Scientific Research Publishing

Journal of Computer and Communications, 2026, 14(2), 83-105
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

Ensuring Ethical Governance in Autonomous Agentic Systems: A Zero Trust Approach to Safeguarding Future Technologies

Eyong Atem

Business and Information Studies, Capitol Technology University, Laurel, USA
Email: eatem@captechu.edu

How to cite this paper: Atem, E. (2026) Ensuring Ethical Governance in Autonomous Agentic Systems: A Zero Trust Approach to Safeguarding Future Technologies. *Journal of Computer and Communications*, **14**, 83-105.
https://doi.org/10.4236/jcc.2026.142005

Received: January 26, 2026
Accepted: February 11, 2026
Published: February 14, 2026

Abstract

The rapid advancement of autonomous agentic systems (AAS) has revolutionized critical sectors such as healthcare, finance, defense, and cybersecurity. However, their integration into these domains raises significant ethical and security concerns, mainly when deployed within Zero Trust (ZT) frameworks, which assume no entity should be trusted by default. This paper explores the ethical governance of AAS within ZT frameworks, focusing on accountability, bias mitigation, and privacy protection. We examine the challenges posed by AAS, including decision-making accountability, inherited biases, and data privacy risks, and propose mitigation strategies such as explainable AI (XAI), human oversight, and privacy-preserving technologies. Additionally, we discuss the implementation of ZT principles, including continuous authentication, least privilege access, and network segmentation, and highlight the complexities of integrating AAS into these frameworks. The paper concludes with policy recommendations and future research directions, emphasizing the need for regulatory standards, ethical AI governance, and interdisciplinary collaboration to ensure the responsible deployment of AAS in ZT environments.

Keywords

Autonomous Agentic Systems, Zero Trust Framework, Ethical Governance, Explainable AI, Privacy-Preserving Technologies, Accountability Mechanisms

1. Introduction

The rapid advancement of autonomous agentic systems (AAS) has ushered in a new era of technological innovation, particularly in critical infrastructure and cybersecurity. These systems, which operate with varying degrees of autonomy, are

increasingly deployed to manage complex tasks, optimize operations, and enhance security protocols. However, their integration into critical sectors such as energy, healthcare, transportation, and defense has raised profound ethical and security concerns. As organizations increasingly adopt Zero Trust architectures—a security model that assumes no entity, internal or external, should be trusted by default—the need to integrate robust ethical governance frameworks for AAS becomes imperative. This article delves into the ethical considerations of deploying AAS within Zero Trust frameworks, focusing on accountability, bias mitigation, and privacy protection.

Autonomous agentic systems are characterized by their ability to perform tasks with minimal human intervention, leveraging artificial intelligence (AI), machine learning (ML), and other advanced technologies. In critical infrastructure, AAS are being used to monitor and control power grids, manage water supply systems, and optimize transportation networks. For instance, smart grids employ AAS to balance energy supply and demand in real-time, while autonomous drones are used to inspect pipelines and other infrastructure components. However, the increasing reliance on these systems has exposed vulnerabilities that could be exploited by malicious actors, leading to catastrophic consequences.

2. Understanding Autonomous Agentic Systems

Autonomous agentic systems (AAS) represent a significant leap in the evolution of artificial intelligence (AI) and machine learning (ML) technologies. These systems are designed to operate with a high degree of autonomy, enabling them to perform complex tasks, make decisions, and adapt to changing environments with minimal human intervention. Their applications span a wide range of sectors, including healthcare, cybersecurity, finance, and defense, where they are increasingly being deployed to enhance efficiency, accuracy, and security.

2.1. Definition and Characteristics

Autonomous agentic systems are AI-driven entities that exhibit agency, meaning they can perceive their environment, make decisions, and take actions to achieve specific goals without continuous human oversight. These systems are characterized by their ability to learn from data, adapt to new situations, and operate in dynamic and often unpredictable environments. Key characteristics of AAS include:

1) Autonomy: AAS can perform tasks independently, relying on pre-programmed rules, machine learning models, or reinforcement learning algorithms to guide their behavior. This autonomy allows them to operate in environments where human intervention is impractical or impossible [1].

2) Adaptability: These systems are designed to adapt to changing conditions, learning from new data and experiences to improve performance over time. This adaptability is particularly valuable in sectors such as cybersecurity, where threats evolve rapidly [2].

3) Decision-Making Capability: AAS are equipped with decision-making algorithms that enable them to evaluate multiple options and choose the most appropriate course of action based on predefined objectives. This capability is critical in applications such as healthcare diagnostics and financial trading [3].

4) Interactivity: AAS often interacts with other systems, humans, or their environment. For example, autonomous drones in defense operations interact with ground control systems and other drones to coordinate missions [4].

5) Scalability: These systems can be scaled to handle large volumes of data and complex tasks, making them suitable for deployment in critical infrastructure and large-scale industrial operations [5].

2.2. Applications in Various Sectors

The versatility of autonomous agentic systems has led to their widespread adoption across multiple sectors. Below, we explore their applications in healthcare, cybersecurity, finance, and defense, highlighting their transformative impact and the challenges associated with their deployment.

2.2.1. Healthcare

In healthcare, AAS is revolutionizing diagnostics, treatment, and patient care. These systems leverage AI and ML to analyze medical data, assist in decision-making, and perform complex procedures with precision. Key applications include:

- AI-Driven Diagnostics: AAS are being used to analyze medical images, such as X-rays, MRIs, and CT scans, to detect diseases like cancer, cardiovascular conditions, and neurological disorders. For example, AI algorithms have demonstrated the ability to identify breast cancer in mammograms with accuracy comparable to or exceeding that of human radiologists [6].
- Robotic Surgery: Autonomous surgical robots, such as the da Vinci Surgical System, assist surgeons in performing minimally invasive procedures with enhanced precision and control. These systems reduce the risk of human error and improve patient outcomes [7].
- Personalized Medicine: AAS analyzes patient data, including genetic information and medical history, to recommend customized treatment plans. This approach has shown promise in managing chronic diseases such as diabetes and cancer [8].

2.2.2. Cybersecurity

In cybersecurity, AAS detect, prevent, and respond to cyber threats in real time. Their ability to process vast amounts of data and identify patterns makes them invaluable in safeguarding digital infrastructure. Key applications include:

- Automated Threat Detection: AAS uses machine learning algorithms to monitor network traffic, identify anomalies, and detect potential cyberattacks. For instance, intrusion detection systems (IDS) powered by AI can identify and mitigate threats such as malware, phishing, and distributed denial-of-service

(DDoS) attacks [9].

