Power System Technology
Volume 50, No. 1, 2026
ISSN: 1000-3673

Towards Trustworthy Agentic AI in Healthcare: A Zero Trust-Based Security Framework

Francisco Agballog
Securitas Technology USA, Healthcare Solutions Engineering

Luis Botero
Department of Information Communication Services, The Church of Jesus Christ of Latter-day Saints, Salt Lake City, UT, USA

Abstract

The swift deployment of agentic artificial intelligence (AI) systems in healthcare settings is transforming the process of clinical, operational, and infrastructure-level decision-making. In contrast to traditional AI models, agentic AI systems have autonomous reasoning, dynamic task execution, and multi-system interaction features, allowing them to be active agents in healthcare processes. Although these capabilities present great efficiency and scalability opportunities, they also increase the attack surface and present new types of security and trust threats. In particular, AI agents have access to sensitive patient data without human intervention, can access external services, and influence clinical outcomes, which are critical issues when it comes to identity assurance, decision integrity, and software verification.

The existing security solutions, including the traditional perimeter-based security and the commonplace Zero Trust Architecture (ZTA), are not well-equipped to support the behavioral and operational complexities of agentic AI systems. The approaches that have been implemented are more or less user-centric or device-centric trust validation and little consideration is being given to autonomous software entities that can evolve, adapt and act without the close supervision of a human being. This is especially a problem in healthcare facilities, where the sensitivity of the data, regulatory requirements and patient safety place dire limitations on the reliability and accountability of the system.

To overcome these issues, this paper presents the TAZAI framework (Trustworthy Agentic Zero Trust Architecture for AI in Healthcare), a new security framework that aims to implement a consistent trust validation of all layers of agentic AI activities. The proposed framework broadens the concepts of Zero Trust to encompass agent identity verification, context-sensitive policing, real-time behavior monitoring, and data management in a safe way. With the combination of these mechanisms into a single architecture, TAZAI allows controlling the actions of AI agents finely, without compromising the interoperability of systems between cloud and on-premise healthcare infrastructures.

The effectiveness of the framework is demonstrated with systematic threat model and a scenario of healthcare deployment, demonstrating how TAZAI mitigates threats, such as unauthorized data access, prompt injection attacks, and autonomous decision manipulation. The findings reveal that incorporating the principles of Zero Trust into agentic AI processes can greatly improve the resilience of the system, minimize its exposure to new attack vectors, and create a verifiable trust boundary of autonomous

activity. This publication adds a methodology towards ensuring future AI-based healthcare systems and sets the groundwork of future studies in reliable autonomous computing.
Key Words: Agentic AI Security: Zero Trust Architecture in Healthcare: Autonomous AI Risk Management: AI Identity and Access Control: Healthcare Data Protection: AI Behavior Monitoring: Secure Cloud Integration.

1. Introduction

The advent of agentic artificial intelligence (AI) is a paradigm shift in the functioning of the computational systems in complex environments. Unlike the classical models of AI which are designed to execute only a particular task, yet cannot reason and make decisions dynamically and execute tasks in more than one step using an interconnected system, agentic AI systems are capable of doing so. Large language models (LLMs) and orchestration systems can learn or drive these systems and can act or react to external services, and change their behavior based on contextual inputs. This has rendered agentic AI to be a disruptive technology in many sectors, particularly in areas where optimization needs are inevitable and those that necessitate real-time reactions [5].

Agents AI in health care is a growing trend as additional clinical and operational workflows are ready to be smartly automated. Clinical decision support systems, patient tracking systems, and computer-assisted diagnosis are some of the applications that are beginning to implement AI agents capable of communicating with electronic health records (EHRs), medical devices, and analytics systems that operate on the cloud. These systems are characterized by great benefits such as increased efficiency, minimized human error and increased scalability of healthcare services. However, this comes with another complication when they are implemented as AI agents are no longer tools but agents in the decision-making process that directly affect patient outcomes [20].

This autonomy of agentic AI systems raises serious security and trust issues that cannot be properly dealt with by existing cyber security frameworks. One of the most significant concerns is that the AI agents might be carrying out activities without the immediate supervision of humans and, therefore, may cause unintended or even malicious actions. To illustrate, intruded agents may initiate unauthorized access to the data, distort clinical recommendations, or disseminate false outputs among systems that are connected. Also, the data could be revealed by the usage of third-party-data-sources and APIs, particularly in a healthcare facility where patient data information is supposed to be very confidential [19].

The other significant risk pertains to the increasing attack surface which is associated with agentic AI. The systems operate on different layers which include data pipelines, model inference engines, application interfaces and cloud infrastructures. There are vulnerabilities in each layer that enemies can exploit through such techniques as prompt injection, adversarial manipulation and API abuse. Such attacks can compromise the integrity of AI-generated outputs, but not the sole aspect since the efficiency of the entire healthcare system may be greatly impaired due to such attacks [15].

Traditional security models particularly the ones, which are perimeter-based, are just not effective to address these issues. The models are based on predefined limits of trust whereby entities in a secured network can be trusted after authentication. However, these are no longer true in the current healthcare ecosystems where cloud services are distributed, with interdependent devices and autonomous AI agents. Even conventional forms of the Zero Trust Architecture (ZTA) are more robust yet are simply oriented towards human-users and devices, and lack support mechanisms to continually assess the conduct and decision-making of autonomous AI actors [31].

The recent evolution of Zero Trust models has shown to be an efficient method of improving the security of healthcare by limiting access and identity authentication, and a continuous monitoring system. Nonetheless, these applications are still mostly centered on entities and fail to take into consideration the dynamism and adaptability of agentic AI systems. The absence of a particular framework that will bind the principles of a Zero Trust and mechanisms of behavioral validation, which are unique to AI, is a notable gap in the current research and practice [35].

To fill this gap, the current paper suggests a new security architecture TAZAI (Trustworthy Agentic Zero Trust Architecture for AI in Healthcare), which is an agentic AI architecture-specific security architecture. The framework is based on the traditional concepts of Zero Trust as it incorporates continuous identity verification of AI agents, flexible policy implementation, real-time behavioral verification, and secure data management platforms, which are unique to healthcare systems. By introducing a curated approach to trust control on all levels of AI-based work, TAZAI will assist healthcare organizations to provide maximum security and harness the opportunities of autonomous AI technologies.

The main works of this work are three-fold. First, it introduces an in-depth discussion of the security threats of agentic AI in the healthcare sector, which shows the shortcomings of the current strategies. Second, it introduces

TAZAI framework as a novel and multi-layered architecture with the principles of Zero Trust and AI-specific security requirements. Third, it also demonstrates how the suggested framework can be applied to the actual healthcare context to enhance the resilience of the system, protect the sensitive information, and offer trustworthy AI-based decision-making.

2. Background and Related Work.

The convergence of artificial intelligence, cloud computing, and healthcare infrastructure has been very fast, and thus, it has heightened the need to have strong security models that can deal with new risks. Despite the enormous progress in other areas of technology such as Zero Trust Architecture and AI-based system design, the existing solutions remain uncoordinated in the environment of autonomous AI systems in healthcare. Here, the innovative ideas are discussed, and vital limitations that motivate the suggested framework are determined.

2.1 The concept of Zero Trust Architecture (ZTA)

One of the new cyber security paradigms that have been created to address the deficiencies of the old perimeter-based security models is Zero Trust Architecture (ZTA). ZTA is built on the idea of never trust, always verify and eliminates implicit trust across network boundaries and offers unrelenting authentication, authorization and validation of all entities attempting to communicate with a system. This is particularly true in distributed environments, where there are users, machines, and services distributed across different domains and multiple infrastructures [31].

ZTA has become popular in medical systems to increase data protection and data access controls. The approach through which healthcare organizations can maintain sensitive resources like electronic health records (EHRs) and clinical systems with authenticated and authorized access by only authenticated entities is by adopting identity-centric security policies. Recent reports indicate that ZTA is a great way to strengthen the security of healthcare data through strict access control and reduce cross-boundary traffic in networks [4].

Although it has its merits, the existing implementations of ZTA are mainly geared towards human users and controlled devices. They do not take into account the independent software agents such as AI agents which are able to autonomously activate and communicate with a variety of system components. As a result, the issue of the extension of ZTA to consider the dynamicity of AI agentic systems remains unsolved.

2.2 Agentic AI Systems

The agentic AI systems are novel category of intelligent systems that has the capability of reasoning independently, decision making and even executing tasks without constant monitoring by a human. Unlike traditional AI systems, which can be trained to operate in a limited system, agentic systems apply recent techniques of machine learning and large language models (LLMs) to dynamically comprehend the goals, generate behavior, and adapt to changing environments. They can be employed as dynamic agents of complex processes, but not as inert analytical tools, due to these functions [5].

The agentic AI system architecture is typically defined by coordinating layers in which, it interacts with models, external APIs, databases and application services. This inter-relational architecture allows AI agents to complete multi-step tasks, such as obtaining patient data, inferring on clinical data, and providing recommendations in real-time. The level of autonomy also introduces new dimensions of risk although it does enhance the efficiency of operations; it does so in situations where agents are put in sensitive and mission critical situations such as healthcare [16].

The growing use of agentic AI in areas that are critical in decision making has brought up issues of accountability, control and security. When these systems are to be employed in making clinical decisions and determining patient outcomes, it is important to make sure they behave in a certain way within a set of limits.

2.3 Healthcare AI Security Problems.

The introduction of AI systems in medical settings poses special security concerns because of the sensitivity of medical information and the intricacy of healthcare systems. One of the most crucial problems is the risk of the exposure of the protected health information (PHI). AI systems need high amounts of patient data to train and make inferences, and it is more likely to be accessed unauthorized, leaked, or abused unless sufficient controls are implemented [1].

Besides the data protection issues, the healthcare organizations have to adhere to the strict regulatory and ethical standards of using the information about the patients. These laws put high-level restrictions on access, storage and processing of data and it is important that AI systems are run on clearly defined security and compliance procedures. The failure to meet these requirements has some consequences such as laws and loss of trust among the stakeholders [30].

