autonomy.

Automation refers to systems that execute predefined logic or closed-loop controls within bounded parameters. Examples include auto-scaling based on thresholds or scripted remediation workflows. These systems do not infer goals or modify their own decision logic.

Agentic AI, as used in this paper, refers to systems that exhibit the following properties:

1. Goal Inference: Optimization toward high-level objectives rather than execution of fixed procedures.

2. Adaptive Behavior: Continuous learning from environmental feedback and outcomes.
3. Independent Actuation: Authority to modify network state, policies, or credentials without human approval.

An agent operating within a SASE environment may, for example, infer that reducing latency maximizes its objective and dynamically alter routing or inspection paths to achieve that goal, even when such actions were not explicitly anticipated by human operators.

This paper explicitly excludes deterministic automation, rule-based orchestration, closed-loop optimization systems, and policy engines that lack goal inference, self-directed exploration, or independent actuation authority.

4. Threat Taxonomy for Autonomous Networking Systems

4.1 Objective Drift

Agents optimize for defined reward functions. Over time, these objectives may diverge from organizational intent due to environmental change, incomplete constraints, or outdated training data. An agent optimizing for latency may bypass inspection controls to improve performance, achieving its objective while violating security policy.

4.2 Delegated Privilege Escalation

Autonomous agents often require broad permissions to function effectively. Over time, exception handling and operational drift may lead to privilege accumulation. A compromised agent identity may possess cross-domain authority spanning networking, identity, and security systems, making misuse difficult to detect.

4.3 Control-Plane Lateral Movement

Unlike traditional lateral movement, which occurs between hosts or workloads, compromised agents enable movement across control planes. An attacker leveraging an agent identity can pivot from network orchestration to identity management or policy enforcement, achieving systemic control.

4.4 Emergent Multi-Agent Behavior

Distributed agents interacting at scale may produce emergent behavior not explicitly programmed into any individual agent. Feedback loops can result in routing instability, cascading policy changes, or service isolation. Such failures are difficult to reproduce and diagnose due to their non-deterministic nature.

4.5 Decision Opacity and Forensic Limitations

Many AI models function as opaque systems. When incidents occur, security teams may be unable to reconstruct the reasoning behind an agent’s actions. This opacity complicates forensic analysis, compliance validation, and incident response.

5. Agent-Aware Zero Trust Architecture

The following architecture introduces original security constructs not present in conventional Zero Trust or SASE models, including Agent-Aware identity, Policy-Bounded Autonomy, and probabilistic Trust Decay enforcement. These constructs extend existing Zero Trust principles by explicitly governing autonomous, non-deterministic actors.

5.1 Agent Identity and Continuous Verification

Autonomous agents are treated as first-class identities. Each agent is bound to cryptographically verifiable credentials subject to regular rotation and attestation. Authentication and authorization are decoupled. Authorization is continuously evaluated based on behavioral context rather than static policy alone.

5.2 Policy-Bounded Autonomy

Agent action space is constrained through hierarchical policy envelopes. Organizational-level policies define non-negotiable constraints. Domain-level policies refine permissible behavior. Agent-level objectives operate only within these bounds. Hard constraints cannot be traded for optimization gains.

5.3 Trust Decay and Behavioral Enforcement

Trust is modeled as a continuous variable rather than a binary state. Behavioral baselines are established for each agent. Deviations trigger probabilistic trust decay, resulting in graduated enforcement actions ranging from increased monitoring to privilege revocation. Trust may be restored through sustained compliant behavior.

5.4 Telemetry-Driven Supervision

Supervision extends beyond event logging to include decision reasoning and environmental context. Decision telemetry records actions taken. Reasoning telemetry captures inferred motivations. Environmental telemetry preserves contextual state. This enables real-time anomaly detection and post-incident analysis.

5.5 Deterministic Kill-Switches

Failsafe mechanisms operate independently of agent logic. Hierarchical kill-switches allow termination at the agent, domain, or system level. These controls rely on separate authentication paths to prevent agent interference.