- Incident Response: Autonomous systems can automatically respond to cyber incidents by isolating affected systems, blocking malicious traffic, and initiating recovery protocols. This reduces response times and minimizes the impact of cyberattacks [10]. AI improves threat detection by using machine learning to analyze data in real-time, identifying patterns and anomalies to catch new threats early. AI also automates incident response, learning from past breaches and adapting to emerging threat.
- Vulnerability Management: AAS continuously scans systems for vulnerabilities and applies patches or updates to mitigate risks. This proactive approach enhances the resilience of cybersecurity frameworks [11].

2.2.3. Finance

The financial sector has embraced AAS to improve efficiency, reduce risks, and enhance decision-making. These systems are particularly effective in fraud detection, trading, and risk management. Key applications include:

- Fraud Detection: AAS analyzes transaction data to identify fraudulent activities, such as credit card fraud and money laundering. Machine learning models can detect unusual patterns and flag suspicious transactions in real time [12].
- Algorithmic Trading: Autonomous trading systems use AI algorithms to analyze market data, predict trends, and execute trades at optimal times. These systems operate at high speeds and volumes, enabling them to capitalize on market opportunities [13]. AI enhances cybersecurity by swiftly analyzing vast amounts of data to detect anomalies and potential threats.
- Risk Management: AAS assesses financial risks by analyzing market conditions, credit scores, and economic indicators. They provide insights that help organizations make informed decisions and mitigate potential losses [14].

2.2.4. Defense

In defense, AAS are transforming military operations by enhancing situational awareness, precision, and operational efficiency. These systems are deployed in various forms, including autonomous drones, robotic combat units, and surveillance systems. Key applications include:

- Autonomous Drones: Unmanned Aerial Vehicles (UAVs) equipped with AI capabilities are used for reconnaissance, surveillance, and targeted strikes. These drones can operate in hostile environments, reducing the risk to human personnel [4].
- Robotic Combat Units: Autonomous ground robots are deployed in combat zones for bomb disposal, reconnaissance, and logistics support. These systems enhance the safety and effectiveness of military operations [15].
- Surveillance and Intelligence: AAS analyzes data from satellites, drones, and other sources to provide real-time intelligence and situational awareness. This information is critical for decision-making in military operations [16] (**Figure 1**).

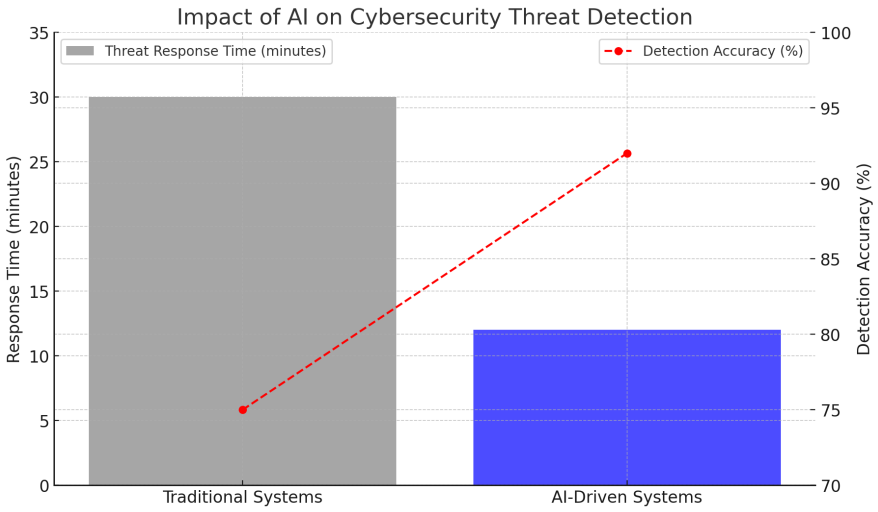

Figure 1. AI-driven threat detection improvements in cybersecurity.

3. Ethical Concerns in Autonomous Systems

The deployment of autonomous agentic systems (AAS) across critical sectors has brought about significant ethical challenges that must be addressed to ensure responsible use. While offering numerous benefits, these systems also raise concerns related to decision-making accountability, bias and discrimination, and privacy and surveillance. This section details these ethical concerns, highlighting their implications and the need for robust governance frameworks.

3.1. Decision-Making and Accountability

One of the most pressing ethical concerns surrounding AAS is the issue of accountability in decision-making. Autonomous systems are designed to operate with minimal human intervention, making decisions based on pre-programmed rules, machine learning models, or real-time data analysis. However, determining responsibility becomes a complex legal and ethical challenge when these systems make erroneous or harmful decisions.

- The Problem of Attribution: In traditional systems, human operators are accountable for decisions and actions. However, in AAS, the decision-making process is often opaque, making it difficult to attribute responsibility. For example, if an autonomous vehicle causes an accident, is the manufacturer, the software developer, or the user at fault? This lack of clarity complicates legal proceedings and undermines trust in autonomous systems [17].
- Moral Responsibility: AAS operate in environments where their decisions can have significant ethical implications, such as healthcare or defense. For instance, an autonomous medical diagnosis system that makes an incorrect recommendation could harm a patient. Similarly, an autonomous weapon system that mistakenly targets civilians raises profound ethical questions about delegating life-and-death decisions to machines [18].
- Regulatory Challenges: Existing legal frameworks are often ill-equipped to ad-

dress the unique challenges posed by AAS. For example, liability laws typically assume human agency, making applying them to autonomous systems difficult. Developing new regulatory frameworks that clearly define accountability for AAS is essential to ensure their ethical deployment [19].

Mitigation Strategies

- Explainable AI (XAI): Developing transparent AI systems that can explain their decision-making processes can help address accountability concerns. Explainable AI allows stakeholders to understand how decisions are made, facilitating accountability and trust [20].
- Human Oversight: Incorporating human oversight mechanisms, such as human-in-the-loop (HITL) systems, ensures that critical decisions are reviewed by humans, reducing the risk of harmful outcomes [21].

3.2. Bias and Discrimination

Bias in autonomous systems is a significant ethical concern, particularly in law enforcement, finance, and healthcare sectors. Machine learning algorithms underpin many AAS and are trained on large datasets that may contain biases, leading to unfair or discriminatory outcomes.