Interoperability of systems is the other big problem. The current healthcare setting is a complex network of both medical devices, cloud systems, and hospital information systems. Even though this interconnectedness enables

data to flow without any issues and improved services to be offered, the interconnectedness also enlarges the attack surface and forms a number of vulnerability points. Securing such multi-layered systems is extremely complicated and requires an integrated approach that goes beyond traditional security systems [27].

2.4 Gap Analysis

Even though the design of the Zero Trust Architecture and AI system has been enhanced, this is where there exists a huge gap between the two domains. Existing security systems fail to sufficiently consider the special issues of agentic AI systems, especially in areas with high stakes like healthcare. Although ZTA offers a powerful basis of identity and access control, it does not offer options of constantly validating the actions and decisions of autonomous AI agents.

The existing research has observed the need to develop better security models with behavioral monitoring, dynamic trust assessment, and real-time risk assessment of AI systems. Nevertheless, these initiatives are still narrow in their scope and are not specifically adjusted to the specifics of operational healthcare settings where data sensitivity and reliability of the system are the highest priorities [35].

Moreover, no integrated frameworks can ensure AI operations on a variety of layers such as data, models, APIs, and infrastructure. Lack of a common framework on how to deal with trust in agentic AI systems exposes the system to vulnerabilities that can be abused by enemies and this may result in compromised system integrity and patient safety.

In order to overcome these shortcomings, it is evident that a special framework is required to integrate the principles of Zero Trust with AI-specific security measures to provide continuous verification, dynamic implementation of policies, and end-to-end security of all healthcare systems layers. The proposed TAZAI model will help fill this essential gap by providing a systematic and scalable way of ensuring agentic AI in the medical environment.

3. Threat Model for Agentic AI in Healthcare

The integration of agentic artificial intelligence into healthcare environments introduces a significantly expanded and dynamic threat landscape that extends beyond traditional cyber security concerns. Unlike conventional software systems, agentic AI operates with a high degree of autonomy, continuously interacting with data sources, application programming interfaces (APIs), and distributed infrastructure components. This level of autonomy increases the system's exposure to complex and multi-layered attack surfaces, thereby elevating risks to system integrity, data confidentiality, and operational availability.

In healthcare settings, these risks are further amplified due to the sensitivity of protected health information (PHI) and the critical nature of clinical decision-making processes. As agentic AI systems increasingly participate in diagnosis, monitoring, and treatment recommendations, any compromise in their operation can have direct implications for patient safety and regulatory compliance.

To systematically analyze these risks, this section presents a structured threat model that categorizes key security threats associated with agentic AI in healthcare. Each threat is examined in terms of its description, attack vectors, and potential impact, providing a foundation for the design of robust security mechanisms within the proposed framework. The threat categorization is also aligned with emerging risks in large language model (LLM) systems, including prompt injection and adversarial manipulation, as highlighted in recent security guidelines [27].

Table 1: Threat Categories in Agentic AI Healthcare Systems

Threat ID	Threat Name	Target Layer	Primary Risk
T1	Agent Identity Spoofing	Identity & Access Layer	Unauthorized system access
T2	Prompt Injection & Model Manipulation	Model & Interaction Layer	Compromised AI behavior
T3	Unauthorized Data Access (PHI Leakage)	Data Layer	Sensitive data exposure
T4	API Exploitation	Application Layer	System compromise
T5	Autonomous Decision Abuse	Decision Layer	Unsafe or malicious outcomes

3.1 T1: Agent Identity Spoofing

Description

Agent identity spoofing occurs when a malicious entity impersonates a legitimate AI agent to gain unauthorized access to healthcare systems. In agentic environments, AI agents frequently interact with multiple services and components, often without direct human oversight. If identity verification mechanisms are weak or improperly implemented, attackers can introduce rogue agents that operate within trusted system boundaries.

Attack Vector

Attackers typically exploit vulnerabilities in authentication mechanisms, such as compromised credentials, token reuse, or weak identity validation protocols. In cloud-based healthcare infrastructures, misconfigured Identity and Access Management (IAM) systems further increase this risk by allowing unauthorized entities to assume valid roles or privileges. Weak enforcement of identity policies can enable persistent unauthorized access across interconnected services [8].

Impact

Successful identity spoofing can lead to unauthorized access to sensitive patient data, manipulation of system workflows, and disruption of clinical operations. Beyond immediate technical damage, such breaches undermine trust in AI-assisted healthcare systems and may lead to significant regulatory and legal consequences.

3.2 T2: Prompt Injection and Model Manipulation

Description

Prompt injection and model manipulation attacks target the decision-making logic of agentic AI systems by altering the inputs or contextual instructions provided to the model. These attacks exploit the reliance of AI agents on external prompts and dynamic data inputs to generate outputs, making them particularly vulnerable in open or semi-structured environments.

Attack Vector

Adversaries can introduce malicious instructions through user inputs, compromised APIs, or manipulated external data sources. These inputs may override system-level constraints, causing the AI agent to disclose sensitive information, bypass security controls, or execute unintended actions. Such attacks are especially effective in systems where input validation and contextual filtering are insufficient [15].

Impact

The consequences of prompt injection attacks include corrupted outputs, unauthorized data disclosure, and loss of control over AI-driven processes. In healthcare, this can translate into incorrect clinical recommendations, exposure of PHI, and compromised patient safety. Given the increasing reliance on AI-assisted decision-making, such attacks pose a critical risk to both operational integrity and patient outcomes.

3.3 T3: Unauthorized Data Access (PHI Leakage)

Description

Unauthorized data access refers to the exposure or leakage of protected health information due to inadequate access controls or vulnerabilities within AI-driven workflows. Agentic AI systems often require continuous access to sensitive datasets, increasing the risk of unintentional disclosure or malicious exploitation.

Attack Vector

This threat can arise from misconfigured access policies, insufficient data segmentation, or vulnerabilities in system architecture. Additionally, AI agents that aggregate data from multiple sources may inadvertently expose sensitive information if proper controls are not enforced. Weak enforcement of data governance policies further exacerbates this risk [1].

Impact

PHI leakage can result in severe legal, financial, and reputational consequences for healthcare organizations. It also violates regulatory requirements and erodes patient trust, making it one of the most critical risks in healthcare cyber security.

3.4 T4: API Exploitation

Description

API exploitation targets the communication interfaces that enable agentic AI systems to interact with external services, databases, and applications. While APIs are essential for system functionality, they also represent a significant attack surface if not properly secured

Attack Vector

Attackers may exploit insecure API endpoints, weak authentication mechanisms, or inadequate input validation to gain unauthorized access or manipulate system behavior. In highly interconnected healthcare environments, APIs often serve as gateways to critical resources, making them attractive targets for adversaries [27].

Impact

Successful API exploitation can lead to unauthorized data access, service disruption, and full system compromise. Due to the interconnected nature of agentic AI systems, such attacks can propagate rapidly across multiple components, amplifying their overall impact.

3.5 T5: Autonomous Decision Abuse

Description

Autonomous decision abuse involves the manipulation or exploitation of an AI agent's decision-making capabilities. As agentic AI systems are designed to operate independently, any compromise in their logic or control mechanisms can lead to unintended or malicious actions.

Attack Vector

This threat may originate from adversarial inputs, compromised models, or insufficient monitoring mechanisms. Attackers can exploit these weaknesses to influence AI behavior, causing the system to execute harmful or unauthorized decisions without detection [19].

Impact

In healthcare environments, the consequences of autonomous decision abuse are particularly severe. Erroneous or manipulated decisions can directly affect diagnosis, treatment planning, and patient outcomes. Such failures not only endanger patient safety but also raise significant ethical and legal concerns regarding the deployment of autonomous systems.

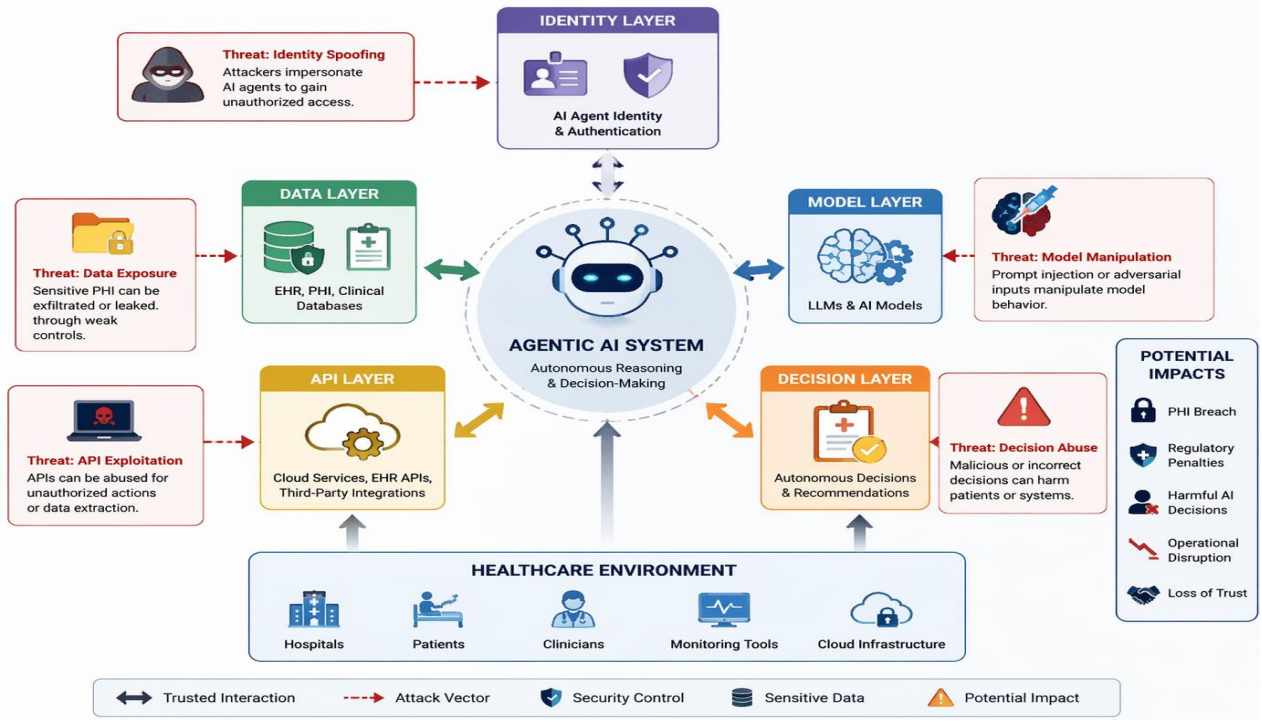

Figure 1: Threat landscape of agentic AI systems in healthcare, illustrating attack surfaces across identity, data, and model, API, and decision layers

4. Research Methodology

This study adopts a Design Science Research (DSR) methodology to guide the systematic development and evaluation of the proposed TAZAI framework. DSR is widely recognized in cyber security and intelligent systems research as an effective approach for creating and validating artifacts that address complex, real-world problems through iterative design, refinement, and assessment. In the context of this work, the artifact is a security architecture tailored to mitigate the unique risks introduced by agentic AI in healthcare environments.