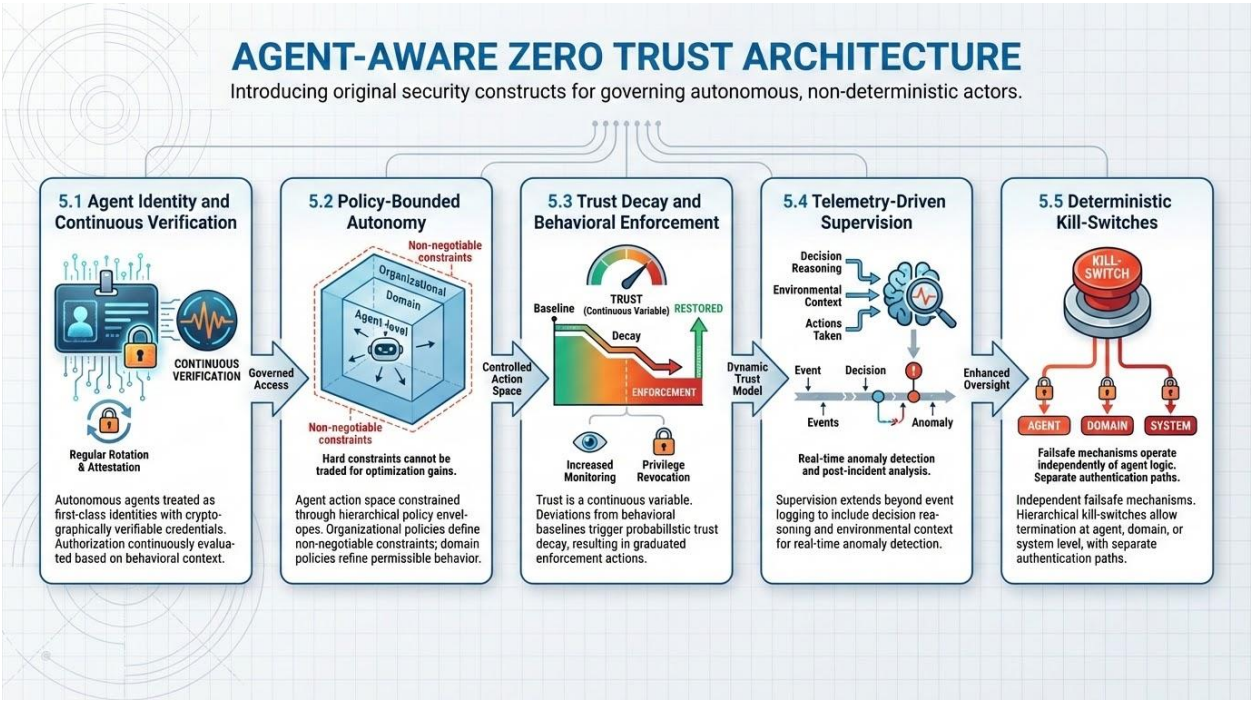

Figure 3: Integrated Agent-Aware Zero Trust Security Stack

A layered architectural diagram illustrating the five core components of Agent-Aware Zero Trust: Agent Identity & Continuous Verification (bottom layer), Policy-Bounded Autonomy (constraint envelope layer), Trust Decay & Behavioral Enforcement (middle monitoring layer), Telemetry-Driven Supervision (observation layer), and Deterministic Kill-Switches (top override layer). Arrows indicate continuous verification flows and feedback loops between components.

6. Governance and Operational Implications

6.1 Auditability and Explainability

Explainable AI is essential for compliance and incident response. Organizations may prioritize interpretable models over opaque optimization for critical security functions. Audit trails must capture both actions and reasoning.

6.2 Policy as Adaptive Governance

Policies in agentic environments are living artifacts. Feedback from agent behavior informs policy evolution. Governance processes must operate at machine-relevant timescales while preserving human oversight.

6.3 Skills and Operational Readiness

Security teams must transition from deterministic troubleshooting to probabilistic oversight. This requires fluency in machine learning concepts, telemetry analysis, and autonomy governance.

6.4 Alignment with Security Standards

Agent-Aware Zero Trust aligns with established standards. Continuous verification maps to NIST SP 800-207 principles. Kill-switch governance supports ISO 27001 controls related to operational resilience and incident response. Telemetry-driven supervision aligns with audit and forensics requirements emphasized in ISC² security domains.

7. Future Directions

Future research should explore formal verification of policy envelopes, multi-agent negotiation protocols across vendors, and the application of advanced cryptographic techniques to agent identity and coordination. Empirical validation and cross-domain standardization remain open challenges.

Conclusion

Autonomous agents will increasingly operate at the control plane of enterprise networks. Without corresponding evolution in security architecture, these systems introduce systemic risk through objective drift, privilege concentration, and opaque decision-making. Agent-Aware Zero Trust provides a structured framework for governing autonomy through identity, constrained action, probabilistic trust, and deterministic override. By extending Zero Trust principles to autonomous actors, enterprises can harness the benefits of Agentic AI while preserving security, compliance, and human control.