- Inherited Biases: AAS can perpetuate or amplify biases in their training data. For example, facial recognition systems have been shown to exhibit racial and gender biases, leading to higher error rates for specific demographic groups [22]. Similarly, predictive policing algorithms may disproportionately target minority communities if trained on biased historical crime data [23].
- Impact on Fairness: Bias in AAS can result in the unfair treatment of individuals or groups. In the financial sector, biased algorithms used for credit scoring or loan approvals may deny opportunities to specific populations. In healthcare, biased diagnostic tools may lead to unequal patient treatment based on race, gender, or socioeconomic status [24].
- Reinforcement of Stereotypes: AAS that rely on biased data can reinforce harmful stereotypes, perpetuating social inequalities. For instance, hiring algorithms trained on biased data may favor specific demographics over others, exacerbating workplace discrimination [25].

Mitigation Strategies

- Bias Detection and Mitigation: Techniques such as fairness-aware machine learning and adversarial debiasing can help identify and mitigate biases in training data and algorithms [26].
- Diverse and Representative Data: Ensuring that training datasets are diverse and representative of the population can reduce the risk of biased outcomes [27].
- Algorithmic Audits: Regular audits of AAS by independent third parties can help identify and address biases, ensuring compliance with ethical standards [28].

3.3. Privacy and Surveillance

The use of AAS in cybersecurity and surveillance contexts raises significant privacy concerns. These systems often collect, process, and store vast amounts of personal data, posing privacy and security risks.

- Data Collection and Usage: AAS in surveillance applications, such as facial recognition systems or smart city technologies, can collect sensitive personal data without individuals' consent. This raises ethical questions about the extent to which such data collection is justified and how the data is used [29].
- Compliance with Privacy Laws: AAS must comply with privacy regulations such as the General Data Protection Regulation (GDPR) in the European Union and the Health Insurance Portability and Accountability Act (HIPAA) in the United States. These laws mandate strict data collection, storage, and usage guidelines, requiring organizations to implement robust data governance frameworks [30].
- Surveillance and Civil Liberties: The widespread use of AAS can infringe on civil liberties, such as the right to privacy and freedom of movement. For example, deploying autonomous drones for public surveillance has sparked debates about the balance between security and individual rights [31].

Mitigation Strategies

- Privacy by Design: Incorporating privacy considerations into the design and development of AAS can help minimize data collection and ensure compliance with privacy laws [32].
- Data Anonymization: Data anonymization and encryption can protect individuals' privacy while allowing AAS to function effectively [33].
- Transparency and Consent: Ensuring transparency in data collection practices and obtaining informed consent from individuals can help build trust and ensure ethical data usage [34].

4. Zero Trust Frameworks: Principles and Application

The Zero Trust (ZT) security model has emerged as a critical framework for modern cybersecurity, particularly in an era where traditional perimeter-based defenses are increasingly inadequate. Zero Trust operates on the principle of "never trust, always verify," requiring continuous authentication and strict access controls to protect networks from both insider and external threats. This section explores the core tenets of Zero Trust, its implementation in cybersecurity, and the challenges and limitations associated with its adoption, particularly in the context of integrating autonomous agentic systems (AAS).

4.1. Implementation in Cybersecurity

Implementing a zero-trust framework involves a comprehensive approach to cybersecurity, integrating multiple technologies and practices to enforce its core principles. Key components of Zero Trust implementation include:

- Identity Verification: Robust identity and access management (IAM) systems

are essential for continuous authentication. Multi-factor authentication (MFA) and biometric verification are commonly used to ensure that only authorized users and devices can access network resources [35]. Challenge: How do you continuously authenticate an autonomous system, such as a drone or an AI-driven diagnostic tool, when it doesn't have a traditional "user" identity?

- Device Security: Zero Trust requires that all devices accessing the network meet strict security standards. This includes ensuring that devices are up-to-date with security patches, have endpoint protection installed, and are free from malware [36].
- Least Privilege Access: Access controls are dynamically adjusted based on user roles, device security posture, and contextual factors such as location and time of access. This ensures that users and devices can only access the resources they need at any given time [37]. How do you enforce least privilege access for AAS without hindering their ability to perform complex tasks? For example, an autonomous trading system may need access to vast amounts of financial data to make decisions, but granting such access could increase the risk of data breaches.
- Network Segmentation: Micro-segmentation is implemented to isolate critical assets and limit lateral movement within the network. Software-defined networking (SDN) and virtual LANs (VLANs) are often used to create these isolated segments [38]. The use of software-defined networking (SDN) to dynamically create and manage network segments based on the operational needs of AAS. This would allow for flexible segmentation that adapts to the system's requirements while maintaining security.
- Continuous Monitoring and Analytics: Zero Trust frameworks rely on continuous network activity monitoring to detect and respond to potential threats in real time. Security information and event management (SIEM) systems and machine learning algorithms are used to analyze traffic patterns and identify anomalies [39].

4.2. Case Study: Google's AI for Breast Cancer Screening

4.2.1. Background & Significance

Breast cancer is one of the leading causes of cancer-related deaths globally. Early detection significantly improves survival rates, yet traditional mammogram analysis has limitations, including false positives and false negatives.

4.2.2. AI Model Development

- Google Health developed a deep learning model trained on 76,000 mammograms from the UK and 15,000 from the US.
- To validate generalizability, the AI model was tested on independent datasets with over 25,000 UK and 3,000 US mammograms.

4.2.3. Key Findings

- Reduced False Positives by 5.7% (US) and 1.2% (UK) → Fewer unnecessary biopsies.

- Reduced False Negatives by 9.4% (US) and 2.7% (UK) → Fewer missed cancer diagnoses.
- Outperformed Radiologists: In retrospective analysis, the AI model detected cancer more accurately than experienced radiologists.
- Consistency: Unlike human experts, whose performance can vary due to fatigue or cognitive biases, AI provided uniform interpretations.

4.2.4. Ethical Governance Challenges in AI Diagnostics

Despite the success, several ethical concerns and governance challenges arise:

1) Transparency & Explainability

- The AI system's decision-making process is not always interpretable.
- Black-box models limit clinicians' ability to understand and trust AI-generated diagnoses.
- Ethical Governance Insight: Implement Explainable AI (XAI) techniques such as saliency maps to visualize which areas in a mammogram influenced the AI's prediction.

2) Bias & Generalizability

- The training dataset was predominantly from Western populations.
- AI models may underperform on underrepresented ethnic groups due to biased training data.
- Ethical Governance Insight: Regulatory frameworks should mandate diverse, representative datasets for training AI in medical applications.

3) Accountability in Misdiagnosis

- If an AI incorrectly classifies a tumor as benign, who is responsible? The developer, the hospital, or the physician?
- Ethical Governance Insight: A structured audit log tracking AI-generated diagnoses and clinician overrides can enhance accountability.