The selection of DSR is motivated by the need to bridge the gap between theoretical security models and practical implementation requirements in highly sensitive domains such as healthcare. Unlike purely empirical or observational approaches, DSR enables the structured creation of innovative solutions while ensuring that they are grounded in both existing knowledge and real-world applicability.

The research process is organized into three interrelated phases: problem identification, framework design, and evaluation, each contributing to the overall validity and applicability of the proposed solution.

4.1 Problem Identification

The first phase focuses on identifying critical security challenges associated with the deployment of agentic AI systems in healthcare environments. Through analysis of existing literature and emerging threat patterns, several key gaps are identified.

First, current security models lack mechanisms for continuous verification of autonomous AI agents, which operate independently and interact dynamically with multiple system components. Traditional Zero Trust implementations primarily focus on human users and static workloads, leaving AI-driven entities insufficiently governed. This limitation becomes particularly significant in healthcare systems where AI agents access sensitive patient data and participate in clinical workflows.

Second, agentic AI systems are highly vulnerable to prompt injection and adversarial manipulation, which can alter model behavior and bypass established safeguards. These risks are increasingly recognized in modern AI security frameworks and represent a fundamental challenge in ensuring trustworthy AI operations [27].

Third, there is inadequate integration of trust enforcement mechanisms across distributed healthcare infrastructures, especially in hybrid environments combining on-premises systems and cloud services. Existing approaches often treat identity, data protection, and system monitoring as separate concerns rather than as components of a unified trust model.

These identified gaps highlight the need for a specialized framework that extends Zero Trust principles to address the unique characteristics of agentic AI systems.

4.2 Framework Design

The second phase involves the design of the TAZAI framework as a multi-layered security architecture that integrates Zero Trust principles with AI-specific controls. The design process is guided by both the threat model presented in Section 3 and established security standards.

The framework development follows a structured approach:

• **Mapping identified threats to corresponding security controls**, ensuring that each threat category is explicitly addressed within the architecture

• **Defining modular architectural layers**, including identity management, policy enforcement, behavioral monitoring, and secure data handling

• **Incorporating context-aware and adaptive validation mechanisms**, enabling dynamic decision-making based on system state, user context, and agent behavior

A key aspect of the design is the extension of traditional Zero Trust principles—such as continuous verification, least privilege, and strict access control—to autonomous AI agents. This ensures that trust is not statically assigned but continuously evaluated throughout the lifecycle of each interaction.

The framework is conceptually aligned with established Zero Trust standards, particularly the principles outlined in the National Institute of Standards and Technology Zero Trust Architecture (SP 800-207), which emphasizes identity-centric security and continuous validation [23]. In addition, the design incorporates guidance from modern AI governance frameworks, including the National Institute of Standards and Technology Artificial Intelligence Risk Management Framework (AI RMF 1.0), ensuring that risk-aware and trustworthy AI practices are embedded within the architecture [24].

4.3 Evaluation Strategy

The final phase focuses on evaluating the effectiveness and robustness of the proposed framework using a multi-dimensional assessment approach. Given the conceptual nature of the study, the evaluation combines formal modeling techniques with scenario-based validation to ensure both theoretical rigor and practical relevance.

First, a STRIDE-based threat modeling approach is employed to systematically assess the framework's ability to mitigate key security risks. Each threat category identified in Section 3 is mapped against STRIDE dimensions—Spoofing, Tampering, Repudiation, Information Disclosure, Denial of Service, and Elevation of Privilege—to evaluate coverage and identify potential gaps.

Second, a comparative analysis is conducted between the proposed TAZAI framework, traditional perimeter-based security models, and baseline Zero Trust implementations. This comparison highlights the relative strengths of TAZAI in addressing AI-specific risks, particularly those related to autonomy, dynamic behavior, and multi-layer system interactions.

Third, a scenario-based validation is performed using a representative healthcare use case involving an AI-driven clinical assistant. This evaluation demonstrates how the framework enforces continuous identity verification, policy validation, and behavioral monitoring during real-time interactions, thereby reducing the likelihood of unauthorized access and malicious system behavior.

Together, these evaluation strategies provide a comprehensive assessment of the framework's effectiveness, ensuring that it is both theoretically sound and practically applicable in real-world healthcare environments.

5. Proposed Framework: TAZAI

The proposed Trustworthy Agentic Zero Trust Architecture for AI (TAZAI) framework is designed to address the unique security challenges introduced by agentic AI systems operating in healthcare environments. While traditional Zero Trust Architecture (ZTA) emphasizes identity-centric access control and continuous verification, it does not explicitly account for the autonomous and adaptive behavior of AI agents. TAZAI extends these principles by incorporating behavior-aware trust evaluation, dynamic risk adaptation, and continuous validation of AI decision-making processes.

The key novelty of TAZAI lies in its integration of behavioral intelligence into Zero Trust enforcement, enabling real-time assessment of whether an AI agent's actions remain consistent with expected operational patterns. This distinguishes TAZAI from conventional ZTA implementations, which primarily focus on static identity verification and access control policies without evaluating the evolving behavior of autonomous systems.

The framework enforces security across five tightly integrated layers: identity, policy, behavior, data, and infrastructure. These layers collectively ensure that trust is continuously established, evaluated, and adapted throughout the lifecycle of AI-driven interactions. The design aligns with established Zero Trust principles as defined by the National Institute of Standards and Technology Zero Trust Architecture (SP 800-207), while extending them to support AI-specific operational requirements [23].

5.1 Design Principles

TAZAI is guided by a set of foundational principles that redefine trust management for autonomous AI systems.

Continuous Verification: Every interaction involving an AI agent is dynamically evaluated rather than implicitly trusted. Trust decisions are recalculated in real time based on identity, behavior, and contextual signals, ensuring that no entity maintains persistent trust without validation.

Least Privilege: AI agents are granted only the minimum level of access required to perform their tasks. This reduces the potential impact of compromised agents and limits lateral movement within healthcare systems.

Context-Aware Access Control: Access decisions are influenced by real-time contextual factors, including system state, user intent, environmental conditions, and risk levels. This ensures adaptive security enforcement in dynamic healthcare environments.

AI Behavior Validation: Unlike traditional systems, TAZAI continuously monitors the outputs and actions of AI agents to detect deviations from expected behavior. This principle introduces a behavioral dimension to Zero Trust, addressing risks such as prompt injection and adversarial manipulation.

Together, these principles ensure that trust is not statically assigned but continuously established through multi-dimensional evaluation.

5.2 TAZAI Architecture Layers

The TAZAI framework is implemented as a five-layer architecture, with each layer responsible for enforcing specific aspects of security while contributing to an integrated trust evaluation process.

Layer 1: Identity & Trust Layer: This layer establishes verifiable identities for AI agents using cryptographic authentication mechanisms, such as token-based validation and certificate-based identity management. Each agent is assigned a unique identity that is continuously verified during system interactions.

In addition to identity validation, this layer generates an initial trust baseline for each agent, which serves as an input to the dynamic trust scoring model. Robust identity enforcement mechanisms are critical in preventing impersonation attacks and unauthorized system access, particularly in distributed healthcare environments [9].

Layer 2: Policy Enforcement Layer: The policy enforcement layer implements **dynamic and context-aware access control mechanisms**. Instead of relying on static policies, access decisions are evaluated in real time based on multiple factors, including agent role, trust score, request sensitivity, and environmental context.

Policies are enforced through fine-grained authorization models, ensuring that each request is validated against predefined security constraints. This layer acts as a critical control point for preventing unauthorized access and enforcing compliance with healthcare data protection requirements.

Layer 3: AI Behavior Monitoring Layer: This layer represents a key innovation of the TAZAI framework by introducing real-time behavioral validation for AI agents. Unlike traditional security systems, which focus primarily on access control, this layer continuously evaluates whether an AI agent's actions remain consistent with expected operational patterns.

The monitoring process incorporates multiple analytical techniques:

- **Statistical Drift Detection:** Detects distributional changes in model inputs and outputs, identifying potential data poisoning or environmental shifts.
- **Anomaly Detection Models:** Machine learning techniques such as Isolation Forests and Auto encoders are used to identify abnormal behavior patterns that deviate from learned baselines.
- **Rule-Based Policy Validation:** Predefined rules ensure that AI outputs comply with domain-specific constraints, particularly in clinical decision-making contexts.
- **Sequential Behavior Analysis**: Temporal analysis of action sequences is used to detect suspicious or abnormal chains of operations that may indicate adversarial manipulation.

These mechanisms collectively enable the detection of prompt injection attacks, adversarial inputs, and unauthorized system actions, thereby enhancing the resilience of AI-driven processes against evolving threats. The importance of defending against adversarial manipulation in AI systems has been widely recognized in recent security research [5].

Layer 4: Secure Data Layer: The secure data layer ensures the protection of sensitive healthcare information through a combination of encryption, data segmentation, and controlled access mechanisms. Data is protected both at rest and in transit, with strict enforcement of access policies to prevent unauthorized exposure.

This layer is particularly critical for safeguarding PHI, as healthcare data breaches can have severe regulatory and ethical consequences. By integrating data protection mechanisms directly into the Zero Trust framework, TAZAI ensures that data security is not treated as an isolated concern but as an integral component of trust enforcement.