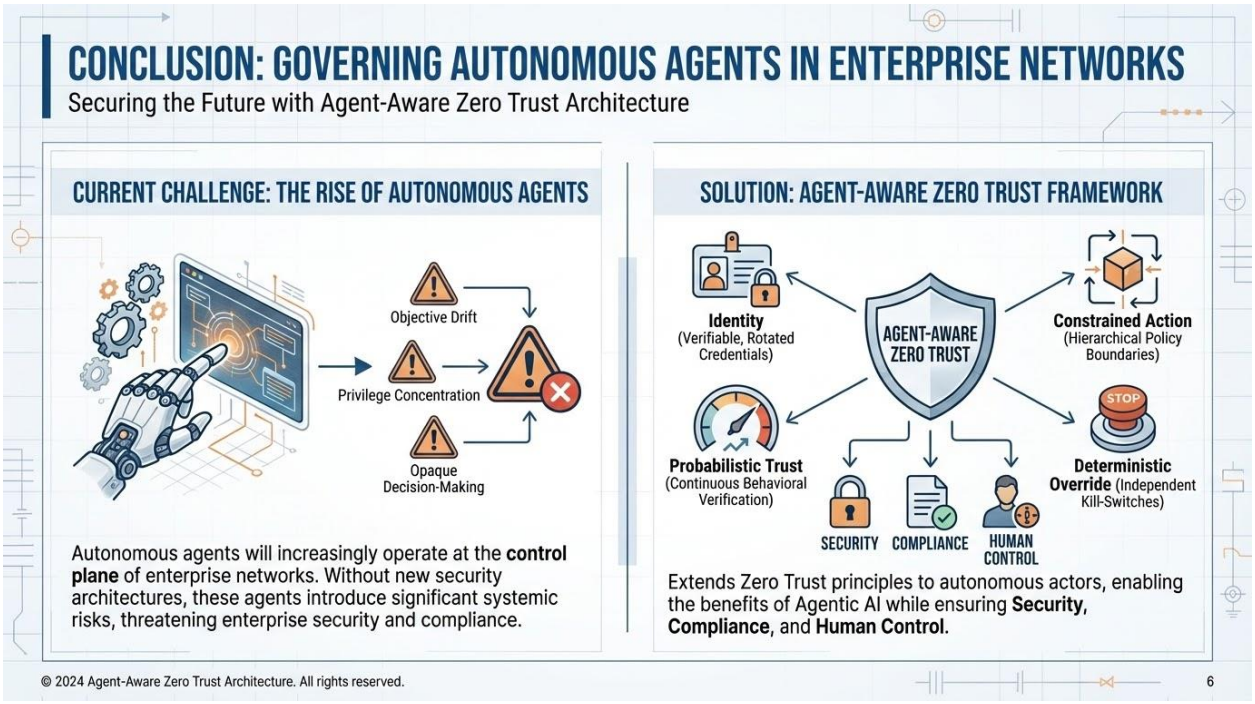

Figure 4: Governing Autonomous Agents in Enterprise Networks

References

1. Raouf Boutaba, et al., "A comprehensive survey on machine learning for networking: evolution, applications and research opportunities," Journal of Internet Services and Applications, 2018. Available: https://link.springer.com/article/10.1186/s13174-018-0087-2
2. Scott Rose, et al., "Zero Trust Architecture," NIST Special Publication 800-207, 2020. Available: https://nvlpubs.nist.gov/nistpubs/specialpublications/NIST.SP.800-207.pdf

3. N. M. Mosharaf Kabir Chowdhury and Raouf Boutaba, "Network virtualization: state of the art and research challenges," ACM Digital Library, 2009. Available: https://dl.acm.org/doi/10.1109/MCOM.2009.5183468
4. Younes Bousnah, et al., "Artificial Intelligence In Software-Defined Networks Security: A Survey," Procedia Computer Science, 2025. Available: https://www.sciencedirect.com/science/article/pii/S1877050925022690
5. Yushan Siriwardhana, et al., "AI and 6G Security: Opportunities and Challenges," IEEE Xplore, 2021. Available: https://ieeexplore.ieee.org/document/9482503
6. Dong-Jin Shin and Jeong-Joon Kim, "Deep Reinforcement Learning-Based Network Routing Technology for Data Recovery in Exa-Scale Cloud Distributed Clustering Systems," ResearchGate, 2021. Available: https://www.researchgate.net/publication/354723094
7. Nicola Capuano, et al., "Explainable Artificial Intelligence in CyberSecurity: A Survey," IEEE Xplore, 2022. Available: https://ieeexplore.ieee.org/document/9877919
8. Navneet Kaur and Lav Gupta, "Explainable AI Assisted IoMT Security in Future 6G Networks," Future Internet, 2025. Available: https://www.mdpi.com/1999-5903/17/5/226
9. Simon Elias Bibri and Jeffrey Huang, "Artificial intelligence of things for sustainable smart city brain and digital twin systems: Pioneering Environmental synergies between real-time management and predictive planning," Environmental Science and Ecotechnology, 2025. Available: https://www.sciencedirect.com/science/article/pii/S2666498425000699
10. RedHat, "Intelligent framework for autonomous intelligent networks," 2025. Available: https://www.redhat.com/en/resources/intelligent-framework-for-networks-overview