4) Privacy & Data Protection

- AI models require access to large datasets, raising concerns about patient data privacy and potential misuse.
- Ethical Governance Insight: Implement Privacy-Preserving Machine Learning (PPML) techniques, such as federated learning, to train AI models without direct access to patient data (Table 1).

Table 1. Practical takeaways for ethical AI governance.

Issue	Challenge	Ethical Governance Strategy
Explainability	AI decisions are hard to interpret.	Require **XAI techniques** for transparency.
Bias	AI may underperform on certain demographics.	Mandate **diverse training datasets**.
Accountability	Unclear responsibility for AI errors.	Develop **audit logs** and human oversight.
Privacy	AI requires sensitive health data.	Use **federated learning** for data protection.

4.3. Case Study: Google's BeyondCorp

Google introduced BeyondCorp as an alternative to traditional perimeter-based security models after a sophisticated cyberattack in 2009 (Operation Aurora). BeyondCorp enforces Zero-Trust principles, ensuring access control based on continuous authentication, device trustworthiness, and dynamic risk evaluation.

4.3.1. How AI Is Integrated into BeyondCorp

- AI-Driven Continuous Authentication
  - AI models assess user and device behavior in real-time, detecting anomalies that might indicate compromised credentials or insider threats.
  - AI can trigger additional authentication steps if a user logs in from an unusual location or at an odd time.
- Automated Risk Scoring & Adaptive Access
  - AI analyzes risk signals such as login patterns, endpoint security posture, and network requests to adjust access levels dynamically.
  - Employees are only granted the minimum access required (Principle of Least Privilege).
- Self-Healing Endpoint Security
  - Google's AI-driven Endpoint Verification system ensures all devices connecting to internal applications are compliant with security policies.
  - Devices found vulnerable (e.g., outdated software, malware infection) are automatically quarantined until they meet compliance requirements.

4.3.2. Ethical Considerations in AI-Driven BeyondCorp

- Transparency & Explainability → Google ensures audit logs and explainability mechanisms are in place to justify AI-driven access control decisions.
- Privacy Protection → AI-driven authentication minimizes data collection, ensuring compliance with privacy laws like GDPR and CCPA.
- Bias Mitigation → AI models are regularly audited to avoid discriminatory access control decisions that might disproportionately affect certain employees.

Google's BeyondCorp initiative is a prominent example of Zero Trust implementation. BeyondCorp shifts access controls from the network perimeter to individual users and devices, ensuring access is granted based on identity and device security rather than network location. This approach has significantly enhanced Google's security posture, reducing the risk of insider threats and external attacks [40].

4.4. Challenges and Limitations

While Zero Trust offers significant security benefits, its implementation is challenging. These challenges are further compounded when autonomous agentic systems (AAS) are integrated into Zero Trust frameworks.

1) Infrastructure Investment: Implementing Zero Trust requires significant investment in new technologies, such as IAM systems, endpoint protection, and network segmentation tools. Organizations must also invest in training and re-

configuring their IT infrastructure, which can be costly and time-consuming [35].

2) System Efficiency: The continuous authentication and monitoring required by Zero Trust can introduce latency and reduce system efficiency. This is particularly problematic in environments where real-time performance is critical, such as financial trading or healthcare [36].

3) Complexity of Integration: Integrating AAS into Zero Trust frameworks introduces additional complexities. Autonomous systems often operate with minimal human intervention, making applying traditional authentication and access control mechanisms difficult. For example, how does one verify the identity of an autonomous drone or ensure that an AI-driven diagnostic tool complies with least privilege principles [38].

4) Trust Verification: AAS rely on machine learning models that may be opaque or difficult to interpret, raising questions about how to verify their trustworthiness. Ensuring these systems comply with Zero Trust principles requires new approaches to explainability and transparency [20].

5) Scalability: Zero Trust frameworks must be scalable to accommodate the growing number of devices and users in modern networks. This is particularly challenging in large organizations or those with distributed workforces, where maintaining consistent security policies across all endpoints can be difficult [37].

4.5. Mitigation Strategies

- Phased Implementation: Organizations can adopt a phased approach to Zero Trust implementation, starting with critical assets and gradually expanding to the entire network. This reduces the initial investment and allows for iterative improvements [35].
- Automation and AI: Leveraging automation and AI can help manage the complexity of Zero Trust frameworks. For example, AI-driven analytics can enhance continuous monitoring, while automation can streamline access control and policy enforcement [36].
- Collaboration with Vendors: Partnering with technology vendors and cybersecurity experts can help organizations overcome implementation challenges and ensure compliance with Zero Trust principles [38].

5. Microsoft's Zero Trust Deployment: AI-Powered Least Privilege Access

5.1. Background

Microsoft has been a leading advocate of Zero Trust security across its enterprise and cloud services. The company applies AI-powered security solutions to manage device security, user authentication, and network access controls [41].

5.2. How AI Is Integrated into Microsoft's Zero Trust Framework

- AI-Driven Least Privilege Enforcement
  - Microsoft uses Azure Active Directory Conditional Access Policies to grant

user access dynamically based on AI-evaluated risk.

- Example: AI detects unusual login activity (e.g., logging in from a high-risk country) and enforces additional security checks.

- Automated Threat Detection with Microsoft Defender
- AI models analyze billions of security signals daily to detect anomalous network behavior, insider threats, and malware intrusions.
- Microsoft Defender for Endpoint uses AI to predict and prevent advanced cyberattacks in real-time [41].

- AI-Powered Compliance Monitoring
- Microsoft Compliance Manager uses AI to track regulatory compliance across industries, ensuring that security policies align with standards like NIST, ISO/IEC 27001, and GDPR.
- AI can automate remediation, reducing compliance gaps without manual intervention (Table 2).

Table 2. Comparison of Google and Microsoft's AI-driven zero trust approaches.

Feature	Google BeyondCorp	Microsoft Zero Trust Deployment
Authentication	AI-driven continuous authentication	AI-powered adaptive conditional access
Threat Detection	AI-based anomaly detection in access requests	AI-powered Defender detects advanced cyber threats
Device Security	AI-driven endpoint verification	AI-enforced device compliance policies
Least Privilege	Dynamic role-based access using AI	AI-managed conditional least privilege access
Compliance Monitoring	Audit logs ensure transparency	AI automates compliance with global regulations
Privacy Protection	AI minimizes personal data collection	AI models trained on anonymized datasets

6. Intersection of Ethical Governance and Zero Trust in Autonomous Systems

Integrating autonomous agentic systems (AAS) into Zero Trust (ZT) frameworks presents unique challenges at the intersection of ethical governance and cybersecurity. As AAS becomes increasingly prevalent in critical sectors, ensuring accountability, transparency, and privacy within these systems is paramount. This section explores how ethical governance principles can be embedded into Zero Trust frameworks to address these challenges, focusing on accountability, transparency, and privacy protection.