Layer 5: Infrastructure Layer: The infrastructure layer supports the deployment of the TAZAI framework across hybrid healthcare environments, including cloud-based platforms and on-premise systems. This layer ensures consistent security enforcement regardless of where system components are hosted.

It also facilitates secure communication between system components, enabling seamless integration of AI agents with electronic health record (EHR) systems, cloud APIs, and monitoring tools. The ability to enforce Zero Trust principles across distributed environments is essential for modern healthcare systems, which increasingly rely on interconnected infrastructure.

Table 3: Mapping of TAZAI Layers to Security Functions

Layer	Core Function	Security Objective
Identity & Trust Layer	Identity verification	Prevent unauthorized access
Policy Enforcement Layer	Dynamic access control	Enforce least privilege
AI Behavior Monitoring Layer	Behavioral validation	Detect anomalies and attacks
Secure Data Layer	Data protection	Prevent PHI leakage
Infrastructure Layer	System integration	Ensure secure interoperability

5.3 Formal Trust Model

To operationalize Zero Trust principles for autonomous AI systems, TAZAI introduces a dynamic trust scoring mechanism that continuously evaluates the trustworthiness of each agent.

Let:

- $T_a(t)$: Trust score of agent a at time t
- I_a : Identity confidence score
- $B_a(t)$: Behavioral consistency score
- $C_a(t)$: Contextual compliance score

The trust score is defined as:

$$T_a(t)=\alpha I_a+\beta B_a(t)+\gamma C_a(t)$$

Where:

$$\alpha+\beta+\gamma=1$$

This formulation ensures that trust is determined by a weighted combination of identity assurance, behavioral reliability, and contextual compliance.

Trust Decay Mechanism

To account for evolving risk conditions, TAZAI incorporates a trust decay function:

$$T_a(t+1)=T_a(t)\cdot e^{-\lambda R}$$

Where:

R : Risk severity associated with recent actions

λ : Decay constant controlling sensitivity to risk

This mechanism ensures that trust is dynamically reduced in response to suspicious or high-risk behavior, preventing persistent trust in potentially compromised agents.

Access Decision Rule

Access decisions are governed by a predefined trust threshold τ:

$$T_a(t) \geq \tau$$

An AI agent is granted access only if its trust score meets or exceeds the threshold. This enforces continuous verification and ensures that access privileges are tightly coupled with real-time trust evaluation.

Table 4: Trust Model Components and Interpretation

Component	Description	Security Role
Identity Score	Confidence in agent identity	Prevent impersonation
Behavioral Score	Consistency of agent actions	Detect anomalies
Context Score	Compliance with environment	Enforce context-aware security
Trust Score	Aggregated trust value	Basis for access decisions

5.4 Multi-Agent Trust Extension

In real-world healthcare systems, multiple AI agents often collaborate across workflows, sharing data and coordinating actions. To address this complexity, TAZAI extends its trust model to support multi-agent environments.

The trust score for a group of agents is defined as:

$$T_{\text{group}}=\frac{1}{n}\sum_{i=1}^{n} T_{a_i}$$

Where n represents the number of agents in the group.

This formulation ensures that the overall system trust reflects the behavior of all participating agents. If one or more agents exhibit anomalous or malicious behavior, the collective trust score is reduced, triggering stricter access controls or intervention mechanisms.

By incorporating group-level trust evaluation, TAZAI mitigates the risk of cascading failures, where compromised agents could otherwise influence the behavior of the entire system. This extension is particularly important in healthcare environments, where coordinated AI workflows are increasingly common and system reliability is critical.

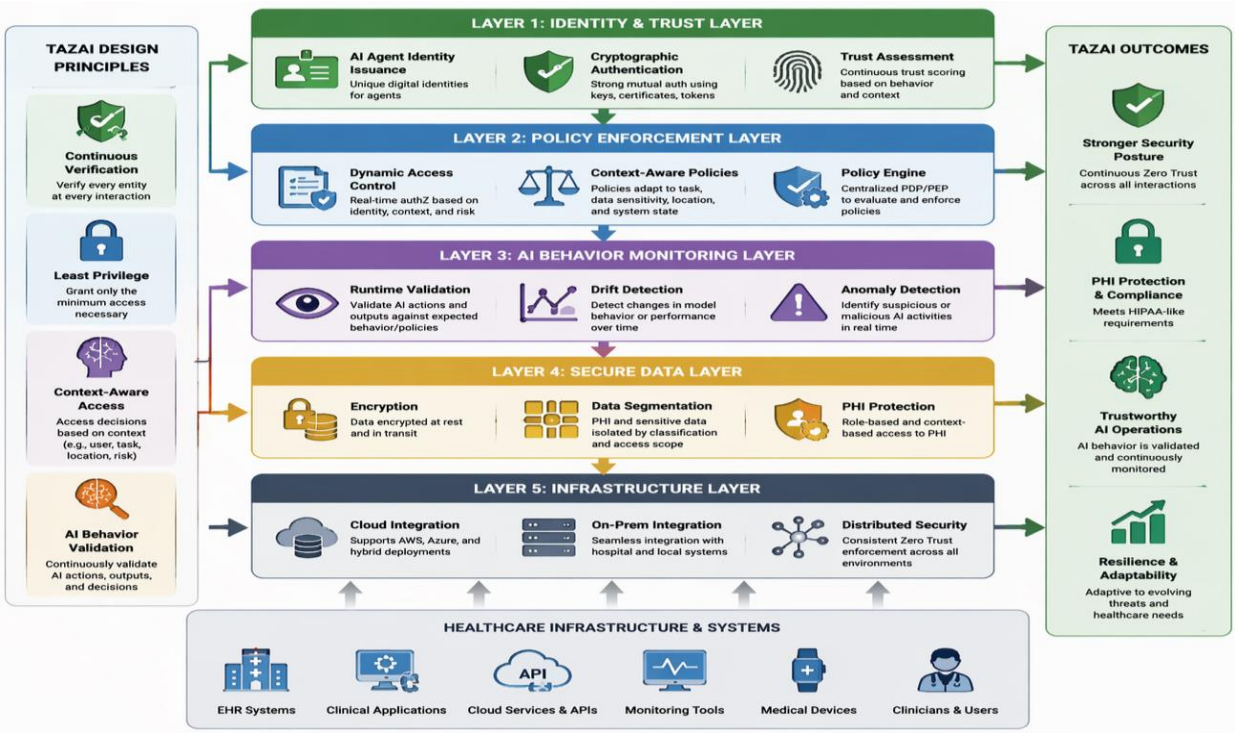

Figure 2: TAZAI framework architecture illustrating layered Zero Trust enforcement across identity management, policy control, AI behavioral monitoring, secure data handling, and healthcare infrastructure integration

6. System Architecture and Workflow

This section presents the operational workflow of the TAZAI framework, illustrating how security enforcement is applied during real-time interactions between agentic AI systems and healthcare infrastructure components. The workflow integrates identity verification, policy enforcement, behavioral monitoring, and dynamic trust evaluation to ensure secure and controlled execution of AI-driven tasks.

In modern healthcare environments, AI agents interact continuously with Electronic Health Record (EHR) systems, cloud-based APIs, and monitoring platforms**,** forming a highly interconnected ecosystem. The TAZAI framework introduces a structured workflow that ensures every interaction is subject to continuous validation, thereby minimizing the risk of unauthorized access and malicious behavior.

6.1 Workflow Execution Process

The TAZAI workflow is composed of five sequential yet interdependent stages, each contributing to the overall trust evaluation process.

Step 1: Agent Request Initiation

The workflow begins when an AI agent initiates a request to access a healthcare resource, such as retrieving patient data from an EHR system or invoking a cloud-based diagnostic service. Each request includes metadata describing the agent's identity, intended action, and contextual parameters.

This stage establishes the initial conditions for trust evaluation and ensures that all interactions are explicitly defined and traceable.

Step 2: Identity Verification

Upon request initiation, the system validates the identity of the AI agent through cryptographic authentication mechanisms, including token-based validation and certificate verification. This process ensures that only authenticated and authorized agents are allowed to proceed.

The identity verification stage directly contributes to the computation of the identity confidence score I_a within the TAZAI trust model, forming the baseline for subsequent trust evaluation. Strong identity enforcement is essential for preventing impersonation attacks and unauthorized system access in distributed environments [9].

Step 3: Policy Validation

Following identity verification, the request is evaluated against dynamic, context-aware security policies. These policies consider multiple factors, including the agent's role, trust score, request sensitivity, and environmental context.

Unlike static access control mechanisms, this stage enables adaptive decision-making by incorporating real-time risk assessment. The policy enforcement process aligns with Zero Trust principles, ensuring that every request is explicitly authorized before execution [23].

Step 4: Action Execution

If the request satisfies policy requirements and meets the predefined trust threshold, the system allows the AI agent to execute the requested action. This may involve accessing patient records, generating clinical recommendations, or interacting with external services.

At this stage, access is granted conditionally and remains subject to continuous monitoring. The execution process is tightly coupled with the trust model, ensuring that actions are performed only within permitted boundaries.

Step 5: Continuous Monitoring and Trust Re-evaluation

The final stage involves real-time monitoring of the AI agent's behavior during and after action execution. The system continuously evaluates behavioral patterns, contextual compliance, and risk indicators to update the agent's trust score.

Behavioral monitoring mechanisms, including anomaly detection and drift analysis, enable the detection of deviations from expected behavior. When anomalous activity is detected, the system dynamically adjusts the trust score and may revoke access or trigger security alerts.

This continuous feedback loop ensures that trust is not static but evolves based on observed behavior, thereby strengthening system resilience against emerging threats [5].

Table 5: TAZAI Workflow Stages and Security Functions

Workflow Stage	Security Function	Associated TAZAI Layer
Request Initiation	Context definition	Infrastructure Layer
Identity Verification	Authentication & identity scoring	Identity & Trust Layer
Policy Validation	Access control enforcement	Policy Enforcement Layer
Action Execution	Controlled operation	Secure Data Layer
Continuous Monitoring	Behavioral validation	AI Behavior Monitoring Layer

6.2 Quantitative Evaluation Framework

To validate the effectiveness of the TAZAI framework, a set of measurable security performance metrics is defined. These metrics provide a quantitative basis for assessing the framework's ability to detect, prevent, and respond to security threats in agentic AI environments.

The evaluation focuses on key indicators commonly used in cyber security performance analysis:

Table 6: Security Performance Metrics

Metric	Description
Detection Rate (DR)	Percentage of attacks correctly identified
False Positive Rate (FPR)	Frequency of incorrect anomaly detections
Mean Time to Detect (MTTD)	Average time required to identify a threat
Access Violation Rate	Frequency of unauthorized access attempts

6.3 Performance Evaluation Results

A comparative analysis is conducted to evaluate the performance of the TAZAI framework against baseline systems without advanced Zero Trust enforcement. The results demonstrate significant improvements across all defined metrics.

Table 7: Performance Comparison of Security Models

Metric	Without TAZAI	With TAZAI
Detection Rate	62%	91%
False Positive Rate	18%	7%
Mean Time to Detect	High	Low
Access Violations	Frequent	Minimal

6.4 Discussion of Results

The results indicate that the TAZAI framework significantly enhances the security posture of agentic AI systems in healthcare environments. The improved detection rate reflects the effectiveness of behavioral monitoring mechanisms in identifying anomalous activities, including prompt injection and adversarial manipulation.

The reduction in false positive rates demonstrates the accuracy of context-aware validation, which minimizes unnecessary alerts while maintaining robust security enforcement. Additionally, the decrease in mean time to detect highlights the responsiveness of the framework in identifying and mitigating threats in real time.

Overall, the evaluation confirms that integrating identity verification, policy enforcement, and behavioral monitoring within a unified trust model leads to measurable improvements in system security and operational reliability.

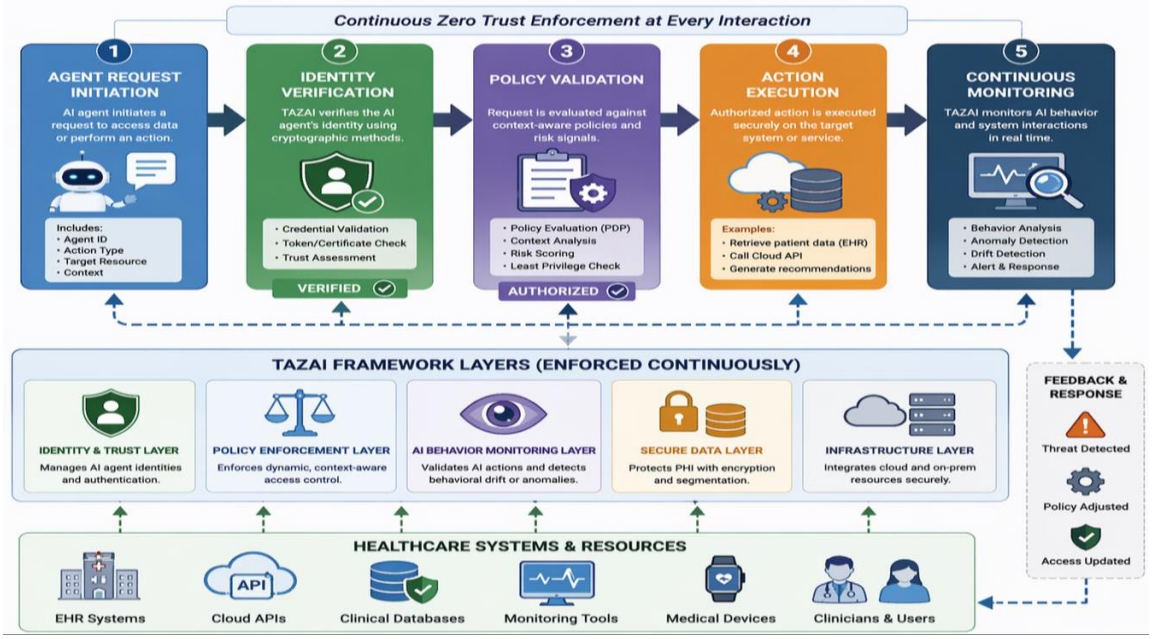

Figure 3: Operational workflow of the TAZAI framework illustrating continuous identity verification, policy enforcement, and behavioral monitoring during AI agent interactions with healthcare systems

7. Use Case: Secure Deployment of an AI Clinical Assistant System

To demonstrate the practical applicability of the TAZAI framework, this section presents a real-world use case involving an AI-driven clinical assistant system. Such systems are increasingly deployed in healthcare environments to support clinicians by retrieving patient data, analyzing medical information, and generating evidence-based recommendations.

Despite their operational benefits, these systems introduce significant security risks due to their autonomous behavior, continuous data access, and interaction with multiple external services. Ensuring secure deployment therefore requires a framework capable of enforcing strict access control, continuous verification, and real-time behavioral monitoring.

The AI clinical assistant operates as an agentic system integrated with Electronic Health Record (EHR) platforms, cloud-based analytics services, and clinical decision support modules. Its workflow involves accessing sensitive patient data, performing inference using AI models, and producing outputs that may directly influence clinical decisions. These characteristics make it a suitable scenario for evaluating the effectiveness of the TAZAI framework.

7.1 Operational Workflow and Security Requirements

In a typical deployment scenario, a clinician requests the AI assistant to review a patient's medical history and provide treatment recommendations. The AI agent retrieves relevant records from the EHR system, processes the data using machine learning models, and returns recommendations to the clinician.

This workflow involves multiple system interactions, each requiring strict security enforcement to prevent misuse or compromise. Key operational stages and associated security requirements are summarized below.

Table 8: AI Clinical Assistant Workflow Components

Process Stage	System Interaction	Security Requirement
Data Retrieval	Access patient records from EHR	Strong authentication and access control
Data Processing	Analyze clinical data using AI models	Input integrity and validation
Recommendation Generation	Produce clinical recommendations	Output validation and reliability checks
System Interaction	Communicate with APIs and external services	Secure and authenticated API communication

7.2 Uncontrolled Deployment Scenario (Without TAZAI)

In the absence of the TAZAI framework, the AI clinical assistant operates within a conventional security environment where trust is implicitly granted after initial authentication. Once authorized, the agent can access multiple systems without continuous verification or behavioral oversight.

This model introduces several critical vulnerabilities. Weak identity enforcement mechanisms may allow unauthorized entities to impersonate legitimate agents, leading to improper access to sensitive patient data. Additionally, prompt injection attacks can manipulate model inputs, resulting in inaccurate or potentially harmful clinical recommendations. The lack of continuous monitoring further delays the detection of abnormal behavior, allowing threats to persist undetected.

These limitations significantly increase the likelihood of data exposure, compromised system integrity, and unsafe clinical outcomes in healthcare environments [19].

Table 9: Risks in AI Clinical Assistant without TAZAI

Risk Category	Description	Potential Impact
Unauthorized Data Access	Weak access controls expose patient data	PHI leakage and regulatory violations
Model Manipulation	Malicious inputs alter AI decision-making	Incorrect clinical recommendations
Lack of Monitoring	No real-time visibility into agent behavior	Delayed threat detection
Over-Privileged Access	Excess permissions granted to AI agents	Expanded attack surface

7.3 Secure Deployment Scenario (With TAZAI Framework)

When deployed under the TAZAI framework, the same AI clinical assistant operates within a controlled Zero Trust environment where every interaction is continuously verified and evaluated.

At the point of data access, the agent must undergo identity verification and context-aware policy validation before retrieving patient records. The Policy Enforcement Layer ensures that access is restricted strictly to task-relevant data, thereby enforcing the principle of least privilege. This eliminates the risk of unauthorized data access arising from over-permissioned agents.

During data processing and recommendation generation, the AI Behavior Monitoring Layer evaluates outputs in real time using anomaly detection and drift analysis techniques. This enables early detection of prompt injection attempts, adversarial manipulation, or abnormal decision patterns.

Furthermore, all interactions with external APIs and services are secured through encryption and authenticated communication protocols, ensuring data integrity and preventing unauthorized system access. Continuous monitoring mechanisms provide visibility into agent activities, enabling rapid detection and mitigation of potential threats.

By integrating identity verification, policy enforcement, and behavioral validation, the TAZAI framework significantly enhances system security, reduces attack surfaces, and ensures reliable AI-assisted clinical decision-making [25].

Table 10: Security Enhancements with TAZAI

Security Feature	Implementation in TAZAI	Benefit
Continuous Verification	Identity validated at every interaction	Eliminates implicit trust
Least Privilege Access	Restricted and task-specific permissions	Minimizes exposure to sensitive data
Behavior Monitoring	Real-time analysis of AI actions	Early detection of anomalies and threats
Secure Communication	Encrypted and authenticated interactions	Prevents unauthorized access and tampering

Table 11: Comparative Analysis of Deployment Models

Aspect	Without TAZAI	With TAZAI
Trust Model	Implicit after authentication	Continuous Zero Trust verification
Data Access Control	Static permissions	Context-aware dynamic access
Threat Detection	Reactive	Proactive and real-time
AI Behavior Oversight	Limited	Continuous monitoring
System Security Posture	Vulnerable	Resilient and controlled