6.1. Ensuring Accountability in Autonomous Agents

Accountability is a cornerstone of ethical governance, particularly in autonomously operating systems. In the context of AAS, accountability ensures that decisions

and actions can be traced back to their source, whether human or machine, and that appropriate measures are in place to address errors or misconduct.

- Audit Trails and Logging: Embedding audit trails within AAS is essential for tracking decision-making processes. These logs record the system's actions, including data inputs, algorithmic decisions, and outputs. For example, an autonomous diagnostic system in healthcare should maintain detailed logs of patient data, diagnostic criteria, and treatment recommendations to ensure accountability in case of errors [17].
- Regulatory Compliance: Accountability mechanisms must align with regulatory requirements, such as the General Data Protection Regulation (GDPR) in the EU or the Health Insurance Portability and Accountability Act (HIPAA) in the US. These regulations mandate that organizations implement measures to ensure transparency and accountability in automated decision-making processes [30].
- Human Oversight: Incorporating human-in-the-loop (HITL) systems ensures that humans review critical decisions made by AAS. This not only enhances accountability but also provides a safeguard against unintended consequences. For instance, human oversight can intervene in autonomous vehicles when the system's decision-making is uncertain or risky [21].

Mitigation Strategies:

- Blockchain for Immutable Logs: Blockchain technology can create immutable audit trails, ensuring that logs cannot be tampered with and providing a transparent record of system actions [42].
- Accountability Frameworks: Developing standardized accountability frameworks for AAS can help organizations ensure compliance with ethical and regulatory standards [19].

6.2. Transparency and Explainability in AI Decision-Making

Transparency and explainability are critical for building trust in AAS, particularly in high-stakes applications such as healthcare, finance, and defense. Explainable AI (XAI) ensures that the decision-making processes of autonomous systems are understandable to stakeholders, including users, regulators, and affected individuals.

- Explainable AI (XAI): XAI techniques, such as decision trees, rule-based systems, and model-agnostic methods, provide insights into how AAS arrive at their decisions. For example, in financial services, explainable algorithms can help regulators understand why a loan application was denied, ensuring compliance with anti-discrimination laws [20].
- Stakeholder Trust: Transparency in AAS fosters trust among users and stakeholders. In healthcare, for instance, patients are more likely to trust AI-driven diagnostic tools if they understand how recommendations are made [8].
- Ethical Guidelines: Ethical frameworks, such as the EU's Ethics Guidelines for Trustworthy AI, emphasize the importance of transparency and explainability

in AI systems. These guidelines recommend that AI decisions be explainable to users and that systems provide transparent information about their capabilities and limitations [43].

Mitigation Strategies:

- Model Interpretability: Using interpretable machine learning models, such as linear regression or decision trees, can enhance transparency in AAS [44].
- User-friendly explanations: User-friendly explanations, such as visualizations or natural language summaries, can make AI decisions more accessible to non-technical stakeholders [45].

6.3. Privacy and Data Protection within Zero Trust Frameworks

Zero Trust frameworks enhance data security by enforcing strict access controls and continuous monitoring. However, integrating AAS into these frameworks requires balancing security with privacy protection, ensuring that sensitive data is handled ethically and complies with regulations.

- Data Minimization: Zero Trust principles align with data minimization, a key tenet of privacy regulations such as GDPR. By limiting data collection and access to what is strictly necessary, organizations can reduce the risk of privacy breaches [30].
- Encryption and Anonymization: Encrypting data at rest and in transit, as well as anonymizing sensitive information, ensures that even if data is accessed, it cannot be easily exploited. For example, patient data can be anonymized in healthcare before training AI models, protecting individual privacy [33].
- Access Control and Least Privilege: Zero Trust's principle of least privilege ensures that AAS only access the data necessary for their tasks. This reduces the risk of unauthorized access and data breaches. For instance, an autonomous financial trading system should only have access to the data required for its trading algorithms, not the entire customer database [35].
- Ethical Data Governance: Ethical governance frameworks should ensure that AAS complies with privacy regulations while maintaining operational efficiency. This includes implementing data protection impact assessments (DPIAs) and ensuring that data usage is transparent and consensual [32].

Mitigation Strategies:

- Privacy by Design: Incorporating privacy considerations into the design and development of AAS ensures that data protection is a core component of the system [32].
- Regular Audits: Regular audits of data access and usage within Zero Trust frameworks help identify and address potential privacy risks [28].

Ethical Considerations in AI-Driven Zero Trust

- Fairness & Non-Discrimination → AI models undergo bias testing to prevent unfair enforcement of access controls (e.g., discriminatory blocking of specific users or locations).
- Accountability & Explainability → Microsoft ensures that AI-driven security

decisions are transparent, allowing security teams to audit and override when necessary.

- Data Security & Privacy → AI models are trained on anonymized, aggregated data, ensuring privacy is not compromised while detecting security threats.

7. Policy Recommendations and Future Directions

Integrating autonomous agentic systems (AAS) into Zero Trust (ZT) frameworks requires a proactive approach to policymaking, ethical governance, and research. As these systems become more pervasive in critical sectors, it is essential to establish regulatory standards, adopt best practices for ethical AI governance, and identify future research directions to address emerging challenges. This section outlines policy recommendations and future directions to ensure the responsible deployment of AAS within ZT environments.

7.1. Regulatory Standards and Compliance

To ensure the ethical and secure deployment of AAS in Zero Trust frameworks, governments, and industry bodies must establish robust regulatory standards and compliance mechanisms. These frameworks should address cybersecurity and ethical considerations, ensuring AAS operates transparently, accountably, and in alignment with societal values.