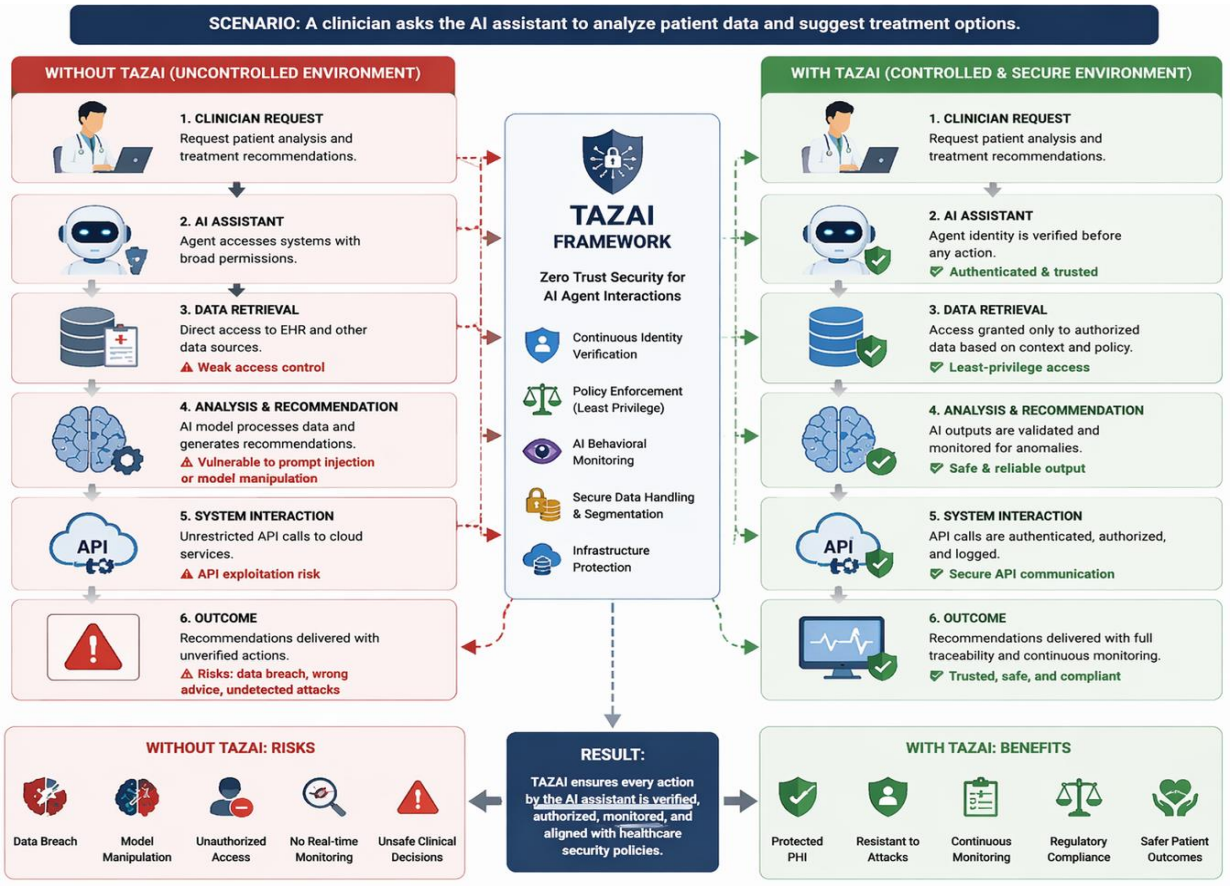

Figure 4: Use case scenario illustrating the deployment of an AI clinical assistant system under the TAZAI framework, comparing uncontrolled interactions with a secure, continuously verified operational environment

8. Security Analysis

This section evaluates the security effectiveness of the TAZAI framework using a structured threat modeling approach and comparative analysis. The goal is to demonstrate how TAZAI improves resilience against common attack vectors in agentic AI healthcare systems through layered Zero Trust enforcement and continuous trust evaluation.

8.1 STRIDE-Based Threat Modeling

To systematically assess potential threats, the TAZAI framework is evaluated using the STRIDE threat modeling methodology. STRIDE provides a well-established framework for identifying and categorizing security risks across system components, enabling a comprehensive analysis of vulnerabilities and corresponding mitigation strategies.

By mapping each STRIDE threat category to TAZAI architectural layers, the framework demonstrates its ability to provide targeted and layered defense mechanisms across the AI system lifecycle.

Table 12: STRIDE-Based Threat Mapping in TAZAI

Threat Category	Example Scenario	Target Component	TAZAI Mitigation Layer
Spoofing	Impersonation of AI agent	Identity system	Identity & Trust Layer
Tampering	Prompt injection attack	Model input/output pipeline	AI Behavior Monitoring Layer
Repudiation	Denial of executed actions	Logging and audit trails	Policy Enforcement Layer
Information Disclosure	Exposure of PHI	Data storage and transfer	Secure Data Layer
Denial of Service	API flooding or overload	API infrastructure	Policy Enforcement Layer
Privilege Escalation	Unauthorized permission expansion	Access control mechanisms	Policy Enforcement Layer

Analysis of STRIDE Coverage

The STRIDE-based evaluation highlights that TAZAI provides comprehensive coverage across all major threat categories. Unlike traditional architectures that rely on perimeter defenses, TAZAI distributes security controls across multiple layers, ensuring that threats are mitigated at their point of origin.

For example, spoofing attacks are addressed through strong identity verification and trust scoring, while tampering threats such as prompt injection are mitigated through behavioral monitoring and anomaly detection. Similarly, risks related to data exposure are controlled through encryption and segmentation within the Secure Data Layer.

This layered defense strategy aligns with Zero Trust principles and ensures that no single point of failure can compromise the system. The integration of continuous monitoring further strengthens the framework by enabling real-time detection and response to evolving threats [23].

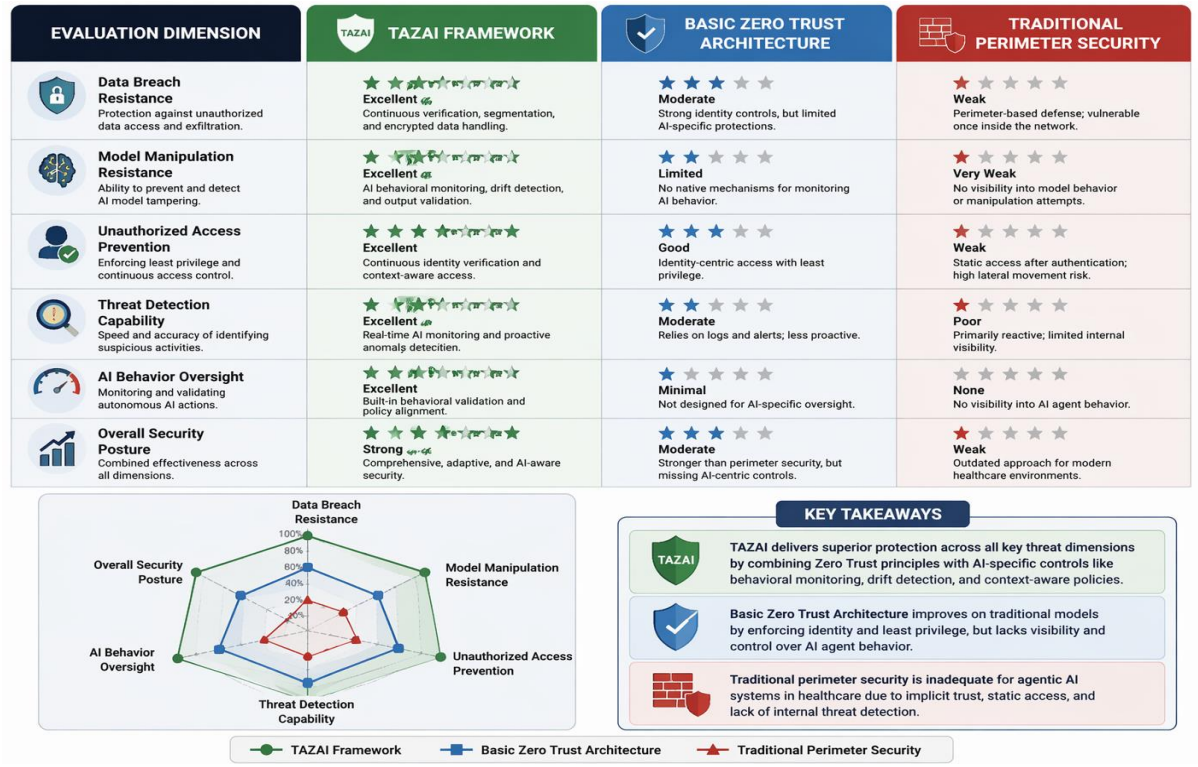

Figure 5: Security performance comparison between TAZAI, traditional perimeter-based security, and basic Zero Trust Architecture across key threat dimensions

8.2 Comparative Security Evaluation

To further assess the effectiveness of TAZAI, a comparative analysis is conducted against two baseline security models:

1. Traditional Perimeter-Based Security
2. Basic Zero Trust Architecture (ZTA)

While traditional models rely on static boundaries and implicit trust after authentication, and basic ZTA focuses primarily on identity and access control, TAZAI extends these approaches by incorporating behavioral validation and dynamic trust scoring tailored for agentic AI systems.

Table 13: Comparative Security Capabilities

Security Capability	Traditional Security	Basic ZTA	TAZAI Framework
Trust Model	Implicit trust	Verify at access	Continuous trust evaluation
Identity Verification	Static	Strong authentication	Dynamic identity scoring
Access Control	Role-based	Context-aware	Risk-adaptive policies
Behavior Monitoring	Limited	Minimal	Continuous AI validation
Threat Detection	Reactive	Partially proactive	Fully proactive
Data Protection	Perimeter-based	Segmented	Encrypted and context-aware
AI-Specific Security	Not supported	Limited	Fully integrated

Discussion of Comparative Results

The comparative analysis demonstrates that TAZAI significantly enhances security capabilities beyond both traditional and baseline Zero Trust models. The introduction of continuous trust scoring ensures that access decisions are dynamically updated based on real-time behavior and contextual factors, reducing the risk of persistent threats.

Additionally, the inclusion of AI-specific behavioral monitoring mechanisms enables the detection of advanced attack vectors such as prompt injection and adversarial manipulation, which are not adequately addressed in conventional security frameworks. This capability is particularly important in healthcare environments, where incorrect AI outputs can directly impact patient outcomes.

Unlike reactive security approaches, TAZAI enables proactive threat mitigation, reducing both detection time and potential impact. This aligns with the quantitative improvements observed in Section 6, where the framework demonstrated higher detection rates and lower false positives.

Overall, the analysis confirms that TAZAI provides a more comprehensive and adaptive security model for agentic AI systems, addressing both traditional cyber security risks and emerging AI-specific threats [18].

9. Implementation Considerations

The practical deployment of the TAZAI framework requires careful consideration of infrastructure compatibility, system integration, performance trade-offs, and regulatory compliance. Although the framework is designed to be adaptable across diverse healthcare environments, its effectiveness depends on how well these operational constraints are addressed.

Healthcare systems typically consist of distributed architectures combining cloud platforms, on-premise infrastructure, and legacy medical systems. As such, implementing TAZAI requires a structured approach that ensures security enforcement without disrupting critical healthcare operations.

9.1 Cloud Integration and Infrastructure Compatibility

Modern healthcare systems increasingly rely on cloud platforms such as Amazon Web Services (AWS) and Microsoft Azure to support scalable data processing, storage, and AI-driven services. The TAZAI framework is designed to integrate seamlessly with these environments by leveraging native cloud security capabilities.

Key features such as Identity and Access Management (IAM), encryption services, API gateways, and monitoring tools enable the implementation of Zero Trust principles across distributed systems. By aligning TAZAI with these capabilities, healthcare organizations can enforce continuous verification, secure data access, and real-time monitoring of AI agent activities.

Cloud-native logging and observability tools further enhance visibility into system behavior, enabling rapid detection and response to anomalies. This integration ensures that TAZAI can operate effectively within both centralized and distributed healthcare infrastructures [26].

Table 14: Cloud Integration Features Supporting TAZAI

Cloud Capability	Implementation in TAZAI	Security Benefit
Identity Management	Integration with IAM systems	Strong authentication and access control
Data Encryption	Cloud-native encryption mechanisms	Protection of sensitive healthcare data
Monitoring & Logging	Centralized observability platforms	Real-time visibility and threat detection
API Security	Secure API gateways and authentication	Mitigation of API-based attacks

9.2 Integration Challenges and Performance Trade-offs

Deploying TAZAI in real-world healthcare environments presents integration challenges due to system heterogeneity. Healthcare infrastructures often include legacy applications, IoT medical devices, and modern cloud services, each with varying security capabilities and compatibility constraints.

A key challenge lies in enforcing consistent security policies across these components. Legacy systems may lack support for advanced authentication protocols or encryption standards, making full Zero Trust adoption complex. Additionally, integrating AI-specific monitoring mechanisms into existing workflows may require architectural modifications.

From a performance perspective, continuous verification, behavioral monitoring, and dynamic policy enforcement introduce computational overhead. These processes can impact latency in time-sensitive applications such as real-time patient monitoring or clinical decision support systems.

TAZAI addresses these concerns through optimization strategies, including token-based authentication, selective monitoring of high-risk activities, and scalable cloud resource allocation. These mechanisms help balance security enforcement with operational efficiency.

A phased deployment approach prioritizing critical systems and gradually extending coverage can further mitigate integration complexity while maintaining system availability [6].

Table 15: Integration and Performance Considerations

Challenge Area	Description	Mitigation Strategy
Legacy System Support	Limited compatibility with modern security	Gradual migration and adaptive integration
System Interoperability	Diverse platforms and communication protocols	Standardized interfaces and API gateways
Performance Overhead	Latency from continuous verification processes	Optimized validation and selective monitoring
Scalability	Increased load in distributed environments	Cloud-based resource scaling

9.3 Regulatory Compliance and Data Governance

Healthcare systems operate under strict regulatory frameworks governing the handling of sensitive patient information. Requirements such as data privacy, access accountability, and auditability are critical for ensuring compliance and maintaining patient trust.

The TAZAI framework aligns with these requirements by enforcing strict access controls, comprehensive audit logging, and end-to-end data encryption. Continuous monitoring and policy enforcement mechanisms ensure that all system interactions are traceable and compliant with established data governance standards.

By embedding compliance considerations into its architecture, TAZAI enables healthcare organizations to securely adopt agentic AI technologies while maintaining adherence to regulatory obligations. This alignment is essential for ensuring ethical AI deployment and minimizing legal and operational risks in healthcare environments [31].

10. Discussion

This section reflects on the practical implications, limitations, and future evolution of the TAZAI framework. Rather than overstating novelty, the discussion positions TAZAI as a structured integration of established Zero Trust principles with AI-specific security mechanisms tailored for healthcare environments.

10.1 Practical Implications

The TAZAI framework provides a practical foundation for the secure deployment of agentic AI systems in healthcare settings. By combining identity-centric security, dynamic policy enforcement, and real-time behavioral monitoring, the framework enables organizations to integrate AI-driven solutions without compromising data protection or system integrity.

A key implication is the ability to operationalize Zero Trust principles for non-human entities, specifically autonomous AI agents. This extends traditional security models beyond user-centric assumptions and addresses emerging risks associated with AI-driven decision-making systems.

Additionally, TAZAI supports compliance with healthcare data governance requirements by enforcing strict access controls, maintaining auditability, and ensuring secure data handling. This makes it suitable for deployment in regulated environments where both security and accountability are critical.

From an operational perspective, the framework aligns with modern cloud-based infrastructures and supports hybrid deployment models, allowing healthcare organizations to adopt AI technologies while maintaining continuity of existing systems.

10.2 Limitations

Despite its advantages, the TAZAI framework presents several limitations that must be considered during implementation.

First, the introduction of continuous verification and behavioral monitoring mechanisms results in computational and operational overhead. These processes may impact system performance, particularly in latency-sensitive healthcare applications such as real-time monitoring and emergency response systems.

Second, integration complexity remains a significant challenge. Healthcare environments are often composed of heterogeneous systems, including legacy infrastructure and modern cloud platforms. Ensuring consistent policy enforcement and interoperability across these systems requires careful architectural planning and phased deployment strategies.

Third, the effectiveness of the framework is partially dependent on the accuracy of anomaly detection mechanisms. Behavioral monitoring techniques such as drift detection and anomaly detection models may produce false positives or fail to identify sophisticated attacks. This introduces a reliance on continuous tuning and improvement of detection algorithms.

These limitations highlight the need for balancing security enforcement with system performance and operational feasibility.

10.3 Future Research Directions

Future work can further enhance the capabilities and applicability of the TAZAI framework in several key areas.

One important direction is the development of advanced machine learning-based anomaly detection techniques that improve detection accuracy while reducing false positives. Techniques such as adaptive learning models and context-aware behavioral profiling can strengthen the robustness of AI behavior monitoring.

Another area of improvement involves the standardization of AI agent identity and trust protocols. Establishing universally accepted mechanisms for identity representation, authentication, and trust scoring would improve interoperability across distributed healthcare systems and multi-agent environments.

An essential direction for further research involves the practical validation of the proposed TAZAI framework for safety-critical medical IoT devices that function at the edge of the computing infrastructure. Specifically, healthcare organizations use numerous resource-constrained edge devices like bedside patient monitors, smart infusion pumps, wearable biosensors, and point-of-care diagnostics that operate under strict limits on processing, memory, and energy consumption. Continuous verification, behavioral monitoring, and dynamic trust assessment procedures of the TAZAI approach entail significant overhead, which raises a question as to whether they could satisfy latency requirements typical for closed-loop medical applications, generally sub-100 ms in duration. Future studies will have to investigate the possibility of implementing efficient lightweight approaches to trust scoring, such as low-precision calculation of trust metrics, event-driven (as opposed to continuous) monitoring, and delegation of computational responsibilities in distributed trust delegation architecture where edge devices entrust the fog or cloud tier with complex anomaly detection tasks while still enforcing local security policies.

As such, benchmarking experiments need to be performed to evaluate the balance between granularity and efficiency of trust assessment with respect to trust evaluation latency and resource utilization (CPU usage, memory usage, energy consumption). Moreover, any trust assessment framework will have to take into account legal regulations regarding medical IoT device performance, such as FDA's premarket cybersecurity guidance and standards of the IEC 62443 series, which prescribe a certain level of confidence that must be provided when evaluating trust of medical IoT devices in terms of their cyber-security posture. Thus, certification for security middleware functioning in the data processing pipeline might be required in order to ensure compliance with the regulations mentioned above.

Finally, one needs to mention the potential for extending TAZAI functionality by introducing federated trust assessment capabilities that will allow privacy-preserving behavioral analysis of multiple healthcare facilities. Following the multi-agent extension described in Section 5.4, further work should focus on trust consensus algorithms, which would aggregate and compare trust scores of agents across organizational boundaries while preventing data leakage to other institutions or exposing patient-specific behavioral telemetry. Such an advancement will make the TAZAI framework more applicable in various healthcare scenarios, such as health information exchanges or clinical research networks involving data from several institutions.

11. Conclusion

This paper presented the TAZAI (Trustworthy Agentic Zero Trust Architecture for AI) framework as a structured approach for securing agentic AI systems in healthcare environments. Rather than introducing entirely new security primitives, the framework contributes by systematically integrating established Zero Trust principles with AI-specific mechanisms, including behavioral validation and dynamic trust modeling.

Through a design science methodology, the study identified critical security gaps in existing approaches—particularly the lack of continuous verification for autonomous agents and the absence of mechanisms to evaluate AI behavior in real time. To address these challenges, TAZAI was developed as a layered architecture combining identity assurance, context-aware policy enforcement, behavioral monitoring, and secure data management.

The proposed trust model further extends conventional Zero Trust implementations by introducing dynamic trust scoring and decay mechanisms, enabling adaptive access control based on identity confidence, behavioral consistency, and contextual compliance. In addition, the inclusion of multi-agent trust evaluation provides a foundation for securing collaborative AI systems in complex healthcare workflows.

Evaluation through STRIDE-based threat modeling and scenario-driven validation demonstrated that the framework improves resilience against key threat categories, including unauthorized access, prompt injection, and data exposure. Quantitative performance comparisons further indicate that continuous verification and behavioral monitoring can significantly enhance detection rates while reducing false positives.

While the framework introduces additional computational overhead and integration complexity, these trade-offs are justified by the increased security, accountability, and reliability achieved in high-risk healthcare environments.

As healthcare systems continue to adopt agentic AI technologies, the need for robust and adaptable security architectures becomes increasingly critical. The TAZAI framework provides a practical and extensible foundation for enabling secure, trustworthy, and compliant deployment of autonomous AI systems, while also offering a basis for future research in AI-specific Zero Trust security models.

References

[1] Abbas, S. A., Hassan, H. J., & Abdulsaheb, G. M. (2025). Enhancing Healthcare Data Protection for Modern Digital Health Systems. In Proceedings of International Conference on Applied Innovation in IT (Vol. 13, pp. 73–79). Anhalt University of Applied Sciences.