- AI Ethics Guidelines: Regulatory frameworks should incorporate AI ethics guidelines, such as those proposed by the European Commission's High-Level Expert Group on AI. These guidelines emphasize principles such as fairness, transparency, accountability, and human oversight, which are critical for the ethical deployment of AAS [43].
- Cybersecurity Compliance: AAS operating within ZT frameworks must comply with cybersecurity standards, such as the NIST Cybersecurity Framework and ISO/IEC 27001. These standards provide guidelines for implementing Zero Trust principles, including continuous authentication, least privilege access, and micro-segmentation [36] [46].
- Data Protection Regulations: AAS must adhere to data protection regulations, such as the General Data Protection Regulation (GDPR) in the EU and the California Consumer Privacy Act (CCPA) in the US. These regulations mandate strict data collection, storage, and usage controls, ensuring that AAS handles sensitive information responsibly [30].
- Cross-Border Collaboration: Given the global nature of cybersecurity threats, international collaboration is essential to harmonize regulatory standards and ensure consistent enforcement. Organizations such as the United Nations and the OECD can play a key role in facilitating cross-border cooperation on AI and cybersecurity governance [19].

7.1.1. Critical Assessment of Existing Regulatory Frameworks

1) General Data Protection Regulation (GDPR)

- Strengths: GDPR is one of the most comprehensive data protection frame-

works, emphasizing transparency, accountability, and user consent. It requires organizations to implement measures such as data minimization, encryption, and privacy by design.

- Limitations for AAS and ZT:
  - GDPR focuses primarily on human data subjects and does not explicitly address the challenges posed by autonomous systems. For example, how can consent be obtained from an autonomous drone, or how can transparency be ensured in AI-driven decision-making?
  - The regulation's right to explanation (Article 22) is challenging to apply to complex AI models, especially in real-time AAS applications like autonomous vehicles or financial trading systems.

2) NIST Cybersecurity Framework

- Strengths: The NIST framework provides a flexible, risk-based approach to cybersecurity, emphasizing continuous monitoring, incident response, and access control. It aligns well with Zero Trust principles.
- Limitations for AAS and ZT:
  - The framework does not explicitly address the integration of autonomous systems or the ethical implications of AI-driven decision-making.
  - It lacks detailed guidance on implementing least privilege access for AAS, which often requires broad access to data and systems.

3) EU Ethics Guidelines for Trustworthy AI

- Strengths: These guidelines emphasize fairness, transparency, and human oversight, which are critical for the ethical deployment of AAS.
- Limitations for AAS and ZT:
  - The guidelines are non-binding and lack specific technical or regulatory requirements for implementation.
  - They do not address integrating AI ethics with Zero Trust cybersecurity frameworks, leaving a gap in balancing ethical considerations with security requirements [47] (**Table 3**).

Table 3. Comparison of Google and Microsoft's AI-driven zero trust approaches.

Framework	Strengths	Limitations for AAS and ZT	Proposed Innovations
GDPR	Emphasizes transparency, accountability, and user consent.	Focuses on human data subjects; difficult to apply to autonomous systems.	AAS-specific accountability frameworks; dynamic access control policies.
NIST Cybersecurity	Flexible, risk-based approach; aligns with Zero Trust principles.	Lacks guidance on integrating AAS and addressing ethical implications of AI.	Explainable AI mandates; ethical impact assessments.
EU Ethics Guidelines	Emphasizes fairness, transparency, and human oversight.	Non-binding; lacks technical or regulatory requirements for implementation.	Cross-border regulatory harmonization; mandatory EIAs for AAS.

7.1.2. Policy Recommendations

1) Develop AAS-Specific Regulations: Governments and industry bodies should develop regulations specifically tailored to the unique challenges of AAS, including accountability, transparency, and dynamic access control.

2) Mandate Explainable AI: Regulatory frameworks should require explainable AI techniques in AAS to ensure transparency and accountability.

3) Conduct Ethical Impact Assessments: Organizations deploying AAS should be required to conduct EIAs to evaluate their systems' ethical, privacy, and societal implications.

4) Harmonize International Standards: International organizations should facilitate the development of harmonized regulatory standards for AAS and ZT to address global cybersecurity threats [19].

7.2. Best Practices for Ethical AI Governance

Adopting best practices for ethical AI governance is essential to mitigate risks associated with AAS and ensure their responsible deployment within Zero Trust frameworks. These practices should focus on bias detection, continuous monitoring, and transparency.

- Bias Detection Mechanisms: Implementing bias detection and mitigation techniques is critical to ensuring fairness in AAS. Tools such as fairness-aware machine learning and adversarial debiasing can help identify and address biases in training data and algorithms [24] (**Figure 2**).

	Study	Focus Area	Key Findings	Mitigation Strategies
1	Buolamwini & Gebru (2018)	Bias in Facial Recognition	Commercial facial recognition models showed error rates of up to 34% for	Improve dataset diversity; use fairness-aware training algorithms.
2	O'Neil (2016)	Bias in Predictive Policing & Hiring	Predictive policing algorithms disproportionately targeted minority communities;	Ensure diverse training data; conduct fairness audits; regulate AI use in
3	Mehrabi et al. (2021)	Bias in AI Decision-Making	Surveyed various biases in AI systems and recommended frameworks for	Use adversarial debiasing; ensure explainability and fairness-aware ML techniques.
4	Raji et al. (2020)	Algorithmic Auditing for Fairness	Proposed an internal auditing framework for AI models to address bias	Regular audits and impact assessments before deploying AI models.
5	Zemel et al. (2013)	Fair Representation Learning	Developed a method for training AI models to produce fairer representations of	Fairness constraints during model training to balance outcomes across groups.

Figure 2. AI fairness studies and findings.

- Continuous Monitoring and Auditing: AAS should be subject to constant monitoring and auditing to ensure compliance with ethical and cybersecurity standards. This includes tracking system behavior, identifying anomalies, and addressing potential risks in real-time [28].
- AI Transparency: Promoting transparency in AAS through open-source standards and explainable AI (XAI) techniques can enhance trust and accountability. For example, organizations can publish model architectures, training data, and decision-making criteria to ensure transparency [20].
- Stakeholder Engagement: Engaging stakeholders, including users, regulators, and affected communities, in the development and deployment of AAS ensures that diverse perspectives are considered. This can help identify potential ethical risks and build public trust [17].

Best Practices

- Ethical Impact Assessments: Conduct ethical impact assessments (EIAs) to evaluate AAS's potential risks and benefits before deployment.
- Human-in-the-Loop Systems: Incorporate human oversight mechanisms to ensure that critical decisions made by AAS are reviewed by humans [21].

7.3. Future Research Needs

As AAS and Zero Trust frameworks evolve, further research is needed to address emerging challenges and advance the state of the art in AI ethics and cybersecurity.

- Integrating AI Ethics with Cybersecurity Frameworks: Research is needed to develop integrated frameworks that combine AI ethics principles with zero-trust cybersecurity practices. This includes exploring how ethical considerations, such as fairness and accountability, can be embedded into ZT architectures [19].
- Unbiased Machine Learning Models: Developing unbiased machine learning models is critical to ensuring fairness in AAS. This includes researching techniques for bias detection and mitigation and creating diverse and representative training datasets [24].
- Enhancing Zero Trust Automation: Automating Zero Trust processes, such as continuous authentication and access control, can improve efficiency and scalability. Research is needed to develop AI-driven tools that enhance ZT automation while maintaining security and compliance [35].
- Privacy-Preserving Technologies: Advancements in privacy-preserving technologies, such as federated learning and homomorphic encryption, can enhance data protection in AAS. These technologies enable secure data sharing and analysis without compromising privacy [48].
- Ethical and Legal Implications of AAS: Further research is needed to explore the ethical and legal implications of AAS, particularly in high-stakes applications such as healthcare and defense. This includes addressing questions of liability, accountability, and the societal impact of autonomous systems [18] (**Figure 3**).

	Metric Type	Quantitative Metrics	Reference Framewor
1	Cybersecurity: AI-driven Intrusion Detection	Reduction in threat response times (e.g., AI-based IDS reduced detection time by	NIST Cybersecurity Framework, MITRE ATT&CK
2	Financial Sector: Algorithmic Trading	Increased trading efficiency (e.g., AI-driven trading improved return rates by 15%	Risk-adjusted return models, SEC AI trading regulations
3	Healthcare: AI-assisted Diagnostics	Improved diagnostic accuracy (e.g., AI models improved early cancer	AI in Healthcare standards, WHO AI Ethics guidelines
4	Fairness in AI Decisions	Disparity impact ratio, equal opportunity difference (used in AI Fairness 360	IBM AI Fairness 360 Toolkit, Fairness-aware ML frameworks
5	Transparency in AI Systems	Explainability score, interpretability level of models (LIME, SHAP	Explainable AI (XAI), SHAP, LIME methodologies
6	Accountability in AI Governance	Audit trail completeness, human-in-the-loo p intervention rates	ISO/IEC 38500 AI Governance, AI ethics regulatory compliance

Figure 3. Performance and ethical impact metrics for AAS.

7.4. Research Priorities

- Interdisciplinary Collaboration: Foster collaboration between AI researchers, cybersecurity experts, ethicists, and policymakers to address complex challenges at the intersection of AI and cybersecurity.
- Long-Term Impact Studies: Conduct longitudinal studies to assess the long-term impact of AAS on society, including their effects on employment, privacy, and human rights.

8. Conclusion

Integrating autonomous agentic systems (AAS) into Zero Trust (ZT) frameworks represents a transformative shift in cybersecurity and operational efficiency across critical sectors such as healthcare, finance, defense, and infrastructure. However, this convergence also introduces significant ethical and security challenges that must be addressed to ensure the responsible deployment of these technologies. Throughout this article, we have explored the moral considerations, technical complexities, and governance frameworks necessary to navigate the intersection of AAS and Zero Trust. To ensure ethical deployment, organizations must; Im-

plement explainable AI (XAI) to improve trust and transparency in decision-making. Adopt bias-mitigation strategies and diverse training datasets to ensure fairness. Establish clear accountability frameworks, including audit trails and human oversight. Balance data security with privacy protection using compliance-driven approaches like Privacy-Preserving Machine Learning (PPML). Industry leaders like Google and Microsoft have successfully deployed AI-driven Zero Trust models, demonstrating the potential of AI-powered adaptive security while maintaining ethical compliance. Moving forward, policymakers, AI developers, and cybersecurity experts must collaborate to establish regulatory standards that align innovation with ethical responsibility. Organizations can leverage AAS within Zero Trust frameworks to create a more secure, transparent, and ethically responsible digital future by proactively addressing AI fairness, accountability, and governance.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] Russell, S. and Norvig, P. (2020) Artificial Intelligence: A Modern Approach. 4th Edition, Pearson.

[2] Goodfellow, I., Bengio, Y. and Courville, A. (2016) Deep Learning. MIT Press.

[3] Wooldridge, M. (2009) An Introduction to Multi-Agent Systems. 2nd Edition, Wiley.

[4] Sharkey, N. (1920) Staying in the Loop: Human Supervisory Control of Weapons. In: *Autonomous Weapons Systems*: *Law*, *Ethics*, *Policy*, Cambridge University Press, 23-38. https://doi.org/10.1017/cbo9781316597873.002

[5] Brynjolfsson, E. and McAfee, A. (2014) The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies. W.W. Norton & Company.

[6] McKinney, S.M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., *et al.* (2020) International Evaluation of an AI System for Breast Cancer Screening. *Nature*, **577**, 89-94. https://doi.org/10.1038/s41586-019-1799-6

[7] Lanfranco, A.R., Castellanos, A.E., Desai, J.P. and Meyers, W.C. (2004) Robotic Surgery: A Current Perspective. *Annals of Surgery*, **239**, 14-21. https://doi.org/10.1097/01.sla.0000103020.19595.7d

[8] Topol, E.J. (2019) Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again. Basic Books.

[9] Sarker, I.H., Kayes, A.S.M. and Watters, P. (2019) Effectiveness Analysis of Machine Learning Classification Models for Predicting Personalized Context-Aware Smartphone Usage. *Journal of Big Data*, **6**, Article No. 57. https://doi.org/10.1186/s40537-019-0219-y

[10] Kott, A., Ludwig, J. and Lange, M. (2019) Autonomous Cyber Defense: A Review and a Roadmap. *IEEE Security & Privacy*, **17**, 76-85.

[11] Buczak, A.L. and Guven, E. (2016) A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection. *IEEE Communications Surveys & Tutorials*, **18**, 1153-1176. https://doi.org/10.1109/comst.2015.2494502

[12] Dal Pozzolo, A., Caelen, O., Le Borgne, Y., Waterschoot, S. and Bontempi, G. (2014) Learned Lessons in Credit Card Fraud Detection from a Practitioner Perspective. *Ex-*

pert Systems with Applications, **41**, 4915-4928.
https://doi.org/10.1016/j.eswa.2014.02.026

[13] Treleaven, P., Galas, M. and Lalchand, V. (2013) Algorithmic Trading Review. *Communications of the ACM*, **56**, 76-85. https://doi.org/10.1145/2500117

[14] Kou, G., Lu, Y., Peng, Y. and Shi, Y. (2014) Evaluation of Classification Algorithms Using MCDM and Rank Correlation. *International Journal of Information Technology & Decision Making*, **13**, 407-425. https://arxiv.org/pdf/2304.12408

[15] Arkin, R.C. (2009) Governing Lethal Behavior in Autonomous Robots. CRC Press.