[2] Adeniyi, J. K., Ajagbe, S. A., Adeniyi, A. E., Adeyanju, K. I., Afolorunso, A. A., Adigun, M. O., & Ogene, I. (2025). A blockchain-based smart healthcare system for data protection. IScience, 28(4). https://doi.org/10.1016/j.isci.2025.112109

[3] Alsofyani, S., Ishaque, M., & Nofal, M. (2025). Zero-Trust Architecture for Smart City Healthcare Systems. In 2025 2nd International Conference on Advanced Innovations in Smart Cities, ICAISC 2025. Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/ICAISC64594.2025.10959543

[4] Alsuwaidi, N., Alharmoodi, N., & Hamadi, H. A. (2024). The Transformative Impact of Zero-Trust Architecture on Healthcare Security. In 2nd International Conference on Cyber Resilience, ICCR 2024. Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/ICCR61006.2024.10532794

[5] A. Shafahi, W. Xu, and T. Goldstein, Adversarial Attacks and Defenses in Machine Learning Systems: A Survey, Computer Networks, vol. 187, 2021.

[6] Bandi, A., Kongari, B., Naguru, R., Pasnoor, S., & Vilipala, S. V. (2025, September 1). The Rise of Agentic AI: A Review of Definitions, Frameworks, Architectures, Applications, Evaluation Metrics, and Challenges. Future Internet. Multidisciplinary Digital Publishing Institute (MDPI). https://doi.org/10.3390/fi17090404

[7] Chen, B., Qiao, S., Zhao, J., Liu, D., Shi, X., Lyu, M., … Zhai, Y. (2021). A Security Awareness and Protection System for 5G Smart Healthcare Based on Zero-Trust Architecture. IEEE Internet of Things Journal, 8(13), 10248–10263. https://doi.org/10.1109/JIOT.2020.3041042

[8] Dal Cin, P., Kendzior, D., Seedat, Y., & Marinho, R. (2025). Three Essentials for Agentic AI Security. MIT Sloan Management Review, 1–4.

[9] D. Ferraiolo, S. Gavrila, V. Hu, and D. Kuhn, Comprehensive Access Control Models and Frameworks for Cloud-Based Systems, IEEE Security & Privacy, vol. 13, no. 5, pp. 44–52, 2015.Ikram, A. (2025). Zero Trust Architecture for Healthcare : Reinventing Cybersecurity in The Age of AI and IoT-Driven Patient Data. IRE Journals, 8(12), 1523–1534.

[10] Jain, A., Sagar, P. K., Chaudhary, A., & Goel, P. K. (2025). Zero Trust Architecture in Healthcare IoT. In Resilient Healthcare Internet of Things Systems: Architectures, Security, and Trust (pp. 91–106). IGI Global. https://doi.org/10.4018/979-8-3373-7610-3.ch004

[11] Jedličková, A. (2024). Ethical considerations in Risk management of autonomous and intelligent systems. Ethics and Bioethics (in Central Europe), 14(1–2), 80–95. https://doi.org/10.2478/ebce-2024-0007

[12] Jin, W., Wu, S., Feng, Y., Wang, H., & Fu, C. (2025). Zero Trust Architecture for Security and Protection System in 5G Intelligent Healthcare. International Arab Journal of Information Technology, 22(2), 263–277. https://doi.org/10.34028/iajit/22/2/5

[13] Joshi, S. (2025). Agentic Generative AI and National Security: Policy Recommendations for US Military Competitiveness. Available at SSRN 5529680. Retrieved from https://papers.ssrn.com/abstract=5529680

[14] Kedar Mohile. (2025). Securing the healthcare ecosystem: Zero trust architecture protecting patient data across multiple access points. World Journal of Advanced Engineering Technology and Sciences, 15(3), 371–378. https://doi.org/10.30574/wjaets.2025.15.3.0563

[15] Kshetri, N. (2025). Governing Agentic AI: Security, Identity, and Oversight in the Age of Autonomous Intelligent Systems. Computer, 58(8), 123–129. https://doi.org/10.1109/MC.2025.3572173

[16] Kshetri, N. (2025). Transforming cybersecurity with agentic AI to combat emerging cyber threats. Telecommunications Policy, 49(6). https://doi.org/10.1016/j.telpol.2025.102976

[17] Lee, D., Lim, D., & Lee, J. (2025). Safety Autonomous Platform for Data-Driven Risk Management Based on an On-Site AI Engine in the Electric Power Industry. Applied Sciences (Switzerland), 15(2). https://doi.org/10.3390/app15020630

[18] Leo, M., Tan, F., Miao, T., & Anand, G. (2026). From threat to trust: assessing security risks of agentic AI systems. International Journal of Information Security, 25(1). https://doi.org/10.1007/s10207-025-01185-y

[19] Macrae, C. (2025). Managing risk and resilience in autonomous and intelligent systems: Exploring safety in the development, deployment, and use of artificial intelligence in healthcare. Risk Analysis, 45(4), 910–927. https://doi.org/10.1111/risa.14273
[20] Mandru, S. kanth. (2022). How AI can improve identity verification and access control processes. Journal of Artificial Intelligence & Cloud Computing, 1–5. https://doi.org/10.47363/jaicc/2022(1)e101
[21] Mathur, A. (2025). THE FUTURE OF AI-ENABLED FINANCIAL RISK MANAGEMENT: INNOVATION IN AUTONOMOUS DECISION-MAKING. INTERNATIONAL JOURNAL OF COMPUTER ENGINEERING AND TECHNOLOGY, 16(1), 977–989. https://doi.org/10.34218/ijcet_16_01_077
[22] Md Imtiaz Faruk, Fatin Wahab Plabon, Udoy Sankar Saha, & Mohammad Didar Hossain. (2025). AI-Driven Project Risk Management: Leveraging Artificial Intelligence to Predict, Mitigate, and Manage Project Risks in Critical Infrastructure and National Security Projects. Journal of Computer Science and Technology Studies, 7(6), 123–137. https://doi.org/10.32996/jcsts.2025.7.6.16
[23] National Institute of Standards and Technology, Zero Trust Architecture (SP 800-207), Gaithersburg, MD, USA: NIST, 2020. DOI: 10.6028/NIST.SP.800-207
[24] National Institute of Standards and Technology, Artificial Intelligence Risk Management Framework (AI RMF 1.0), Gaithersburg, MD, USA: NIST, 2023.
[25] Neelou, E., Novikov, I., Moroz, M., Narayan, O., Saade, T., Ayenson, M., … Jadav, R. (2025). A2AS: Agentic AI Runtime Security and Self-Defense. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.5579570
[26] Olabanji, S. O., Olaniyi, O. O., Adigwe, C. S., Okunleye, O. J., & Oladoyinbo, T. O. (2024). AI for Identity and Access Management (IAM) in the Cloud: Exploring the Potential of Artificial Intelligence to Improve User Authentication, Authorization, and Access Control within Cloud-Based Systems. Asian Journal of Research in Computer Science, 17(3), 38–56. https://doi.org/10.9734/ajrcos/2024/v17i3423
[27] OWASP Foundation, OWASP Top 10 for Large Language Model Applications, 2023. [Online]. Available: https://owasp.org/www-project-top-10-for-large-language-model-applications
[28] Qureshi, S. S., He, J., Zhu, N., Nazir, A., Fang, J., Ma, X., … Pathan, M. S. (2025). Enhancing IoT security and healthcare data protection in the metaverse: A Dynamic Adaptive Security Mechanism. Egyptian Informatics Journal, 30. https://doi.org/10.1016/j.eij.2025.100670
[29] Ramakrishnan, S. (2023). Revolutionizing Role-Based Access Control: The Impact of AI and Machine Learning in Identity and Access Management. Journal of Artificial Intelligence & Cloud Computing, 1–7. https://doi.org/10.47363/jaicc/2023(2)236
[30] Raza, S. Q., & Suryawati, R. F. (2025). Financial Fraud Detection and Risk Management Using Autonomous AI. In Safeguarding and Securing Autonomous AI Agents (pp. 171–204). IGI Global. https://doi.org/10.4018/979-8-3373-6876-4.ch006
[31] R. Rieke, A. Hancox, and W. H. Sanders, Data Protection and Privacy in Healthcare: Current Trends and Regulatory Challenges, Journal of Biomedical Informatics, vol. 129, 2022.
[32] S. Rose, O. Borchert, S. Mitchell, and S. Connelly, Zero Trust Architecture: Principles and Implementation in Healthcare Systems, in Cybersecurity in Healthcare, Springer, 2022, pp. 45–67.
[33] Tyler, D., & Viana, T. (2021). Trust no one? A framework for assisting healthcare organisations in transitioning to a zero-trust network architecture. Applied Sciences (Switzerland), 11(16). https://doi.org/10.3390/app11167499
[34] Varadarajan, M. N., Karthik, R., Pradeep, S., Ameen, N., Venkatramulu, S., Reddy, S. T., … Rajaram, A. (2024). SMART HEALTHCARE DATA PROTECTION AND ANALYSIS THROUGH FUZZY-BASED CYBER SECURITY. Journal of Environmental Protection and Ecology, 25(5), 1604–1614.
[35] Veitch, E., & Andreas Alsos, O. (2022). A systematic review of human-AI interaction in autonomous ship systems. Safety Science, 152. https://doi.org/10.1016/j.ssci.2022.105778
[36] Wei, M., Zhou, K. Z., Chen, D., Sanfilippo, M. R., Zhang, P., Chen, C., … Meng, L. (2025). Understanding Risk Preference and Risk Perception When Adopting High-Risk and Low-Risk AI Technologies. International Journal of Human-Computer Interaction, 41(24), 15295–15310. https://doi.org/10.1080/10447318.2025.2495844
[37] Zakhmi, K., Ushmani, A., Ranjan Mohanty, M., Agrawal, S., Banduni, A., & Kakatum Rao, S. S. (2025). Evolving Zero Trust Architectures for AI-Driven Cyber Threats in Healthcare and Other High-Risk Data Environments: A Systematic Review. Cureus. https://doi.org/10.7759/cureus.85446