[16] Cummings, M.L. (2017) Artificial Intelligence and the Future of Warfare. *International Security*, **41**, 7-38.

[17] Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., *et al.* (2018) Ai4people—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines*, **28**, 689-707.
https://doi.org/10.1007/s11023-018-9482-5

[18] Sparrow, R. (2016) Robots and Respect: Assessing the Case against Autonomous Weapon Systems. *Ethics & International Affairs*, **30**, 93-116.
https://doi.org/10.1017/s0892679415000647

[19] Cath, C. (2018) Governing Artificial Intelligence: Ethical, Legal and Technical Opportunities and Challenges. *Philosophical Transactions of the Royal Society A*: *Mathematical, Physical and Engineering Sciences*, **376**, Article ID: 20180080.
https://doi.org/10.1098/rsta.2018.0080

[20] Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S. and Yang, G. (2019) XAI—Explainable Artificial Intelligence. *Science Robotics*, **4**, eaay7120.
https://doi.org/10.1126/scirobotics.aay7120

[21] Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J. and Mané, D. (2016) Concrete Problems in AI Safety.

[22] Buolamwini, J. and Gebru, T. (2018) Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Proceedings of the* 1*st Conference on Fairness, Accountability and Transparency*, New York, 23-24 February 2018, 81-91.

[23] O'Neil, C. (2016) Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. Crown Publishing Group.

[24] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K. and Galstyan, A. (2021) A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys*, **54**, 1-35.
https://doi.org/10.1145/3457607

[25] Raghavan, M., Barocas, S., Kleinberg, J. and Levy, K. (2020) Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices. *Proceedings of the* 2020 *Conference on Fairness, Accountability, and Transparency*, Barcelona, 27-30 January 2020, 469-481. https://doi.org/10.1145/3351095.3372828

[26] Zemel, R., Wu, Y., Swersky, K., Pitassi, T. and Dwork, C. (2013) Learning Fair Representations. *Proceedings of the* 30*th International Conference on Machine Learning*, Atlanta, 16-21 June 2013, 325-333.

[27] Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J.W., Wallach, H., Daumé III, H. and Crawford, K. (2018) Datasheets for Datasets. *Proceedings of the* 5*th Workshop on Fairness, Accountability, and Transparency in Machine Learning*, New York, 23-24 February 2018, 17 p.
https://www.microsoft.com/en-us/research/wp-content/uploads/2019/01/1803.09010.pdf

[28] Raji, I.D., Smart, A., White, R.N., Mitchell, M., Gebru, T., Hutchinson, B., *et al.*

(2020). Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing. *Proceedings of the* 2020 *Conference on Fairness, Accountability, and Transparency*, Barcelona, 27-30 January 2020, 33-44. https://doi.org/10.1145/3351095.3372873

[29] Zimmer, M. (2010) "But the Data Is Already Public": On the Ethics of Research in Facebook. *Ethics and Information Technology*, **12**, 313-325. https://doi.org/10.1007/s10676-010-9227-5

[30] Voigt, P. and Von dem Bussche, A. (2017) The EU General Data Protection Regulation (GDPR): A Practical Guide. Springer.

[31] Lyon, D. (2018) The Culture of Surveillance: Watching as a Way of Life. Polity Press.

[32] Cavoukian, A. (2012) Privacy by Design: The 7 Foundational Principles. Information and Privacy Commissioner of Ontario, Canada.

[33] Narayanan, A. and Shmatikov, V. (2010) Myths and Fallacies of "Personally Identifiable Information". *Communications of the ACM*, **53**, 24-26. https://doi.org/10.1145/1743546.1743558

[34] Acquisti, A., Brandimarte, L. and Loewenstein, G. (2015) Privacy and Human Behavior in the Age of Information. *Science*, **347**, 509-514. https://doi.org/10.1126/science.aaa1465

[35] Rose, S., Borchert, O., Mitchell, S. and Connelly, S. (2020) Zero Trust Architecture. NIST Special Publication 800-207.

[36] NIST (2020) Zero Trust Architecture. National Institute of Standards and Technology Special Publication 800-207. https://nvlpubs.nist.gov/nistpubs/specialpublications/NIST.SP.800-207.pdf

[37] Cunningham, C. (2018) The Zero Trust eXtended Ecosystem. Forrester Research.

[38] TechDemocracy (2024) Zero Trust Framework|Zero Trust Principles|TechDemocracy Blog. TechDemocracy—Leader in Identity Security Solutions—IAM. https://www.techdemocracy.com/resources/zero-trust-framework-79

[39] Kindervag, J. (2010) No More Chewy Centers: Introducing the Zero Trust Model of Information Security. Forrester Research. https://www.forrester.com/report/No-More-Chewy-Centers-The-Zero-Trust-Model-Of-Information-Security/RES56682

[40] Ward, R. and Beye, B. (2014) BeyondCorp: A New Approach to Enterprise Security. Google Cloud.

[41] Microsoft (n.d.) Zero Trust Guidance Center. Microsoft Learn: Build Skills That Open Doors in Your Career. https://learn.microsoft.com/en-us/security/zero-trust/

[42] Yli-Huumo, J., Ko, D., Choi, S., Park, S. and Smolander, K. (2016) Where Is Current Research on Blockchain Technology? A Systematic Review. *PLOS ONE*, **11**, e0163477. https://doi.org/10.1371/journal.pone.0163477

[43] High-Level Expert Group on AI (2019) Ethics Guidelines for Trustworthy AI. European Commission. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

[44] Ribeiro, M.T., Singh, S. and Guestrin, C. (2016) "Why Should I Trust You?" Explaining the Predictions of Any Classifier. *Proceedings of the* 22*nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, 13-17 August 2016, 1135-1144. https://doi.org/10.1145/2939672.2939778

[45] Lipton, Z.C. (2018) The Mythos of Model Interpretability. *Communications of the ACM*, **61**, 36-43. https://doi.org/10.1145/3233231

[46] ISO/IEC (2013) ISO/IEC 27001:2013 Information Security Management. Interna-

tional Organization for Standardization. https://www.iso.org/standard/88435.html

[47] European Commission (2021) Proposal for a Regulation on a European Approach for Artificial Intelligence. European Commission.

[48] Yang, Q., Liu, Y., Chen, T. and Tong, Y. (2019) Federated Machine Learning: Concept and Applications. *ACM Transactions on Intelligent Systems and Technology*, **10**, 1-19. https://doi.org/10.1145/3298981